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

MMTC Communications – Review

Vol. 10, No. 2, April 2019

IEEE COMMUNICATIONS SOCIETY

TABLE OF CONTENTS

Message from the Review Board Directors

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

This issue comprises five reviews that cover multiple facets of multimedia communication research including image padding, 3D sensing, image recognition, and anomaly detection. These reviews are briefly introduced below.

The first paper is published IEEE International Conference on Acoustics, Speech, and Signal Processing 2019 and edited by Dr. Gwendal Simon. It proposes a solution that exploits the awareness of the texture adjacency in the 3D mesh to improve the encoding in giant texture images.

The second paper is published in 29th British Machine Vision Conference and edited by Dr. Jun Zhou. It proposes an explicit boundary handling rule for CNNs, and the filters of CNN are learned as usual at the internal pixels.

The third paper, published in IEEE Transactions on Multimedia and edited by Dr. Ye Liu, investigates how a hybrid deep learning based anomaly detection module can generate real-time anomaly report.

The fourth paper is published in the IEEE International Conference on Multimedia and Expo (ICME) 2018 and edited by Dr. Carsten Griwodz. The paper uses sensor fusion between RGBD video, gyroscope and accelerometer for solving a rather hard challenge (deblurring of depth images) without resorting to the magic of artificial intelligence and machine learning.

The fifth paper, published in IEEE Transactions on Multimedia and edited by Dr. Bruno Macchiavello, presents a simple personalization framework, which is a combination of the nearest class mean classifier and the 1-nearest neighbor classifier based on deep features..

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

Texture Images are not Regular Images: How to Leverage Geometry Information?

A short review for "A Geometry-Aware Framework for Compressing 3D Mesh Textures"

Edited by Gwendal Simon

F. Nasiri, N. Mahmoudian, F. Payan, and T. Maugey, "A Geometry-Aware Framework for Compressing 3D Mesh Textures" IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), 2019.

The multimedia applications that require processing 3D scenes have recently flourished beyond the traditional sector of video games and professional visual effects video edition. The growth of virtual reality and augmented reality applications has been tremendous over the past four years, and both areas are only the tip of an iceberg that includes the generation of photorealistic images and videos by game engines. 3D models are everywhere, from the captured representations of our world to the creations that designers are now able to build in dedicated 3D creation environments. These 3D models are also widely shared on the Internet, which opens a wide bunch of challenges and open problems, a key one being adaptive and fast delivery.

3D models are traditionally represented by both a geometric 3D mesh and the set of textures that apply on the surface of the geometric mesh. The compression of mesh has received a considerable attention in the early 2000's. Mature and efficient compression algorithms are now available, such as the DRACO open-source library. Textures are the heaviest component of a 3D model, and they have naturally become the focus of compression experts. A texture is usually encoded into a set of images, each image encoding a given characteristic of the material. This characteristic is processed by the renderer. It is common to have albedo, smoothness, metal/specular, normal, and ambient occlusion maps for one single texture.

Two approaches have been developed to compress the texture images. A first idea is to leverage image compression techniques such as developed at JPEG. The images get a high compression rate, typically 50:1. However, these compression techniques do not enable a fast access to small blocks of pixels in GPU. The random-access property of a decoder has become a requirement of modern 3D renderers. With JPEG-based image compression, the whole images should be uncompressed at the GPU to be used by the renderer, which results in high GPU memory and bandwidth consumption. A second approach is to use specific texture compression techniques such as DXT, which performs at typical 6:1 compression rate. The compressed images support random access to 4x4, 8x8, or 12x12 blocks of pixels in GPU. However, the network can become a bottleneck with low compression rate. The research community is actively looking for better solutions subject to both system and network constraints.

This paper offers a refreshing view on texture image compression by focusing on atlasing, which is a characteristic of the texture images that has been overlooked so far. It is indeed frequent that a vast texture image contains a collection of a potentially large number of smaller images, each of these smaller images applying on a particular face of the geometric mesh. These images are packed into one common giant texture, with resolution frequently greater than 8k. The authors of the paper highlights here that atlas encoding has not been carefully studied, in particular with respect to the geometric mesh on top of which they are projected.

The goal of the paper is to propose a set of techniques that exploit the awareness of the texture adjacency in the 3D mesh to improve their encoding in the giant texture image. The authors study in particular three techniques: The first one is to improve the intra-prediction a fundamental tool for the encoding of images using video compression techniques by copying pixel information on the borders of some blocks. The second technique addresses the inefficiency of raster scanning for distant blocks. The idea here is to scan blocks while taking into account their geometric proximity rather than their proximity into the giant image. Finally, the third technique is related to the encoding of the residual signal, which is less efficient here since blocks are not squared. The author proposes a graph-based approach to modify the residual.

The paper does not hide the fact that the authors are at the early stages of their research work in the area. It is clear that each of the ideas would deserve a longer development and a comprehensive dedicated performance evaluation. However, the authors present in the paper the result of their experience implementing a combination of the three techniques on a set of textures. The results are encouraging, with improvement ranging from 11% to 29%. Since the developed techniques do not prevent the use of existing GPU image processing pipeline at the decoding side, this result is an exciting achievement.

The paper does however neglect the aforementioned problem of random access to specific blocks in the images. The traditional compression approaches (whether they are based on JPEG or MPEG image compression algorithms) easily outperform any other texture compression method in terms of storage and delivery of the image. But these compressed images can hardly be processed in practice due to the need for full image decompression in GPU, which results in over-consumption of the GPU memory and bandwidth. Although it is an exciting new approach, the use of the information of the geometric mesh on the encoding of texture image is thus still a controversial technique. This paper opens the debate and shows potentials that cannot be overlooked.

Gwendal Simon is a Full Professor at IMT Atlantique, an elite technological university in France. His research interests include multimedia delivery systems (video and gaming) and network management. He has written more than 80 scientific papers, 4 of them having been awarded as best papers in prestigious conferences. His impactful research has also resulted in six patents, several software and datasets (including the awarded Solipsis p2p virtual world), and contributions to multiple innovative collaborative projects (including two awarded projects). He has advised 13 PhD students and he has directed a research lab (including 6 post-docs and research engineers). He graduated from University Rennes 1 (France). He obtained a PhD in Computer Science in 2004 and a Habilitation in 2015. From 2001 to 2006 he was a researcher at the research center of Orange (then France Telecom). Since 2006, he has been Associate Professor, and then a Full Professor at IMT Atlantique. He was a visiting researcher at University of Waterloo in 2011/2012 and he held a position of senior scientist at Adobe in 2018/2019.

Image Padding via Deep Learning

A short review for "Learning on the Edge: Explicit Boundary Handling in CNNs"

Edited by Dr. Jun Zhou

Carlo Innamorati, Tobias Ritschel, Tim Weyrich, Niloy Mitra, Learning on the Edge: Explicit Boundary Handling in CNNs," 29th British Machine Vision Conference, 2018.

When making neighborhood operations on an image, for example image filtering, an unavoidable task is to handle the boundary of the image. While this can be done by processing the internal pixels only so as to avoid explicitly involving boundary pixels, a more widely used practice is image padding, which extends the image by adding new pixels outside the boundary. Commonly adopted image padding strategies include padding with zeros, by repeating boundary values or mean values of a local neighborhood within the boundary, by reflecting the image patches along the boundary, and so on.

Although these methods have been widely accepted - actually some of them have been provided as routine functions in Matlab and OpenCV - people seldom question whether they are the best practice or how they can be improved, maybe because the padding is considered as a trivial operation. The importance of the task, however, is increasing because of the wide adoption of deep convolutional neural networks (CNN) [1]. The convolution runs through both internal and boundary pixels. Losing part of the information or generating inaccurate information at the image boundary may lower the quality of the extracted features, especially when the convolution are applied to low resolution images [2].

In this paper, Innamorati et al. pointed out that the ideal padding operation should extend the image content exactly the same as if the image is taken by a larger sensor. This seems to be an easy task for human being since we can immediately tell whether the padding is natural or not. For automatic padding, however, it is a challenging task because the operation is relied on the context of the image, e.g. semantics or textures, which requires understanding and modelling of the dependency between the orginal image and the expected extension, as in learning based image inpainting [3]. Fortunately, the success of deep learning and large amount of training data provides a potential solution to meet this requirement.

This paper proposes an explicit boundary handling rule for CNNs. The filters are learned as usual at the internal pixels. For border pixels, a set of special filters are learned to replace the normal filters. The contents of the special filters depend on their locations to be applied, i.e., four corners or four edges, for which the part overlapping with the image is learned, and the padded pixels simply follow the common padding methods, e.g., by assigning them to zeros or reflected image values.

Such strategy has several advantages. First, the boundary filters produce exactly the same feature channels as the filters applied to internal image, so no additional runtime cost is incurred. Second, the special filters are seamlessly embedded in the training stage, so both internal and border filters can be jointly trained and optimized. Thirdly, it can be used with most traditional padding methods, simply by changing the values of padded pixels. As expected, this method shows significant advantages in image filtering. Take the learning of a Gaussian filter as an example, this new method generates lower MSE than zero and reflect paddings given different numbers of feature channels and depth values of the network.

This method can be applied to various image processing tasks [4-6]. By changing border handling from traditional padding to explicit boundary rule, the DSSIM for Gauss filtering, the PSNR for de-noising and de-bayering, and the error in pixel distances for disparity and scene flow can be significantly improved. It also show slight drop on the DSSIM for colorization.

This paper has all the merits of a successful publication on image processing: it addresses a common problem from a view point that nobody has thought of; the proposed solution is straightforward in implementation; the method is very effective; and it can be easily integrated with other image processing tasks. It is expected that this method will soon generate impact to the multimedia and computer vision research community and image processing practice.

References:

[1] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In NIPS, 2012.

[2] C. Innamorati, T. Ritschel, T. Weyrich, and N. Mitra. Learning on the Edge: Explicit Boundary Handling in CNNs. In British Machine Vision Conference, 2018.

[3] R. Yeh, C. Chen, T. Lim, M. Hasegawa-Johnson, and M. Do. Semantic Image Inpainting with Perceptual and Contextual Losses. arXiv, 1607.07539, 2016.

[4] A. Dosovitskiy, P. Fischer, E. Ilg, P. Häusser, C. Hazırba V. Golkov, P. Smagt, D. Cremers, and T. Brox. Flownet: Learning optical flow with convolutional networks. IEEE International Conference on Computer Vision, pp. 2758-2766, 2015.

[5] M. Gharbi, G. Chaurasia, S. Paris, and F. Durand. Deep joint demosaicking and denoising. ACM Transactions on Graphics, Volume 35, Issue 63, Article No. 191, 2016.

[6] R. Zhang, P. Isola, and A. Efros. Colorful image colorization. European Conference on Computer Vision, pp. 649-666, 2016.

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.

Secure and Reliable Social Multimedia Delivery in Software-Defined Networks

A short review for "Hybrid Deep-Learning-Based Anomaly Detection Scheme for Suspicious Flow Detection in SDN: A Social Multimedia Perspective"

Edited by Dr. Ye Liu

S. Garg, K. Kaur, N. Kumar and J. J. P. C. Rodrigues, "Hybrid Deep-Learning-Based Anomaly Detection Scheme for Suspicious Flow Detection in SDN: A Social Multimedia Perspective," in IEEE Transactions on Multimedia, vol. 21, no. 3, March 2019.

With the popularity of smartphones and development of social mobile applications, social networks become more and more popular. According to the latest statistics of Global Digital Overview, the number of active social media users in 2019 is more than 3.4 billion [1]. That means nearly half of people in the world are involved into social networks. As a result, social multimedia content grows exponentially, and some of the content are sensitive or private. On the other hand, social networks also become hot beds for cyberattack activities, such as phishing, identity theft and evil twin attacks. Moreover, the huge amount of social multimedia data across Internet challenges the communication network in terms of quality of service (QoS) and energy consumption. Therefore, secure and reliable social multimedia data delivery is a big issue [2, 3].

Compare with traditional network, the decoupled data plane, control plane and application plane in software-defined networks (SDN) make it more promising to improve network security, latency and bandwidth. Many research efforts have been conducted in recent years for anomaly detection based on software-defined networking. However, these approaches face challenges on suspicious flow identification since social multimedia data is heterogeneous and multi-dimensional. In addition, the requirements of QoS and energy consumption have not yet received much attention. Thus, efficient suspicious multimedia flow detection and data delivery is needed.

To address above concerns, a consolidated SDN-based framework is presented in this paper. Specifically, a hybrid deep learning based anomaly detection module is proposed to generate real-time anomaly report and multi-objective flow routing based data delivery module is design to guarantee the quality of end-to-end social multimedia delivery with high energy efficiency. The new anomaly detection module and data delivery module are integrated into the application plane and control plane respectively.

The work process of proposed SDN-based framework is as follow: The SDN controller captures flow statistics after initiated social multimedia request from legitimate user (or malicious user). Then, real-time flow features extraction and type classification are performed in anomaly detection module. After that, anomaly detection module generates detection report and transmits it back to control plane over a secure communication channel. If the request is labeled as anomalous flow, the SDN controller stops the procedure. If the flow is legitimate, the data delivery module figures out an optimize route. Finally, the data are securely delivered to user through the update route.

In anomaly detection module, the hybrid deep learning scheme consists of two main steps: dimensionality reduction and classification. During the first step, Restricted Boltzmann Machine (RBM) is adopted to map the flow vector in low-dimensional space so that the noisy elements could be removed. To reduce the model complexity caused by overfitting, an improved RBM scheme that combines with dropout regularization is designed to random hinder some neural units. With the help of dropout RBM scheme, important flow features can be extracted from training data. During the second step, gradient descent based Support Vector Machines (SVM) is used to classify network flow. These elements are mapped from two-dimensional space to three through weighted mixed kernel trick mechanism for non-linear classification [4].

The decision function is also updated dynamically by gradient descent approach.
In data delivery module, three objective functions are established. The latency objective function consists of propagation delay, transmission delay, queuing delay and processing delay. The bandwidth objective function represents the utilization of flow routes. Energy consumed by fixed components and the dynamic energy consumption of active switch ports are considered to build the energy minimization objective function. Finally, a multi-objective flow routing optimization function (MoFR) is formed for reliable end-to-end multimedia traffic delivery.

Extensive experiments show the proposed hybrid deep learning approach improves the anomaly detection performance in terms of classification error rate, detection rate, accuracy and true positive rate. The MoFR scheme also achieves better quality of service in SDN platform for both latency and bandwidth as well as energy consumption. Furthermore, the robustness of insider threats is also tested with Carnegie Mellon University insider threat dataset [5].

In summary, the proposed framework for suspicious traffic detection based on software-defined networks is demonstrated to achieve high security and transmission reliability in social multimedia applications.

References:

[1] DIGITAL 2019. [Online]. Available: https://datareportal.com/reports/digital-2019-global-digital-overview, accessed on 2019-04-10.

[2] K. Zhang, X. Liang, X. Shen and R. Lu, "Exploiting multimedia services in mobile social networks from security and privacy perspectives," in *IEEE Communications Magazine*, vol. 52, no. 3, pp. 58-65, March 2014.

[3] G. Nan, Z. Mao, M. Yu, M. Li, H. Wang and Y. Zhang, "Stackelberg Game for Bandwidth Allocation in Cloud-Based Wireless Live-Streaming Social Networks," in *IEEE Systems Journal*, vol. 8, no. 1, pp. 256-267, March 2014.

[4] M. Wan, W. Shang and P. Zeng, "Double Behavior Characteristics for One-Class Classification Anomaly Detection in Networked Control Systems," in *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 12, pp. 3011-3023, Dec. 2017.

[5] A. P. Moore, D. M. Cappelli, T. C. Caron, E. Shaw, D. Spooner and R. F. Trzeciak, "A preliminary model of insider theft of intellectual property (No. MU/SEI-2011-TN-013)," Carnegie Mellon University, Pittsburgh, PA, USA, 2011.

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.

Sensor fusion to Sharpen Depth Images in Real-time

A short review for "Image Deblur for 3D Sensing Mobile Devices"

Edited by Dr. Carsten Griwodz

Chung-Hua Chu, "Image Deblur for 3D Sensing Mobile Devices," 2018 IEEE International Conference on Multimedia and Expo (ICME), San Diego, CA, July 2018, pp. 1-6. DOI: 10.1109/ICME.2018.8486596.

While image and video are now pervasive elements of our everyday life and image and video analysis are becoming prevalent in our daily interactions, multimedia research turns increasing to visual content that goes beyond this classical media.

3-dimensional meshes, point clouds and light fields have attracted the attention of multimedia researchers due to the new requirements that they impose on computing, storage and networking infrastructures and the new opportunities that they provide for multimedia applications. As it has been the case with other media in the past, it is not the theoretical basis for the creation and use of 3D data that is investigated by the multimedia community. Instead, multimedia research puts these new media into a context and solves those challenges that prevent its adoption for new and prevalent applications.

The paper "*Image Deblur for 3D Sensing Mobile Devices"* is a good example for this kind of advancement. It starts out with the state-of-the-art in depth image deblurring, and demonstrates the limits of state-of-the-art results as they are developed in the image processing [1] and robotic communities [2]. These approaches achieve remarkable quality in their reconstruction of depth maps from moving cameras by nothing else than the visual information. The *Fast Depth Image Deblur* (FDID) method proposed in the paper "*Image Deblur for 3D Sensing Mobile Devices"*, however, demonstrates much better performance by leaving this artificial limit of pure image input behind.

FDID is meant for Google Tango devices, and even though Project Tango has been discontinued, we have today a multitude of mobile devices that are equipped with hardware that is suitable for implementing the ideas that were enabled by Project Tango. The important change of assumptions that was made for FDID is the availability of sensors that can be used to augment the information that is available for the accurate detection of movement. As such, FDID is an example of a quite frequently applied method in multimedia systems, namely sensor fusion.

Of course, sensor fusion in multimedia has frequently been applied to improve our understanding of situations. True to our community's favorite media, early work combined visual and auditory sensors to achieve a better understanding of a situation [3][4], but a variety of sensors was used by the authors who published in the TOMM Special Issue on Multimedia Sensor Fusion [5]. Also a recent survey [6] supports situational awareness and context recognition as a main application of sensor fusion, but even though work has remained fairly focused on media that resembles video in its (assumed) frame-based and temporal properties, there has been an inclusion of systems that record depth information accurately in this way, such as radar and lidar [7][8].

FDID, however, uses gyroscopes and accelerometers built into the mobile device to extend the input information for deblurring the depths maps recorded by the device. The multimedia community has used these sensors to increase the speed of coding tasks for quite a while, for example in motion vector search for image compression [9]. But those uses were just given classical algorithms a speed advantage by improving their starting points with a probably beneficial direction.

In contrast, FDID builds the entire depth map deblurring algorithm around the assumption of sufficiently reliable data acquisition from gyros and accelerometers. The problem of depth map from blurry RGDB images estimation is reformulated one that requires a search for a solution to one that computes individual pixels from an observed motion directly.

This works by first formulating the resulting depth image in terms of an integration of projected 3D positions that change of time into an image frame. At every point in time, such a 3D position is projected onto the frame, but due to the movement of the mobile device, the position and thus the projection matrix is changing. However, the change of projection matrix is derived from the measured sensor values. The motion during the capture could be erratic, but since sensor sample rates are limited and the likelihood of erratic movement during the 40ms of a frame's capture time isn't very high, the author makes the assumption of linear movement. This provides a very fast and quite accurate estimate for the linear and angular motion between two point positions at the start and end of a frame's recording time. Consequently, the distance and angle allow the creation of a blur kernel for the entire frame.

Now, the ideal depth image I that is compatible with the assumption of linear motion is the image that can be convoluted with the computed blur kernel to regenerate the blurry input frame, although the proposed minimization of the difference image's norm is modified by the second-order difference operator derived from the blur kernels, thus providing a minor advantage to a more smooth solution. Interestingly, the minimal solution can be found by calculating the partial derivative in I and setting it to 0. By resolving the equation to I, the ideal depth image has been found. This provides a direct solution without iterative elements directly from the blurred input image and the sensor information.

In a set of experiments, it is clearly shown that FDID has a major advantage over slower state-of-the-art solutions because it retains an inherent understanding of local smoothness by constructing the blur kernel as well as its derivatives directly from the assumption of smooth linear movement. It does thus always find a reasonable smooth depth image, whereas other state-of-the-art solutions estimate motion in individual pixels.

With this direct approach to depth map recovery, FDID is a highly promising tool for improved depth map creation on mobile devices that sport the required sensor information as well as an RGBD camera. These preconditions are apparently fulfilled by many mobile phones of the current generation, which will make FDID generally applicable.

But for the purpose of the R-Letter, it should serve as a reminder that the multimedia community should explore the option of sensor fusion to address problems head-on. We can today may use of such sensors, and by combining them with well-known computer vision algorithms, we can solve problems efficiently even on smaller devices. To me, this seems like a general challenge that is suitable for the multimedia community, which has always solved practical challenges by combining ideas without sticking to the constraints of a single technique or community.

References:

[1] Hui Ji and Kang Wang, "Robust Image Deblurring With an Inaccurate Blur Kernel," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 1624–1634, Apr. 2012.

[2] S. Tourani, S. Mittal, A. Nagariya, V. Chari, and M. Krishna, "Rolling shutter and motion blur removal for depth cameras," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*, 2016, pp. 5098–5105.

[3] S. Nishiguchi, K. Higashi, Y. Kameda, and M. Minoh, "A sensor-fusion method for detecting a speaking student," in *2003 International Conference on Multimedia and Expo. ICME '03. Proceedings (Cat. No.03TH8698)*, 2003, pp. I–129.

[4] G. Friedland, C. Yeo, and H. Hung, "Dialocalization," *ACM Trans. Multimed. Comput. Commun. Appl.*, vol. 6, no. 4, pp. 1–18, Nov. 2010.

[5] T. Haenselmann, "Foreword to the special issue on multimedia sensor fusion," *ACM Trans. Multimed. Comput. Commun. Appl.*, vol. 6, no. 4, pp. 1–2, Nov. 2010.

[6] R. Gravina, P. Alinia, H. Ghasemzadeh, and G. Fortino, "Multi-sensor fusion in body sensor networks: State-of-the-art and research challenges," *Inf. Fusion*, vol. 35, pp. 68–80, May

2017.

[7] R. O. Chavez-Garcia and O. Aycard, “Multiple Sensor Fusion and Classification for Moving Object Detection and Tracking,” *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 2, pp. 525–534, Feb. 2016.

[8] D. Xu, D. Anguelov, and A. Jain, “PointFusion: Deep Sensor Fusion for 3D Bounding Box Estimation,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 244–253.

[9] X. Chen, Z. Zhao, A. Rahmati, Y. Wang, and L. Zhong, “SaVE,” in *Proceedings of the seventeen ACM international conference on Multimedia - MM ’09*, 2009, p. 381.

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

An Incremental Personalized Classifier for Food Classification

A short review for "Personalized Classifier for Food Image Recognition"

Edited by Dr. Bruno Macchiavello

S. Horiguchi, S. Amano, M. Ogawa and K. Aizawa, "Personalized Classifier for Food Image Recognition," in IEEE Transactions on Multimedia, vol. 20, no. 10, pp. 2836-2848, Oct. 2018.

Recently, the use of convolutional neural netwroks (CNNs) for image classification has been the focus of several works. Most methods are based on fixed-class training, where the training step is performed only once using a large dataset and a fixed number of classes [1]-[3]. However, in a real-word scenario there could be cases in which the number of samples in a class continues to increase and/or samples of a new class appear. Therefore, for such cases a system that can incrementally continue to lean after its first training is more adequate. Personalizing a classifier incrementally for each user is a promising way achieve this goal. In this work, the authors present a simple personalization framework, which is a combination of the nearest class mean classifier and the 1-nearest neighbor classifier based on deep features.

In addition to incremental learning, the authors state that personalization of classifiers should also fulfill two other aspects. One is domain adaptation. Domain adaptation is needed when there are variations in the content of classes between source and target domains, there are many previous proposed methods for domain adaption [4]-[5], but most of them cannot learn incrementally. The class definitions are different; thus, it is important to assume that each personal data item is from a different domain. The other is one-shot learning. One-shot learning aims to learn new classes using only a few samples. When focusing on a given person, the number of images is limited; therefore one-shot learning becomes a requirement.

The case study chosen by the authors is food image recognition. There are many studies in food image recognition [6]–[9]. Nevertheless, almost all of the previous studies followed the general method of using fixed food image datasets, which may not appropriate for a real-world purpose. In order to conduct the experiments the authors prepared a large-scale labelled image dataset that contains owner IDs and time stamps. When labeling an image region, the user was able to select from a list of candidates that the system produced. When the user did not find any appropriate entry in them, then the user created a new label by selecting from default label set or entering a user-specific label. The database was created over two years and 1,508,171 food images were collected from 20,820 users.

The proposed personalization framework is referred as sequential personalized classifier (SPC). The proposed system is actually an hydrid classifier that combines an initial fixed-class training (using CNNs) with common nearest class mean classifiers for each user that will perform incremental learning. The training setup of the original CNN was done empirically, the authors used various sizes of datasets. For incremental learning each user has his/her own database. This database starts empty. As time passes by, the user adds samples to his database. When the user records the i-th dish (x_i), its class is predicted by a combination of a common feature vector and a user's own feature vector. The common feature vector is shared between all users, and it is basically a mean of all the previous individual vectors. The similarities between x_i and both vectors are computed. The similarities are combined using a weighting factor, basically the similarity of the users own feature vector is multiplied by the weighting factor and then the maximum value between both similarities is considered as the combined value. The weighting factor basically controls the degree of presonalization, that is, the balance between the common mean and the users vector. When an entered vector is far from every initial vector, its visual concept is novel. Thus, weighting the user's vector more heavily is effective because the CNN is trained without images of such a novel visual concept. When an entered vector is a new sample for existing classes, weighting is also effective because the new vector covers similar initial vectors. After the combination of the similarities the system predicts the class. If the class is correct the sample is added to that class, if it is wrong the user corrects the predicted result, which can create a new class.

During experiments the authors show that SPC outperforms conventional methods in terms of accuracy. The experimental setup includes a set of images that will be added incrementally. The authors show the average accuracy result of several consecutive images added at each time step. The authors also provided results varying the weighting factor and conclude that 0.85 was the best value. However, an actual optimization method for obtaining the weighting factor is not provided. As mentioned by the authors, an important assumption of SPC is that users' records should not be not very noisy. If a user's records are randomly labeled, SPC does not work. Also, SPC has still limitations. It cannot cope with the changes within a user, because it does not have any mechanism to forget old records. Also, complexity may became an issue in the proposed architecture, which may generate slow responses.

In conclusion, this paper introduced a personalization problem in food image recognition and proposed a method for classifier personalization. Personalized food recognition contains problems of incremental learning, domain adaptation, and one-shot learning. The proposed method outperforms existing ones in terms of accuracy. The authors stated that they plan to accelerate this work on personalization by using information from other users whose labeling tendencies are similar to that of the target user.

References:

[1] W. J. Scheirer, A. Rocha, A. Sapkota, and T. E. Boult, "Toward open set recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 7, pp. 1757–1772, Jul. 2013.

[2] A. Bendale and T. Boult, "Towards open world recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1893–1902.

[3] M. Ristin, M. Guillaumin, J. Gall, and L. Van Gool, "Incremental learning of random forests for large-scale image classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 3, pp. 490–503, Mar. 2016.

[4] J. Hoffman, E. Rodner, J. Donahue, K. Saenko, and T. Darrell, "Efficient learning of domain-invariant image representations," in Proc. Int. Conf. Learn. Represent., 2013.

[5] M. Kan, J. Wu, S. Shan, and X. Chen, "Domain adaptation for face recognition: Targetize source domain bridged by common subspace," Int. J. Comput. Vis., vol. 109, no. 1, pp. 94–109, 2014

[6] M. Chen et al., "PFID: Pittsburgh fast-food image dataset," in Proc. 16th IEEE Int. Conf. Image Process., 2009, pp. 289–292.

[7] F. Zhu et al., "The use of mobile devices in aiding dietary assessment and evaluation," IEEE J. Sel. Topics Signal Process., vol. 4, no. 4, pp. 756–766, Aug. 2010.

[8] Y. Matsuda, H. Hoashi, and K. Yanai, "Recognition of multiple-food images by detecting candidate regions," in Proc. IEEE Int. Conf. Multimedia Expo, 2012, pp. 25–30.

[9] H. Kagaya, K. Aizawa, and M. Ogawa, "Food detection and recognition using convolutional neural network," in Proc. 22nd ACM Int. Conf. Multimedia, 2014, pp. 1085–1088.

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.

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.