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

MMTC Communications – Review

Vol. 7, No. 4, August 2016

IEEE COMMUNICATIONS SOCIETY

TABLE OF CONTENTS

Message from the Review Board Directors

Welcome to the August 2016 issue of the IEEE ComSoc MMTC **Communications – Review** including a special section comprising **reviews for the best papers awarded at IEEE ICME 2016** in Seattle Washington, USA (July 11-15, 2016, http://www.icme2016.org/) preceded by an guest editorial by ICME'16 general chair ***Cha Zhang*** from Microsoft Research. Additionally, this issue comes with **four regular reviews** in the area of *voice conversation*, *vehicle re-identification*, *aging face recognition*, and *video characteristics for routing and scalable transmission*.

We hope that this issue **stimulates your research in the area of multimedia communication**.

An overview of all reviews is provided in the following:

The **first paper**, published in the *Proceedings of IEEE ICME'16 as best paper* and *edited by **N.N.***, describes means for phonetic posteriorgrams for m-to-1 voice conversion without parallel data training.

The **second paper**, published in the *Proceedings of IEEE ICME'16 as best student paper* and *edited by **N.N.***, describes an approach for large-scale vehicle re-identification in urban surveillance videos.

The **third paper** is *edited by **Bruno Macchiavello*** and has been published within the *IEEE Transactions on Image Processing*. It proposes a feature-based learning model for aging face recognition.

The **forth paper**, published in *IEEE Multimedia* and *edited by **Frank Hartung***, helps to develop a proper understanding of the properties of the data that occupies the Internet.

The **fifth paper**, published in *IEEE Transactions on Multimedia* and *edited by **Roger Zimmermann***, proposes means for video traffic routing between datacenters with a software defined network architecture.

Finally, the **sixth paper** is *edited by **Koichi Adachi*** and has been published in *IEEE Wireless Communication Letter*. It describes caching methods for scalable video transmission.

We would like to thank all the authors, nominators, reviewers, editors, and others who contribute to the release of this issue.

Finally, we would like to highlight upcoming conferences in 2016 which are related to MMTC:

- **ACM Multimedia**, October 15-19, Amsterdam, The Netherlands: http://www.acmmm.org/2016/
- **IEEE GLOBECOM**, December 4-8, Washington, DC, USA: http://globecom2016.ieee-globecom.org/

IEEE ComSoc MMTC Communications – Review

Director: Christian Timmerer
Alpen-Adria-Universität Klagenfurt, Austria
Email: christian.timmerer@itec.aau.at

Co-Director: Yan Zhang
Simula Research Laboratory, Norway
Email: yanzhang@simula.no

IEEE ICME 2016 Bester Paper Awards

*Guest Editorial Introduction by **Cha Zhang** (IEEE ICME'16 General Co-Chair)*

The 2016 IEEE International Conference on Multimedia and Expo (ICME 2016) was held in Seattle, Washington, USA from July 11 to July 15, 2016. The conference received a total of 512 submissions, among which 152 papers were accepted (30%). After a rigorous evaluation process, the Technical Program Chairs and the ICME 2016 Paper Award Committee selected three papers as best paper award recipients. They are:

Best Paper Award:
Phonetic Posteriorgrams for Many-to-One Voice Conversion without Parallel Data Training, by *Lifa Sun, Kun Li, Hao Wang, Shiyin Kang and Helen Meng*.

According to the Award Committee, this paper is a clear winner for the best paper award. It proposed a novel approach to voice conversion (converting the speech of one speaker to another speaker's voice) using unpaired source speaker and target speaker training data. The algorithm bridges between speakers by means of Phonetic PosteriorGrams (PPGs), which is obtained from a speaker-independent automatic speech recognition system. The idea is novel and can be very flexible for many applications such as personalized speaking aids, movie dubbing, etc.

Best Student Paper Award:
Large-Scale Vehicle Re-Identification in Urban Surveillance Videos, by *Xinchen Liu, Wu Liu, Huadong Ma and Huiyuan Fu*.

This paper examines the problem of vehicle re-identification (Re-Id) across large number of traffic cameras. One of the major contributions is a new large-scale benchmark dataset for vehicle Re-Id. It contains over 40,000 bounding boxes of 619 vehicles captured by 20 cameras in unconstrained traffic scene. Each vehicle is captured by 2-18 cameras in different viewpoints, illuminations, and resolutions to provide high recurrence rate. The paper also presented a vehicle Re-Id method which combines the texture, colors, and high-level attributes information as a baseline on the dataset.

Blind Quality Assessment of Compressed Images via Pseudo Structural Similarity, by Xiongkuo Min, Guangtao Zhai, Ke Gu, Yuming Fang, Xiaokang Yang, Xiaolin Wu, Jiantao Zhou and Xianming Liu.

This paper explores the pseudo structures when images are compressed using block-based methods. It presented an interesting way to evaluate the quality of compressed images via the similarity between pseudo structures of two images. The proposed pseudo structural similarity (PSS) model works well not only on natural scene images, but also on screen content images.

Paper quality at ICME has been improving steadily over the past few years, thanks to a revised charter in 2010 that set paper page limit to 6 and paper acceptance rate to no more than 30%. Many other exciting works were presented at this year's ICME, and it is impossible to list all of them here. Please check out the conference proceedings when you have time, and be sure to submit your future work to ICME!

Cha Zhang is currently a Principal Researcher in the Multimedia, Interaction and eXperience Group at Microsoft Research. He received the B.S. and M.S. degrees from Tsinghua University, Beijing, China in 1998 and 2000, respectively, both in Electronic Engineering, and the Ph.D. degree in Electrical and Computer Engineering from Carnegie Mellon University, in 2004. His current research focuses on applying various audio/image/video processing and machine learning techniques to multimedia applications. Dr. Zhang has published more than 80 technical papers and holds 20+ U.S. patents. He won the best paper award at ICME 2007, the top 10% award at MMSP 2009, and the best student paper award at ICME 2010. He currently serves as an Associate Editor for IEEE Trans. on Circuits and Systems for Video Technology, and IEEE Trans. on Multimedia. He is a Senior Member of the IEEE.

IEEE ICME'16 Best Paper: Phonetic Posteriorgrams for Many-to-One Voice Conversion without Parallel Data Training

*A short review for "Phonetic Posteriorgrams for Many-to-One Voice Conversion without Parallel Data Training" (Edited by **Christian Timmerer**)*

Lifa Sun, Kun Li, Hao Wang, Shiyin Kang and Helen Meng, "Phonetic Posteriorgrams for Many-to-One Voice Conversion without Parallel Data Training", Proceedings of the IEEE International Conference on Multimedia and Expo (ICME'16), Seattle, WA, USA,

The authors of this paper target the application of voice conversion (VC) which "aims to modify the speech of one speaker to make it sound as if it were spoken by another specific speaker". In this context, authors developed an approach – inspired by [1] – using Phonetic PosteriorGrams (PPGs) which are obtained from a speaker-independent automatic speech recognition (SI-ASR) system. Therefore, authors obtain PPGs of the target speech. The relationships between the PPGs and acoustic features of the target speech are modelled using a Deep Bidirectional Long Short-Term Memory based Recurrent Neural Network (DBLSTM) structure. Finally, arbitrary source speech can be converted by obtaining its PPGs from the same SI-ASR and provide them to the trained DBLSTM for generating the converted speech.

The advantages of the proposed approach are as follows: (*i*) no parallel training data is required and (*ii*) a trained model can be applied to any other source speaker for a fixed target speaker allowing for many-to-one conversions.

The proposed approach comprises three stages, i.e., *training stage 1*, *training stage 2*, and the *conversion stage*:

- The training stage 1 extracts parameters from a standard ASR corpus to obtain the PPGs representation of the target speech.
- The training stage 2 models the relationships between the PPGs and mel-cepstral coefficients (MCEPs) features of the target speaker for speech parameter generation.
- The conversion stage uses the PPGs from the source speech (using the same model as for the target speech) as an input to the trained DBLSTM model in order to generate the converted speech.

The authors use the CMU ARCTIC corpus [2] for the evaluation of the proposed PPGs system and compare it with a baseline approach. The baseline approach is a DBLSTM-based approach with parallel training data. The evaluation is split into an objective and subjective evaluation. The former adopts the mel-cepstral distortion (MCD) which is used to measure how close the converted speech is to the target speech. The latter authors use the Mean Opinion Score (MOS) test and an ABX preference test for measuring the naturalness and speaker similarity of converted speech.

The results indicate that both baseline and proposed approach have similar performance in terms of the objective measure. On the other hand, the results of the subjective tests show that the proposed PPGs-based approach performs better than the baseline approach with respect to both speech quality and speaker similarity.

Finally, authors have looked into applying their proposed model for cross-lingual voice conversion and have obtained interesting preliminary results. However, further investigations on cross-lingual applications will be conducted as part of future work.

References

[1] S. Aryal and R. Gutierrez-Osuna, "Articulatory-based conversion of foreign accents with Deep Neural Net- works," in Proc. Interspeech, 2015.

[2] J. Kominek and A. W. Black, "The CMU Arctic speech databases," in *Fifth ISCA Workshop on Speech Synthesis*, 2004.

IEEE ICME'16 Best Student Paper: Large-Scale Vehicle Re-Identification in Urban Surveillance Videos

A short review for "Large-Scale Vehicle Re-Identification in Urban Surveillance Videos"
*(Edited by **Christian Timmerer**)*

Xinchen Liu, Wu Liu, Huadong Ma and Huiyuan Fu, "Large-Scale Vehicle Re-Identification in Urban Surveillance Videos", Proceedings of the IEEE International Conference on Multimedia and Expo (ICME'16), Seattle, WA, USA, July, 2016.

Vehicles are an important object class in (urban) surveillance applications and raise many research issues in computer vision such as detection, tracking, and classification. An interesting aspect is the re-identification (Re-Id) which has been extensively studied for persons [1][2] but not for vehicles. Therefore, this paper proposes a large-scale benchmark dataset for vehicle Re-Id in a real-world urban surveillance scenario which is referred to as "VeRi". Additionally, authors evaluated six vehicle Re-Id methods on VeRi and proposed a baseline algorithm which combines color, texture, and high-level semantic information extracted by a deep neural network.

The authors define vehicle Re-Id based on a given probe vehicle image and the ability to search in a database for images that contain the same vehicle captured by possibly multiple cameras. Furthermore, vehicle Re-Id in urban surveillance video can be found as a near duplicate image retrieval (NDIR) problem [3] which is different from vehicle detection, tracking, or classification.

The characteristics of the dataset can be described as follows:

- It contains more than 40,000 images of 619 vehicles which have been captured by 20 cameras covering an area of 1km^2 area within 24 hours. This makes the dataset scalable enough for vehicle Re-Id but also other related research.
- All images are captured in a real-world unconstrained surveillance environment and labeled with varied attributes such as BBoxes, types, colors, and brands. Therefore, complicate models can be learnt and evaluated for vehicle Re-Id.
- Finally, each vehicle is captured by a minimum of two and by up to 18 cameras in different viewpoints, illuminations, resolutions, and occlusions. This provides high recurrence rate for vehicle Re-Id in practical surveillance environments.

The dataset has been used to evaluate six competing vehicle Re-Id methods which can be clustered based on color, texture, and semantic features:

- Texture based feature (BOW-SIFT) [4]
- Color based feature (BOW-CN) [2]
- Semantic feature extracted by deep neural network: AlexNet [5], and GoogleLeNet [6]
- Feature fusion: AlexNet + BOW-CN and the method proposed by authors based on the fusion of Attributes and Color feaTures (FACT).

The evaluation is done by mean average precision (mAP), HIT@1, and HIT@5 which reveals that fusion-based methods provide superior results compared to others. However, this dataset and this evaluation are only the baseline for future work with hopefully more complicate models leading to better results.

References

[1] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in CVPR, 2014, pp. 152–159.
[2] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jing- dong Wang, Jiahao Bu, and Qi Tian, "Scalable person re- identification: A benchmark," in ICCV, 2015.
[3] Tao Mei, Yong Rui, Shipeng Li, and Qi Tian, "Multimedia search reranking: A literature survey," *ACM CSUR*, vol. 46, no. 3, pp. 38, 2014.
[4] Liang Zheng, Shengjin Wang, Wengang Zhou, and Qi Tian, "Bayes merging of multiple vocabularies for scalable image retrieval," in *CVPR*, 2014, pp. 1963–1970.
[5] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Im- agenet classification with deep convolutional neural networks," in *NIPS*, 2012, pp. 1097–1105.
[6] Linjie Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang, "A large-scale car dataset for fine-grained categorization and verification," in CVPR, 2015, pp. 3973–3981.

Feature-based Learning Model for Aging Face Recognition

*A short review for: "Aging Face Recognition: A Hierarchical Learning Model Based on Local Patterns Selection" (Edited by **Bruno Macchiavello**)*

Z. Li, D. Gong, X. Li and D. Tao, "Aging Face Recognition: A Hierarchical Learning Model Based on Local Patterns Selection," in IEEE Transactions on Image Processing, vol. 25, no. 5, pp. 2146-2154, May 2016.

Humans are able to obtain a wide variety of information from a face image, including identity, age, gender, ethnicity, etc. [1]. Automated systems often need to emulate the human ability of retrieving such information from a digital face image. Specifically, automatic identification based on face recognition is a very important problem [2-5]. However, face recognition is still a challenging task due to different factors, like aging. Aging face recognition is an emerging research topic, that has several useful applications, e.g. finding missing children and identifying criminals based on photographs.

Age related face image analysis has only been studied in recent years. Several works focus on age estimation, which is the determination of a person's age based on biometric features [1,6], while other focus on aging simulation [7-8]. Aging simulation is often used for face recognition, where the input image is modified to the same age as the gallery image, prior to recognition. However, this method has a high computational load and therefore feature-based methods can be desirable in certain situations.

In this paper, the authors propose a novel two-level hierarchical learning model to address this problem. In the proposed model, initially features are extracted from low-level pixel structures. Low-level information is widely believed to be very beneficial to cross-age face recognition. At the second level, higher-level visual information is refined by the learning algorithm.

The proposed low-level pixel structure used in this work is referred as Local Pattern Selection (LPS). Several steps are follow in order to create the LPS sets during training. Two different images, at different ages, are required per subject in the training database. First, each pixel in both face images is associated with a corresponding pixel feature, which is formed by sampling its eight neighbors at a certain radius *r*. Then, each pair of pixel features, called matching pixel feature, is created by associating the pixel features of the two different images of the same subject extracted at the same spatial location. In order to do so, the input images are cropped to the same size with faces aligned to ensure good correspondence. However, the authors stated that strict pixel-level correspondence is not required due to the large amount of pixels that are used. Once, the matching pixel features are obtained, each pair goes through a encoding tree which produces the final LPS code. The encoding tree consists in a binary decision three. The internal nodes are associated with an attribute and a threshold. The attribute refers to one element in a pixel feature, if that element is lower than the threshold then this pixel feature will be directed to the left branch (right branch otherwise). The leafs in the encoding tree are associated with distinct decimal codes.

In order to evaluate the decimal code that is generated, the authors proposed utility measure. This measure is composed of two different terms, the first one indicates common information between cross-age faces. That is, a large number indicates that a high fraction of the matching pixel features pairs have the same code. The second term, in the utility measure formulation, serves as regularization. This regularization encourages even partitioning of the called pixel feature space. A pixel feature space is a region on the image where the different pixel features generate the same decimal code. The higher value of this term means higher entropy, and thus evener distribution.

The main idea of the LPS algorithm is to construct local encoding threes of overlapping small patches of each individual image that maximize the utility measure. This is done using a greedy algorithm. The encoding three grows incrementally, during training, until the expected number of leaf nodes is reached. At each step, the best node that maximizes the increase in the utility measure is selected. The number of expected leaf nodes in the tree is determined by cross validation. This process is repeated for different values of *r* during pixel feature creation and for different patches sizes. The final feature vector of an image is formed by concatenating the local features at all sampling radii.

The features extracted based in LPS are usually of very high dimension due to the employment of both multiple scaling and dense sampling techniques. This extremely high dimension of the facial features is not good for both storage and face matching. Therefore,

as a high-level learning step a bagging algorithm is defined. The bagging repeatedly selects a series of training subsets of samples and based on these subsets it trains a series of classifiers, which will be fused into a unified classifier.

In order to verify the proposed method, the authors first analyze each level of the proposed technique separately. For the low-level structure it was verified that LPS can achieve a more uniform distribution than other low-level techniques. While the high-level learning model is compared to a previously proposed technique [9]. Finally, the overall benchmark comparison is done using the MORPH database. The LPS technique was able to outperformed several previously proposed methods, achieving a recognition accuracy of 92.11%. Moreover, the proposed technique was combined with others in order to increase the recognition accuracy rate up to 94.87%.

While, this work certainly indicates that feature-based techniques can be used for aging face recognition, and that high accuracy rates can be obtained using low-level pixel structures. The proposed method uses a large number of images (10,000) at different ages for training, which may not always be available. And, it requires that the face images at different ages should have been acquire at similar conditions (resolution, illumination, etc.). Therefore, future works can focus on improving the generalization capacity of this proposed learning method.

References

[1] H. Han, C. Otto and A. K. Jain, "Age estimation from face images: Human vs. machine performance," 2013 International Conference on Biometrics (ICB), Madrid, 2013, pp. 1-8.

[2] L. Wolf, T. Hassner, and Y. Taigman, "Descriptor based methods in the wild," in Proc. Workshop Faces Real-Life Images ECCV, 2008, pp. 1–14.

[3] K. Simonyan, O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Fisher vector faces in the wild," in Proc. BMVC, 2013, pp. 1–12.

[4] L. Shao, D. Wu, and X. Li, "Learning deep and wide: A spectral method for learning deep networks," IEEE Trans. Neural Netw. Learn. Syst., vol. 25, no. 12, pp. 2303–2308, Dec. 2014.

[5] Y. Li, L. Meng, J. Feng, and J. Wu, "Downsampling sparse representation and discriminant information aided occluded face recognition," Sci. China Inf. Sci., vol. 57, no. 3, pp. 1–8, 2014.

[6] Y. Fu and T. S. Huang, "Human age estimation with regression on discriminative aging manifold," IEEE Trans. Multimedia, vol. 10, no. 4, pp. 578–584, Jun. 2008.

[7] J. Suo, S.-C. Zhu, S. Shan, and X. Chen, "A compositional and dynamic model for face aging," IEEE Trans. Pattern Anal. Mach. Intell., vol. 32, no. 3, pp. 385–401, Mar. 2010.

[8] U. Park, Y. Tong, and A. K. Jain, "Age-invariant face recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol. 32, no. 5, pp. 947–954, May 2010.

[9] B. Klare, Z. Li, and A. K. Jain, "Matching forensic sketches to mug shot photos," IEEE Trans. Pattern Anal. Mach. Intell., vol. 33, no. 3, pp. 639–646, Mar. 2011.

Bruno Macchiavello is an assistant 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. He also was co-organizer of a special session on Streaming of 3D content in the 19th International Packet Video Workshop (PV2012). His main research interests include video and image coding, image segmentation, distributed video and source coding, multi-view and 3D video processing.

Understanding the Properties of the Data that Occupies the Internet

*A short review for: "A Survey of Current YouTube Video Characteristics" (Edited by **Frank Hartung**)*

Xianhui Che, Barry Ip, and Ling Lin, "A Survey of Current YouTube Video Characteristics", IEEE MultiMedia, 22.2, 2015, pp. 56-63

It is a well-known fact that video has become the dominant content type that uses most resources of the Internet. According to the latest Cisco Visual Networking Index [1], 70 percent of all Internet traffic is video. This share will further increase, and for 2020 it is expected that even 82 percent of all consumer Internet traffic will be video. Among video services, YouTube is the service generating the highest amount of traffic. Thus, for video compression and video networking engineers, it is interesting to know how exactly the proprietary YouTube service works, and what the characteristics of the video traffic it generates are.

The authors of the publication under discussion have analyzed YouTube video traffic. Although similar analyses have been done before, for example around 2007/2008 [2][3][4], it is necessary to re-do them, because the technology used by services such as YouTube, and hence the video characteristics, change. For example, H.264 video compression is being replaced by HEVC [5]. Also, video services introduce improvements in their compression mechanisms [6] and their transport mechanisms [7]. However, it should be noted that the authors have done their analysis of YouTube video traffic in the first half of 2013. All following statements refer to YouTube characteristics 2013.

The authors developed and deployed a web crawler that, starting from the frontpage with most popular YouTube content, retrieved metadata and the first 2 Kbytes of videos, and then iteratively followed the links to the top 20 related popular content items that were presented on the page of the YouTube video. The crawler changed IP addresses frequently. YouTube tries to prevent harvesting URLs by use of JavaScript programs, and the authors had to do programming to get access to the URLs anyway. About 1.2 million videos data sets (not the whole videos) were downloaded, stored, and later analyzed.

In the main part of the paper, the authors describe the insights they have gained from the analysis of this data set.

YouTube videos are categorized. Dominant categories in 2013 were music (23 percent), entertainment (16 percent), gaming (9 percent), and people/blogs (8 percent).

Regarding the duration of YouTube videos, the authors present distribution plots. It is eminent that some "typical" video durations occur often. Music videos are often 200 to 240 seconds, a typical duration of music pieces. Entertaining videos are often shorter, peaking at 60-100 seconds. Gaming videos are typically long, peaking at the maximum allowed duration (previously 600 seconds, later increased to 900 seconds and even relaxed for authorized customers). Still, the vast majority of videos are below 300 seconds.

Video resolution has increased during the last years. Almost all types of cameras can capture HD resolution, and even UHD/4K cameras are more and more available. Still, YouTube hosts a lot of "old" content, formerly captured in low resolution. In average, each YouTube video (of 2013) is available in 3.4 different resolution variations. All videos provide at least a resolution of 320 x 240 pixels. More than 70 percent are available in SD resolution of 640 x 360 pixels, but only 14 percent in 1280 x 720 pixels HD, and as little as 3 percent in 1920 x 1080 pixels HD.

File size distribution relates to duration and resolution. The average file size is 17.6 Mbytes for a resolution of 640 x 360 and 6.5 Mbytes for a resolution of 320 x 240. However, the distribution is wide spread with a long tail.

The authors have determined that YouTube video is encoded at constant bit-rate, which is known to be preferable for streaming or streaming-like transmission. However, the results indicate that the rate at least for higher resolutions is not fixed; rather, different videos have different rates. This speaks for content-dependent rate control in the compression. Content at low resolution like 320 x 240 pixels often has a rate of around 320 Kbps. Audio data rates are also distributed, but peak at 64 and 128 Kbps.

Summarizing, we can say that the publication gives important practical information about typical characteristics of YouTube video. This should be

used by other researchers, for example when selecting parameters for simulations. For example, it is wise to assume a duration distribution of video that resembles real video traffic.

A drawback is certainly that the data is already, shortly after its publication, slightly outdated. Since 2013, cameras have evolved; HD resolution is nowadays the normal resolution for video capture, and UHD/4K is about to take over soon. Also, recent developments like the introduction of HEVC and DASH based streaming, both important developments with major impact, have not yet been considered. Also, some other characteristics are missing in the paper. Video compression researchers would have liked to get more information about used compression characteristics beyond bit rate, like codec profiles and parameters, GOP sizes, etc. For DASH transmission, used segment durations and typical quality levels would be interesting to know.

The topic is to be continued, and more recent characteristics of Internet video traffic are still of further interest.

References

[1] Cisco Visual Networking Index: Forecast and Methodology, 2015–2020, White Paper
[2] X. Cheng, C. Dale, and J. Liu, "Understanding the Characteristics of Internet Short Video Sharing: YouTube as a Case Study," Cornell Univ., 2007; http://arxiv.org/abs/0707.3670.
[3] P. Gill et al., "YouTube Traffic Characterization: A View from the Edge," Proc. 7th ACM Conf. Internet Measurement, 2007, pp. 15–28
[4] X. Cheng, C. Dale, and J. Liu, "Statistics and Social Network of YouTube Videos," Proceedings 16th International Workshop on Quality of Service, 2008, pp. 229–238
[5] Mathias Wien, *High Efficiency Video Coding*, Springer-Verlag Berlin, 2015
[6] Zhi Li, Anne Aaron, Ioannis Katsavounidis, Anush Moorthy and Megha Manohara , "Toward A Practical Perceptual Video Quality Metric", http://techblog.netflix.com/, June 2016
[7] "Delivering Live YouTube Content via DASH", https://developers.google.com/youtube/v3/live/guides/encoding-with-dash, May 2016

Frank Hartung is a full professor of multimedia technology at FH Aachen University of Applied Sciences, Aachen, Germany. He received a MSc in electrical engineering from RWTH Aachen University, Germany, and a PhD in Telecommunications from University of Erlangen, Germany. He has been working with Ericsson Research, as a research team leader in Multimedia Technologies, from 1999 to 2011. In 2008, he was a visiting researcher at Stanford University, Palo Alto, USA, and in 2016, he was a visiting researcher at Eurecom, Sophia-Antipolis, France. His research interests include media security and forensics, networked multimedia, immersive multimedia communication, streaming, and mobile video. He has authored or co-authored more than 50 publications in this domain, and is the co-inventor of 22 granted patents. Dr. Hartung is a member of IEEE and VDE.

Video Traffic Routing between Datacenters with a Software Defined Network Architecture

A review for "Delay-Optimized Video Traffic Routing in Software-Defined Interdatacenter Networks"
*(Edited by **Roger Zimmermann**)*

Yinan Liu, Di Niu, and Baochun Li, "Delay-Optimized Video Traffic Routing in Software-Defined Interdatacenter Networks", IEEE Transactions on Multimedia, vol. 18, no.5, pp. 865–878, May 2015.

The feasibility of video streaming and distribution via the Internet has led to the emergence of a number of very large video sites such as Netflix, Hulu, etc., which transmit content world-wide. Rather than installing their own infrastructure, some of these companies rely on large-scale, distributed datacenters, such as Amazon's EC2 cloud, to reach their many customers. With a global user base, a significant amount of video traffic occurs between datacenters (e.g., for moving content closer to a group of users). In addition to liberating a streaming company from tasks such as hardware maintenance, another benefit is that these globally distributed datacenters are generally very well connected among each other and the capacity is often over-provisioned. Rather than focusing on throughput or link utilization, in this manuscript the authors present a framework to optimize the routing of inter-datacenter network flows assuming that some flows are very delay-sensitive (e.g., videos) while other traffic may not have very stringent latency requirements (for example data backups or database maintenance traffic). As one of the differentiations to earlier work, the authors' approach utilizes a Software Defined Networking (SDN) architecture to compute a globally optimal solution.

The authors motivate the problem by observing that solutions currently used in production, such as Multiprotocol Label Switching Traffic Engineering (MPLS TE) [1], do not compute globally optimal flow allocations and also are rather static in that once a flow route has been computed, it will not be changed, even if the dynamic situation evolves and, for example, a highly delay-sensitive flow enters the system and should be given priority on a low delay route. In order to compute a globally optimal solution, the authors use an application-level SDN architecture which separates the control-plane and the forwarding-plane that have traditionally co-existed in network hardware. In an SDN architecture, the control-plane is managed by a *controller* that is separate from the packet-level *forwarders*, with a standardized control protocol between the two planes [2] (very often the OpenFlow [3] protocol is used for this purpose).

The authors consider an *application-level* SDN architecture which implies that both the controller and the forwarders are implemented at the application level in virtual machines. This seems reasonable as not all the networking switches currently in use are SDN-capable and furthermore if datacenters are globally distributed then the latencies between them, with links that are thousands of miles long, are realistically in the hundreds of milliseconds and the software implementation of the forwarders should have a negligible impact.

The authors focus on two main criteria in their flow optimization framework: (a) path sparsity and (b) low end-to-end delay. In the first criteria, the aim is not to split a flow over multiple paths as much as possible. The reason for this is that the re-merging of partial flows generates additional overhead and delay in packet sequencing and should be avoided. The second criteria, achieving a low latency, is the other primary goal of the optimization.

The latency optimization is posed via three different functions. The objective of the first is to minimize the weighted sum of session delays by controlling the flow rate assigned to each distribution tree (if it is a multicast transmission) or path (unicast) within each session. A Boolean indicator variable describes whether a path will be chosen or not and a penalty function is designed to enforce a sparse tree selection as the preferred solution. In the second optimization function the worst-case latency of each session is replaced by the sum of latencies of all chosen trees. Finally, in the third formulation the measuring of latencies is completely avoided by substituting the sum of all distinct tree depth values instead. The solution in this case prefers trees with low depths, which are assumed to be equivalent to trees with a short latency. Because the three optimization problems are all non-convex, the authors solve them with a Log-det heuristic. In essence they replace the 0-1 valued indicator functions by a smooth log function and minimize the linearization of this log function iteratively. The authors present a proof that

this solution does indeed converge. The authors assume that the network capacity is over-provisioned to accommodate all flows. If this is not the case, they schedule delay-sensitive flows first, and then allocate the background flows. If the capacity is not enough for all the delay-sensitive flows, then there is an optimization step for the bandwidth allocation.

The proposed design of the three optimization functions is evaluated with an application-level SDN implementation where the forwarders are executed in six different Amazon EC2 datacenters in different world regions (Oregon, Virginia, Ireland, Tokyo, Singapore and Sao Paulo) with the SDN controller also being located in Oregon. The authors first perform measurement to show that as assumed the inter-datacenter bandwidth is high and quite stable. In addition to the authors' three methods, the evaluation comparison techniques include shortest path routing (specifically the CSPF algorithm, which is commonly adopted in MPLS TE today) and multi-commodity flow (MCF) as used in the SWAN algorithm [4]. Among the results that the authors present they show that, indeed, the three optimization functions perform very similarly in trend, and hence that the tree depth provides a good approximation of the path latency. As the main results, compared with the shortest path and the MCF methods, the authors' techniques achieve their goal of reducing the latency of the delay-sensitive flows while providing still adequate performance for the background flows.

In summary, the proposed framework enables lower latencies for latency constraint network flows, such as live videos. The system is software-implementable in an SDN-type architecture, but at the application level. Hence it should be practical to deploy with currently existing networks and datacenters.

References

[1] D. O. Awduche and J. Agogbua, "Requirements for Traffic Engineering Over MPLS," RFC 2702, September 1999.

[2] Open Networking Foundation (ONF), Palo Alto, CA, USA, "Software Defined Networking: The New Norm for Networks," 2012.

[3] N. McKeown, T. Anderson, H. Balakrishnan, G. Parulkar, L. Peterson, J. Rexford, S. Shenker, and J. Turner, "OpenFlow: Enabling Innovation in Campus Networks," *SIGCOMM Computer Communication Review*, vol. 38, no. 2, pp. 69–74, March 2008.

[4] C.-Y. Hong et al., "Achieving high utilization with software-driven WAN," in Proceedings of ACM SIGCOMM, 2013, pp. 15–26.

Roger Zimmermann is an associate professor with the Department of Computer Science at the School of Computing with the National University of Singapore (NUS) where he is also an investigator with the Interactive & Digital Media Institute (IDMI). His research interests are in both spatio-temporal and multimedia information management, for example distributed and peer-to-peer systems, spatio-temporal multimedia, streaming media architectures, georeferenced video management, mobile location-based services and geographic information systems (GIS). He has co-authored a book, six patents and more than two hundred conference publications, journal articles and book chapters in the areas of multimedia and databases. He has received the best paper award at IEEE ISM 2012 and was part of the team who won second place at the ACM SIGSPATIAL GIS Cup 2013. He has been involved in the organization of conferences in various positions, for example general co-chair of IEEE ISM 2015. He co-directs the Centre of Social Media Innovations for Communities (COSMIC) at NUS and is an investigator with the NUS Research Institute (NUSRI) in Suzhou, China. Roger Zimmermann is an Associate Editor of the ACM Transactions on Multimedia journal (TOMM) and the Multimedia Tools and Applications (MTAP) journal. He is a Senior Member of the IEEE and a member of ACM. For more details, see http://www.comp.nus.edu.sg/~rogerz.

Caching Methods for Scalable Video Transmission

A short review for "Caching-Based Scalable Video Transmission Over Cellular Networks"
*(Edited by **Koichi Adachi**)*

L. Wu and W. Zhang, "Caching-Based Scalable Video Transmission Over Cellular Networks", IEEE Wireless Communication Letter, vol.20, no.6, pp.1156-1159, June 2016.

In mobile network, the capacity limited backhaul link and time-varying wireless channel are the major bottlenecks in the high-definition and low-latency mobile video service. One possible solution is pre-storing some popular videos in local cache of base station (BS). The users can retrieve their requested videos mainly from the local cache, which does not require backhaul transmission, and retrieve the other videos that are not locally cached from the Internet video server via the capacity limited backhaul. This local caching avoids the unnecessary transmission of video traffic over capacity limited backhaul and reduces the transmission delay, thus enhancing user satisfaction. Due to the increased capability of computing and storage of BSs, wireless caching approaches have been proposed for improving mobile video performance [1]-[3]. In order to improve the so-called hit probability of cached contents and system throughput, the optimization of video cache placement is required for different video contents that have uneven preferences.

In previous video caching algorithms, the differentiated quality requests for the same video with respect to different user equipment (UE) have been largely neglected. Scalable video coding (SVC) [4] can provide scalability in signal-to-noise power ratio (SNR) quality, frame rate, and resolution dimensions to cate for various UE requests. Furthermore, previous works addressed the video cache placement only by exploiting the non-uniformity of video popularity while neglecting the differentiated requests for video quality. In addition, most existing works have neglected the mutually dependent relationship between the backhaul capacity and caching-based transmission.

In this letter, the authors propose an analytical framework of caching-based video transmission which takes into account the layered coding structure. The impact of video popularity and quality on the scalable video caching are firstly analyzed. Furthermore, the mutually dependent relationship between the backhaul capacity and caching-based transmission is studied.

Each video is encoded into two layers, *base layer* (*BL*) and *enhancement layer* (*EL*). By receiving the different layers, the UE can decode the video with different quality. If only BL is decoded, the video quality is standard-definition (SD). If both BL and EL are decoded, the video quality becomes high-definition (HD). Each video layer is encoded into different packet. In this letter, the video on demand streaming service is considered, therefore the transmission delay of each packet should be below predetermined threshold. A probabilistic caching model [5] is considered, where a video is independently cached at different BSs according to the same distribution. Two specific cache placements are considered, which represent the most deterministic and random cases, respectively. The first caching placement method is *popularity priority* (*PP*), where the most popular K_2 videos are cached with both the BL and the EL, while the next most popular K_1 videos are cached with only the BL. The second caching placement is *random caching* (*RC*), where K_1 videos are randomly chosen with only the BL cached, and K_2 videos are randomly chosen with the BL and the EL cached.

The authors consider five scenarios regarding the availability of different video layers at local cache of nearby BS. For example, one scenario is when the UE requests the HD quality of video V_i but only BL packet is cached at the local BS. In that case, the remaining EL packets should be transmitted from content server via capacity limited backhaul.

As a performance metric, the authors consider two indices, namely *the backhaul offloading index* and *the user satisfaction index*. The backhaul offloading index, denoted by O, means the proportion of the requested video traffic which is transmitted from the cached BSs directly among the total requested video traffic. In view of network operating cost, the larger the value of O is, the better the performance of video cache placement will be. Since the video steaming service is delay sensitive, each packet (corresponding to a frame or a slice of the video) should be transmitted within a certain time D_δ. The transmission delay for each packet consists of the backhaul delay

and the BS transmission delay. With different cache methods and differentiated UE requests, the video stream transmission processes through the backhaul and wireless link are different. The user satisfaction index, denoted by $\mathcal{U}$, means the probability that the requested video plays fluently. In other words, $\mathcal{U}$ is the probability that the transmission delay is lower than the certain delay limitation D_δ.

Since the preferences for SD and HD qualities exhibit different trends for different video types, such as sport events, films, and others, the authors introduce a variable $g_\tau(i)$, which denotes the preference for SD quality for the video V_i of type τ. For example, if type τ corresponds to sport event, the viewer may prefer SD quality rather than HD quality in order to ensure the fluent play of video. Then, $g_\tau(i)$ becomes a decreasing function over i.

By numerical simulation, the authors show that the different property of $g_\tau(i)$, i.e., whether it is an increasing function or decreasing function over i, prefers different cache placement methods. Furthermore, the results show that both PP method and RC method can improve backhaul offloading as well as provide differentiated user services and reduced video transmission delay. It is shown that the adoption of cache placement method highly depends on layered encoding parameter and preference for different video quality.

This letter provides the analytical framework for caching method for video transmission in wireless network which can be good reference for researchers to evaluate the performance of caching method under capacity limited backhaul link. In this letter, the video type τ is fixed and common among the UEs. However, in the practical scenario, different UEs may have different request of type τ. The scenario where the multiple types of videos should be transmitted is an interesting future study.

References

[1] U. Niesen, D. Shah, and G. W. Wornell, "Caching in wireless networks," *IEEE Trans. Inf. Theory*, vol. 58, no. 10, pp. 6524–6540, Oct. 2012.

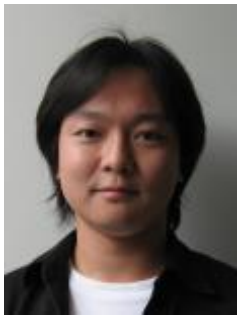
[2] K. Shanmugam, N. Golrezaei, A. G. Dimakis, A. F. Molisch, and G. Caire, "FemtoCaching: Wireless content delivery through distributed caching helpers," *IEEE Trans. Inf. Theory*, vol. 59, no. 12, pp. 8402–8413, Dec. 2013.

[3] X. Wang, M. Chen, T. Taleb, A. Ksentini, and V. C. M. Leung, "Cache in the air: Exploiting content caching and delivery techniques for 5G systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 131–139, Feb. 2014.

[4] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the scalable video coding extension of the H.264/AVC standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 9, pp. 1103–1120, Sep. 2007.

[5] B. Blaszczyszyn and A. Giovanidis, "Optimal geographic caching in cellular networks," *in Proc. IEEE Int. Conf. Commun.*, Jun. 2015, pp. 3358–3363.

Koichi ADACHI received the B.E., M.E., and Ph.D degrees in engineering from Keio University, Japan, in 2005, 2007, and 2009 respectively. From 2007 to 2010, he was a Japan Society for the Promotion of Science (JSPS) research fellow. From May 2010 to May 2016, he was with the Institute for Infocomm Research, A*STAR, in Singapore. Currently, he is an associate professor at The University of Electro-Communications, Japan. His research interests include cooperative communications and energy efficient communication technologies. He was the visiting researcher at City University of Hong Kong in April 2009 and the visiting research fellow at University of Kent from June to Aug 2009.



Dr. Adachi served as General Co-chair of the 10th and 11th IEEE Vehicular Technology Society Asia Pacific Wireless Communications Symposium (APWCS) and Track Co-chair of Transmission Technologies and Communication Theory of the 78th and 80th IEEE Vehicular Technology Conference in 2013 and 2014, respectively. He was recognized as the Exemplary Reviewer from IEEE COMMUNICATIONS LETTERS in 2012 and IEEE WIRELESS COMMUNICATIONS LETTERS in 2012, 2013, 2014, and 2015. He was awarded excellent editor award from IEEE ComSoc MMTC in 2013.

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:

Christian Timmerer (christian.timmerer@aau.at) and Yan Zhang (yanzhang@simula.no).

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: Yonggang Wen, Singapore
Steering Committee Chair: Luigi Atzori, Italy
Vice Chair – North America: Khaled El-Maleh, USA
Vice Chair – Asia: Liang Zhou, China
Vice Chair – Europe: Maria G. Martini, UK
Vice Chair – Letters: Shiwen Mao, USA
Secretary: Fen Hou, China
Standard Liaison: Zhu Li, 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.