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

E-LETTER

Vol. 6, No. 6, June 2011

IEEE COMMUNICATIONS SOCIETY

CONTENTS

SPECIAL ISSUE ON 3D DATA ACQUISITION, DISPLAY AND COMPRESSION

3D Data Acquisition, display and Compression

Irene Cheng, University of Alberta, Edmonton Alberta, Canada
locheng@ualberta.ca

Developing and appreciating 3D graphics and animations has become a requirement not only in research projects, but also in many practical applications and daily activities, such as advertising, movies, games, TV, training and rehabilitation. 3D content has also been integrated into many social web applications such as Second Life, GOOGLE and so on. Evolving from traditional 2D videos, the current development trend aims toward enriching human viewing experience by introducing higher dimension domains. Other than a totally immersive virtual reality environment, such as the CAVE utilizing comparatively heavy head mounted and sensing devices, wearing light-weight 3D glasses is becoming commonplace, with the added convenience of enjoying 3D movies at ease in IMAX theatres and even watching 3D televisions leisurely at home. Initiated as research goals, many 3DTV hardware and software have been delivered as commercial products.

Viewing in 3D perspectives can provide additional insight in the understanding and interpretation of multimedia contents. Given the increased popularity and interests in 3D movies and 3DTV technologies in academia, industries and the general public, the goal of our 3D Rendering, Processing and Communications Interest Group is to provide an international forum for researchers, developers, manufacturers, students and end-users to exchange knowledge, discuss R&D results, explore the latest state-of-the-art methodologies and study the effectiveness of these findings in terms of time, space and quality. We have special interest in multi-view video and stereoscopic 3D data. Our target areas also include 3D data acquisition, registration, visualization and transmission, as well as data quality and standards.

There are two special issues proposed by our Interest Group in 2011. We dedicate this issue to feature the theme on *3D data acquisition, display and compression*. The four position papers are contributed by world class researchers, who discuss many interesting and challenging problems covering real-time 3D applications on hand-held devices, object identification using stereo vision, multi-view video compression standards and the performance of stereo algorithms.

The first paper, co-authored by Dr. Wenwu Zhu, Dan Miao and Hongzhi Li (Microsoft, Asia), discusses the challenges and trends of implementing real-time 3D applications on hand-held devices. The challenges involve bandwidth, latency, computational capability, memory size, battery life and delivered quality. A possible solution of using cloud-based 3D rendering approach is suggested. Free viewpoint TV/video is used as an example to illustrate the concept.

Stereo vision can enhance human viewing experience. It can also improve the detection, tracking and identification of objects, such as recognizing people in surveillance systems. In the second paper, Dr. Gabriel Taubin (Brown University) and Dr. Yong Zhao (Google), discuss using 3D sensors to help solving computer vision problems traditionally based on 2D sensors, and describe a real-time stereo vision system to perform object identification tasks.

While efficient processing and rendering are important, compatibility of 3D hardware, software and multi-modality data, as well as effective network/storage usage are also attributed factors to the success of 3DTV. The third paper authored by Dr. Frederic Dufaux (Telecom ParisTech) reviews current and discusses forthcoming standards of video compression – which is one of the key issues to be addressed in order to ensure interoperability and hence mass adoption of the emerging 3D stereo and multi-view video technology.

At the end, Dr. Stefano Mattoccia (University of Bologna) and Leonardo De-Maeztu (Public University of Novarre) present a paper to discuss the challenges in finding correspondence points in stereo images. They propose a framework that integrates block based and point based incremental calculation schemes to obtain accurate disparity maps with dramatic reduction in execution time.

Undoubtedly, there are other challenging R&D issues to resolve. By coordinating this special issue, we intend to inspire further discussions on 3DRPC related topics, and thus make technological advances. In association, our Interest Group also coordinates a J-STSP special issue on “Emergng Techniques in 3D: 3D Data Fusion, Motion Tracking in Multi-View Video, 3DTV Archives and 3D Content Protection.” We invite your high quality submissions. The call for paper was published in the E-Letter April, 2011 issue.

On behalf of the 3DRPC IG, special thanks to all contributors and the E-Letter Editorial Board in coordinating this special issue.

Irene Cheng, SMIEEE, is the Scientific Director of the iCORE Multimedia Research Centre and an adjunct faculty in the Faculty of Science as well as the Faculty of Medicine & Dentistry, University of Alberta, Canada. Her research interests, among others, include incorporating human perception – JND– following psychophysical methodology, to improve multimedia, graphics and computer vision techniques. She completed her PhD at the University of Alberta and conducted postdoctoral research at the University of Pennsylvania. Before joining academia, she was a regional Information Technology executive in Lloyds Bank International Far East Division. She received an Alumni Recognition Award in 2008 from the University of Alberta for her R&D contributions. She has received, or been offered, many scholarships and fellowships from NSERC, iCORE and others. Dr. Cheng is the Chair of the IEEE Northern Canada Section, EMBS Chapter (2009-2011), Board Member of the IEEE System, Man and Cybernetics (SMC) Society, Human Perception in Vision, Graphics and Multimedia TC, and the Chair of the IEEE Communication Society, MMTC Interest Group on 3D rendering, processing and communications (2010-2012). She is a General Chair in IEEE ICME 2011 and is a visiting professor funded at Institut National des Sciences Appliquees (INSA) de Lyon, France 2011. She has over 100 publications including two books.

Real-time 3D Applications on Handheld Devices: Challenges and Trend

Wenwu Zhu, Dan Miao, and Hongzhi Li, Microsoft Research Asia
{wenwuzhu, *v-danmia,v-ansli}@microsoft.com*

1. Introduction and Challenges of Real-time 3D Applications on Handheld Devices

With the development of computer vision, graphics, and display technology, 3D applications have drawn great attentions and are getting into people's daily life. Real-time 3D applications usually focus on the interaction with users, and render a 3D scene with objects in real-time according to users' action, thereby providing richer user experience comparing with those 2D applications. In the past decades, most 3D applications run on the fat clients who have strong computing ability (e.g., PC). Thanks to the researchers' efforts, many of the challenges in real-time 3D applications, such as 3D reconstruction and rendering, etc., on the fat clients have already been solved or made good progress.

As the result of rapid developments in wireless communications and mobile devices, there has been an increasing demand to view real-time 3D applications on handheld devices. Real-time 3D applications on handheld devices can be roughly grouped into the following categories: 1) augmented / virtual reality applications; 2) 3D image applications; 3) 3D video applications. The first category includes the applications to reconstruct a virtual or real+virtual 3D world, such as Telepresence [1], second-life (http://secondlife.com/), etc. The second category includes stereo images and 3D scenes which are built using images as the source media, such as Photosynth (http://www.photosynth.net), street-side (http://www.bing.com/maps), or street view (http://maps.google.com/maps). In the last category, 3D video applications include stereo video and free –view / multi-view videos [2].

To develop real-time 3D applications on handheld devices is not a simple extension from the traditional 3D applications on PC, and presents new challenges. First, real-time 3D media is usually of great volume, imposing requirements on real-time processing and high-bandwidth & low-latency delivery. Second, mobile devices have limited computational capabilities, memorize size, and battery life, which makes 3D computing on mobile phone difficult. Third, wireless networks' characteristics change dynamically, caused by fading, etc., and network bandwidth of mobile phones is usually narrow.

2 Cloud-based Real-time 3D Rendering for Mobile Devices

For a 3D application system, the entire processing chain is mainly composed of *capturing*, *compression*, *transmission*, *rendering*, and *representation*. Due to page limitation, we omit the details. Various technical challenges and state-of-the-art solutions can be found in [3-8].

Even with recent advances in the above technologies, it is still not easy to build real-time 3D application systems on mobile phones, which can capture, store, and process a large number of 3D data in real-time, due to the limitation of wireless network bandwidth and computing capability of mobile phones. To overcome the limitations, people started to work toward a trend of using server/proxy to do view-rendering computation for mobile phones. For example, Shi et al. proposed to use proxy to do 3D rendering for mobile devices [9].

Multimedia cloud computing is an emerging technology aiming to provide a variety of computing and storage services over the Internet and wireless networks [10]. To address the challenges mentioned above, in this paper, we present a new trend of real-time 3D applications on mobile phones based on multimedia cloud computing. This is motivated by the fact that mobile phones have limited computation capability, battery life, and memory size while cloud has abundant computation and storage resources. To demonstrate this concept, next we will use cloud-base rendering of free viewpoint TV/video (FTV/FVV) as an example. FTV/FVV is an innovative visual media that enables us to view a 3-D scene by freely changing our viewpoints. The architecture of FTV/FVV based on multimedia cloud computing is shown as follows.

In this architecture, Media Edge Cloud (or Media Edge Cloudlet) is proposed, in which media contents and computation are pushed to the edge of the cloud to reduce latency. MECs can be managed in a centralized way or in a Peer-to-Peer (P2P) manner. Encoding and rendering computation of FTV/FVV or multi-view video can be executed in CPU and GPU clusters in an MEC. Traditionally, rendering is conducted at client side. However, rendering on mobile devices or computationally constrained devices impose great challenges due to limited battery life and computing power as well as narrow wireless bandwidth. In essence, an optimal resource allocation strategy between the cloud and mobile devices is needed such that some portion of rendering task can be shifted from client to cloud. Considering the tradeoff between computing and communication, there are two types of cloud-based rendering. One is to conduct all the rendering in cloud; the other is to conduct part of the rendering in cloud while the rest in client [10]. In the former case, cloud will do all the rendering computing, e.g. for the case of thin-client. In the latter case, the key problem is how to allocate rendering task between clients and cloud, which will involve client-cloud resource partition / optimization for multimedia computing. Rendering allocation could be performed based on different criteria. How to formulate the resource allocation problem and find an efficient and dynamical rendering task allocation algorithm between mobile devices and the cloud to optimize QoE (Quality of Experience), such as video quality and interaction delay, is one of our current researches.

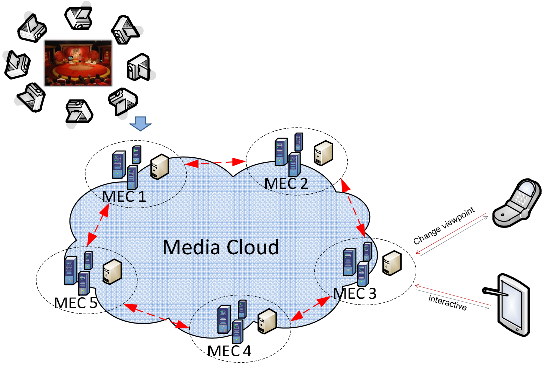


Fig. 1. Architecture of cloud-based FTV/FVV for handheld devices.

The bandwidth limitation problem could also be solved by cloud-based rendering, in that rendering task is performed mainly in the cloud; while just novel view or some parts of multi-view streams are needed to transmit to user. Moreover, the rate adaptation could be performed from an MEC's proxy to different clients in heterogeneous network for achieving better QoE.

3. Conclusions

In this paper, we first introduced real-time 3D applications on mobile devices and presented the challenges. Then we presented cloud-based 3D rendering on mobile device as a future trend, and related research directions were discussed.

References

[1] Telepresence: our remote controlled robotic Future, Special Issue Report, IEEE Spectrum, 2010.

[2] Tanimoto M., Tehrani M.P., Fuji T., Yendo T, “Free-Viewpoint TV,” IEEE Signal Processing Magazine, pp. 67-76, Jan. 2011.

[3] E. Stoykova, A. Alatan, P. Benzie, N. Grammalidis, S. Malassi-otis, J. Ostermann, S. Piekh, V. Sainov, C. Theobalt, T. Thevar,and X. Zabulis, “3-D Time-Varying Scene Capture Technologies - A Survey,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 17, no. 11, pp. 1568–1586, 2007.

[4] A. Smolic, K. Mueller, N. Stefanoski, et al., “Coding algorithms for 3DTV - A survey,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 17, no. 11, pp. 1606–1620, 2007.

[5] G. B. Akar, A. M. Tekalp, C. Fehn, and M. R. Civanlar, “Transport methods in 3DTV - A survey,” IEEE Trans. Circuits and System for Video Technology, vol. 17, no. 11, pp. 1622–1630, Nov. 2007.

[6] S.C. Chan, H.Y. Shum, and K.T. Ng, “Image-based rendering and synthesis,” IEEE Signal Processing Magazines, pp. 22–33, Nov. 2007.

[7] P. Benzie, J. Watson, P. Surman, I. Rakkolainen, K. Hopf, H. Urey, V. Sainov, C. von Kopylow, "A Survey of 3DTV Displays: Techniques and Technologies," IEEE Transactions on Circuits and Systems for Video Technology, vol.17, no.11, pp.1647-1658, Nov. 2007.

[8] Alatan, A., Yemez, Y., Gudukbay, U., Zabulis, X., Muller, K., Erdem, C., Weigel, C.2007. “Scene Representation Technologies for 3DTV—A Survey,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 17, no. 11, pp.1587-1605, Nov. 2007.

[9] Shu Shi, Won J. Jeon, Klara Nahrstedt, and Roy H.Campbell, “Real-time remote rendering of 3d video for mobile devices,” in MM ’09. 2009, pp. 391–400, ACM.

[10] Wenwu Zhu, Chong Luo, Jianfeng Wang, Shipeng Li, “Multimedia Cloud Computing: Application and Direction”, to appear, Special Issue on Distributed Image Processing and Communications, IEEE Signal Processing Magazine, May 2011.

Wenwu Zhu is a Senior Researcher at Internet Media Group of Microsoft Research Asia currently. Prior to his current post, he was a Principal Architect at Microsoft Advanced Technology Center from 2008 to 2009. He was the Director and Chief Scientist of Intel Communication Technology Lab China as well as an Intel’s

Principal Engineer from 2004 to 2008. He was with Microsoft Research Asia's Internet Media Group and Wireless and Networking Group as Research Manager from 1999 to 2004. He was with Bell Labs at AT&T/Lucent Technologies as a Member of Technical Staff during 1996-1999. He has published over 200 refereed papers in the areas of wireless/Internet multimedia delivery and wireless communications and networking. He participated in the IETF ROHC WG on robust TCP/IP header compression over wireless links and IEEE 802.16m WG standardization. He is inventor or co-inventor of over 40 patents. His current research interest is in the area of multimedia communication and computing.

Dr. Zhu has severed and been serving on various editorial boards of IEEE journals, such as Guest Editor for the Proceedings of the IEEE, IEEE JSAC, and IEEE Wireless Communication Magazine, Associate Editor for IEEE Transactions on Mobile Computing, IEEE Transactions on Multimedia, and IEEE Transactions on Circuits and Systems for Video Technology. He received the Best Paper Award in IEEE Transactions on Circuits and Systems for Video Technology in 2001 from IEEE Circuits and Systems Society. He also received the Best paper award from Multimedia Communication Technical Committee in IEEE Communications Society in 2004. Dr Zhu served as Chairman and Secretary of Visual Signal Processing and Communication Technical Committee in IEEE Circuits and Systems Society from 2006-2008 and 2004-2006 respectively, and served on the Steering Committee board of IEEE Transactions on Mobile Computing from 2007-2010. Currently he serves as Chairman of Beijing Chapter of IEEE Circuits and System Society, and serves on advisory board of International Journal of Handheld Computing Research. He will serve as TPC co-chair for IEEE ISCAS 2013. He is a Fellow of the IEEE.

Wenwu Zhu received the B.E. and M.E. degrees from National University of Science and Technology, China, in 1985 and 1988, respectively, the M.S. degree from Illinois Institute of Technology, Chicago, and the Ph.D. degree from Polytechnic Institute of NYU, New York, in 1993 and 1996, respectively, in Electrical and Computer Engineering.

.

Dan Miao received the B.E. degree from University of Science and Technology of China (USTC) in 2009. He is currently working toward the Ph.D. degree in Department of Electrical Engineering, USTC. He has been a research intern in Microsoft Research Asia since 2010. His research interests include video coding, video streaming, and transmission.

Hongzhi Li is an undergraduate student in Computer Science in Zhejiang University, China. He has been a research intern in Microsoft Research Asia since 2009. His research interests are in the areas of multimedia search, computer vision and mobile multimedia computing.

Using Stereo Vision to Detect, Track and Identify People

Yong Zhao and Gabriel Taubin, Brown University, Providence, Rhode Island, USA
{yong_zhao, taubin}@brown.edu

1. Introduction

This article briefly describes a real-time stereo vision system used to detect, track, and identify people for surveillance applications. The use of stereo vision results in a dramatic performance improvement in this traditional computer vision application, compared with traditional system based on monocular sensors. The additional depth information provided by the stereo vision system transforms a 2D image-domain problem into a decision-making problem in the 3D world.

2. Capture the 3D Scene in Real-time

Surveillance systems usually operate without interruptions 24 hours a day. In addition, these systems often need to respond immediately to certain events which require prompt handling. We argue that a sensor able to capture the 3D structure of a scene in real-time is required to accomplish these tasks.

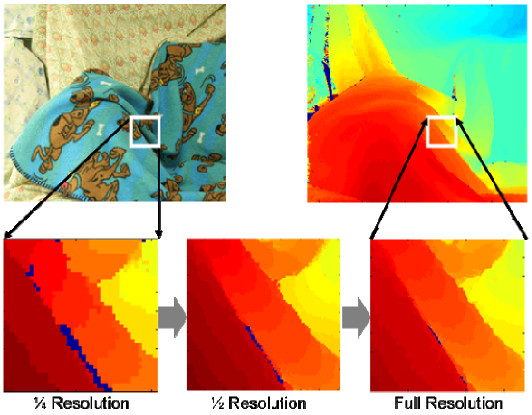


Fig.1 Progressive refinement of disparity map

In order to meet the speed requirement, a GPGPU-based real-time dense stereo matching algorithm was developed. The basic idea of this algorithm is illustrated in Fig.1. A stereo pyramid is generated from the captured stereo image pair, and the stereo matching process is started at very low resolution. Based on the low resolution result, the disparity values are progressively refined at higher resolutions. The underling matching algorithm is the “adaptive window” approach [1] because of it high accuracy with relatively small matching window. Taking the advantage of the fact that only moving objects are of interest for surveillance applications, an appearance based background model is used to limit the processing to the region composed of image pixels labeled as foreground. By carefully mapping the computation to an off-the-shelf commodity graphics card with GPGPU architecture, our implementation is able to perform 7200M disparity evaluations per second, which leads to a speed of 36Hz for the stereo matching algorithm on 1024 × 768 stereo video with disparity range of 256 pixels, which is very fine compared with earlier competing algorithms.

3. People Detection and In-camera Tracking

Shadows of moving objects change the photometric appearance background pixels, but they do not affect their depths. Therefore an appearance-based background model is unable to handle moving shadows. Fig.2 illustrates how our algorithm, which uses a depth-based background model, detects a real moving foreground object and its shadows.

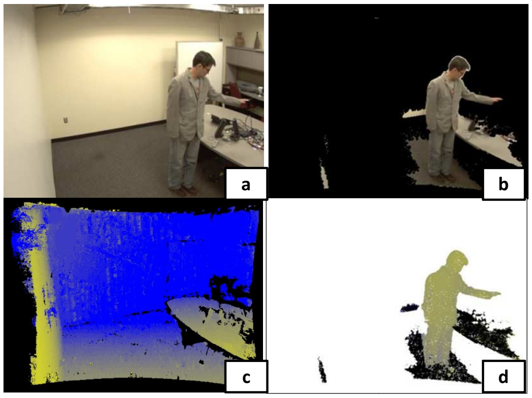


Fig.2 (a) A moving person and his shadows casted on furniture and floor. (b) Foreground regions detected by the appearance-based background model. (c) Pre-measured depth map of background. (d) Foreground regions detected by depth-based background model. Shadows are highlighted with black color.

After the shadows are removed, a 3D point cloud can be generated from the foreground pixels and the estimated depth map. Like some previous works [2,3], our algorithm then projects the 3D point cloud on a top-down plan-view coordinate system for further processing. The justification for this step is that it is much easier to segment individuals who are typically standing from such a view point, regardless of their distances to the camera and occlusions amongst them. The top-down view also provides a good direct summary impression of the objects’ behaviors in world coordinates. Fig.3 illustrates this process. This technique works very well, even in very crowded scenes in which objects are tightly connected and partially occluded in the camera-view.

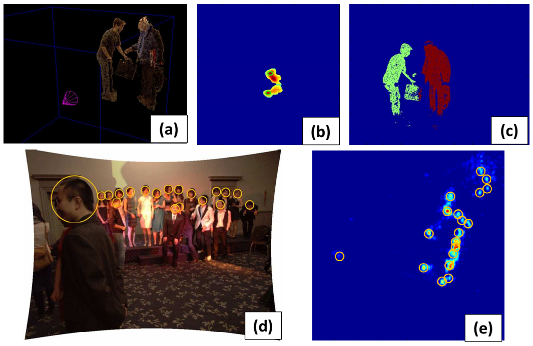


Fig.3 (a) 3-d point clouds of foreground objects with shadows being removed. (b) Objects are detected by segmenting blobs on plan-view occupancy map. (c) The camera-view visualization of segmentation results. (d) A very crowded scene with people detection results highlighted with circles on head positions. (e) The plan-view occupancy map of the crowded scene.

After objects are detected, the algorithm performs in-camera tracking. In-camera tracking means tracking objects which are observed from the point of view of one camera node. A simplified Maximum Likelihood Estimation (MLE) framework is used to obtain the most likely connections of objects' identifications in consecutive frames. The likelihood in the framework is approximated by a similarity score which is defined by the spatial continuity, and similarity of visual cues. Two kinds of visual cues are used here: 3D local feature and 3D localized color histogram.

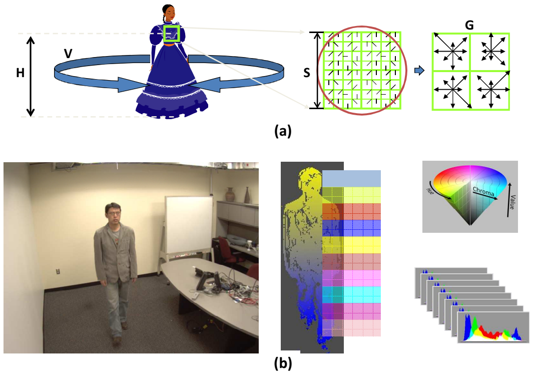


Fig. 4 (a) 3-d enhanced SIFT feature. (b) 3-d localized color histogram.

Fig.4 illustrates the 3D enhanced visual cues used for object tracking. SIFT features [4] are detected on 2-d image. 3D information is used to select a subset of features that are from a relatively flat surface with a nearly head-on direction to the camera. These features are much more repeatable than other ones. Besides, the height, view-point and physical size of the feature are used to enhance the strength of the original HoG based SIFT feature descriptor. 3D information is also used to split an object into multiple parts based on the height. Then the color histograms are computed for each individual part. Experiments have shown that these 3D enhanced visual cues significantly improve the tracking performance. Fig.5 shows an example of tracking multiple moving persons in both the camera view and the top-down plan-view.

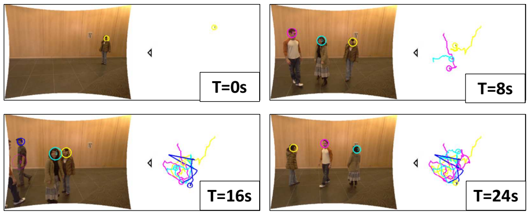


Fig.5 In-camera tracking of multiple moving persons.

4. Person Identification

Identifying people is necessary for tracking and searching individuals in a sparsely deployed camera network. Compared with in-camera tracking, tracking objects across spatially disconnected camera fields of view is more difficult because: 1) there is no spatial continuity; 2) color histograms are less stable due to the lighting variations across different camera locations; and 3) object matching has to be performed on a larger database.

For cross-camera tracking, each object is represented by an object descriptor which contains a set of distinctive 3D local features and their spatial relationship. The idea is that for two instances of the object to be matched, not only their features, have to be largely matched, but also the spatial structures of features must agree with each other.

However such object descriptor contains many local features which are not effective when used in a large scale database. We address the scalability problem with a multi-resolution compression scheme for object descriptors. First the high dimensional HoG-based feature descriptor is compressed to 1D index using the technique developed in content-based image searching [5]. A low resolution object descriptor contains only a subset of the 1D index. A medium resolution object descriptor contains the 1D index set and the spatial structure of features. And a full resolution object descriptor contains the 128D HoG descriptor and the spatial structure of features. When matching object in a large database, low resolution object descriptors are first used to find a small set of candidates. Median and full resolution object descriptor are then used to progressively find the best match.

We have conducted an experiment to test the object identification using low, medium and full resolution descriptors. Object descriptors are created for 20 people in a supervised environment. We then ask each person to show up again in front of the camera and to try to re-identify them. Fig.5 shows the performance using three different resolutions of object descriptors.

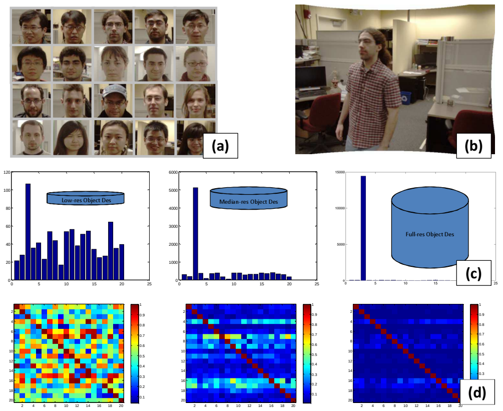


Fig.6 (a) A database of pre-generated object descriptors for 20 people. (b) A new instance of person #3. (c) Similarity scores computed between the new instance and 20 objects in database, using low, median and full resolution object descriptor. All three resolutions give correct answer. The full resolution performs with highest confidence. (d) From left to right, the confusion matrixes of object re-identification experiment on every person using the low, median and full resolution object descriptors.

6. Conclusion

We have shown that 3D information can be used to handle the problem of shadows, occlusion and crowdedness in object detection and tracking applications. 3D information also helps to enhance the strength of 2D local image features and color histograms, resulting in significant performance improvements in object identification tasks. We believe that using 3D sensors will ultimately help to solve many computer vision problems traditionally based on 2D sensors.

References

[1] R. Yang and M. Pollefeys. Improved real-time stereo on commodity graphics hardware. In Proceedings of Conference on Computer Vision and Pattern Recognition Workshop on Realtime 3D Sensors and Their Use, 2004.

[2] M. Harville. Stereo person tracking with adaptive plan-view templates of height and occupancy statistics. IVC, 22(2):127–142, February 2004.

[3] Rafael Munoz-Salinas, Eugenio Aguirre, and Miguel Garc´ıa-Silvente. People detection and tracking using stereo vision and color. Image Vision Comput., 25(6):995–1007, 2007.

[4] D. Lowe. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 2004.

[5] David Nister and Henrik Stewenius. Scalable recognition with a vocabulary tree. In CVPR '06: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pages 2161–2168, Washington, DC, USA, 2006. IEEE Computer Society.

Gabriel Taubin earned a Licenciado en Ciencias Matemáticas degree from the University of Buenos Aires, Argentina, and a Ph.D. degree in Electrical Engineering from Brown University. In 1990 he joined IBM, where during a thirteen years career in the Research Division he held various positions, including Research Staff Member and Research Manager. In 2003 he joined the Brown University School of Engineering as an Associate Professor of Engineering and Computer Science. He was Visiting Professor of Electrical Engineering at the California Institute of Technology during the 2000-2001 academic year, and Visiting Associate Professor of Media Arts and Sciences at MIT in 2010. He is the current Editor-in-Chief of the IEEE Computer Graphics and Applications Magazine, and was named IEEE Fellow for his contributions to the development of three-dimensional geometry compression technology and multimedia standards.

Yong Zhao is a Software Engineer at Google Research Lab. He got his PhD degree at the School of Engineering, of the Brown University. Prior to that, he did research internship at Nvidia Research Lab, EPSON Lab and MERL. His work has ranged over computational photography, multimedia processing, and real-time computer vision processing on SIMD platforms.

Emerging 3D Stereo and Multiview Video Compression Standards

Frédéric Dufaux, CNRS UMR 5141 – LTCI, Télécom ParisTech, Paris, France
frederic.dufaux@telecom-paristech.fr

1. Introduction

Interest for three-dimensional (3D) video is steadily gaining momentum. With the expectation of greatly enhanced user experience, 3D video is widely perceived as the next major advancement in video technology.

Several 3D video formats, coding schemes, and display technologies currently coexist. Standardization is one of the key issues to be addressed in order to ensure interoperability and hence mass adoption of such technology. MPEG has previously standardized and brought to the market several 3D video formats. For instance, industry is already deploying products and services based on Multiview Video Coding (MVC) and frame-compatible stereoscopic formats.

However, new market needs are emerging along with advances in 3D displays and services [1]. Efficient 3D video representations, which enable the reconstruction of an arbitrarily large number of views prior to rendering, are proposed in [2][3]. Free Viewpoint Video (FVV), which allows for interactively varying the viewpoint, is presented in [4].

To address these new requirements, MPEG is initiating a new phase of standardization for 3D Video Coding (3DVC), with the objective to go beyond the capabilities of existing standards. A Draft Call for Proposals has been recently released [5], and the Final Call for Proposals will be formally issued in March 2011. More specifically, 3DVC aims at supporting the synthesis of multiple views and enabling advanced stereoscopic processing.

2. Existing 3D Video Formats

A few standardized formats already support 3D video applications. We briefly discuss them hereafter, along with a discussion on their respective advantages and shortcomings.

Frame-compatible stereoscopic formats, also known as stereo interleaving, consist in a multiplex of the left and right views into a single frame or a sequence of frames [6]. With temporal multiplexing, successive frames correspond to the left and right views respectively. Alternatively, with spatial multiplexing, the left and right views are sub-sampled and combined into a single frame, e.g. in side-by-side or top-bottom formats. Auxiliary information is required in order to correctly interpret frame-compatible formats. For this purpose, Supplementary Enhancement Information (SEI) has been standardized in the framework of H.264/MPEG-4 Advanced Video Coding (AVC) to correctly distinguish the samples corresponding to the left and right views.

However, this representation suffers from two severe drawbacks. First, spatial or temporal resolution is decreased, possibly resulting in reduced quality and user experience. Second, legacy 2D devices are not able to correctly decode SEI messages and hence fail to correctly interpret the interleaved data. In particular, this is a key issue in broadcast environment where it is costly to upgrade devices.

The *2D video plus depth* format is another useful representation. Depth information results in a display-independent representation which enables synthesis of a number of views. Two straightforward advantages of this representation is that the 2D video stream provides backward compatibility with legacy devices and it is independent of underlying coding formats. Depth data has different characteristics when compared to natural video data, usually resulting in better compression performance. However conventional coding methods may not be optimal and coding artifacts also impact the quality of synthesized views. ISO/IEC 23002-3, more commonly known as MPEG-C Part 3, is a specification for the standardized representation of auxiliary video data, including depth map. Its drawback is that it is only capable of rendering a limited depth range.

Multiview Video Coding (MVC) [7] is an extension of the AVC standard [8]. It addresses multiview video representation resulting from multiple cameras capturing the same scene from different viewpoints. This typically generates a tremendous amount of data, and therefore efficient compression is paramount. Since multiview video also exhibits significant inter-view statistical redundancies, MVC combines conventional intra-view temporal prediction and inter-view prediction from neighboring views [9]. Moreover, MVC streams must include an AVC base layer for backward compatibility with legacy 2D devices. Subjective quality assessment tests have shown that MVC can achieve a substantial bit rate saving when compared to AVC simulcast for the same visual quality.

Auto-stereoscopic displays, requiring a large number of views, are one of the primary targets of MVC. However, a major shortcoming of the design is that the bit rate essentially grows proportionally to the number of encoded views. As a consequence, channel bandwidth constraints typically prevent a large number of views.

Nevertheless, MVC still proves useful for stereo video content delivery. Indeed, it has been adopted for stereo coding on the 3D Blu Ray Disc format. Both the good coding efficiency and backward compatibility are essential features. When compared to frame-compatible formats, MVC maintains full resolution. In addition, it usually achieves better rendering quality than 2D video plus depth format.

3. Forthcoming 3D Video Coding

With the goal to go beyond existing standards, MPEG has initiated a new phase of standardization for 3D Video Coding (3DVC).

Two major objectives are targeted [1]. The first one is to support advanced stereoscopic display processing, in order to allow stereo devices to cope with diverse display types and dimensions as well as varying viewing conditions. It includes the adjustment of depth perception by controlling baseline stereo distance, a feature which proves useful to improve user viewing experience and to prevent fatigue. The second objective is to improve support for high quality auto-stereoscopic multiview displays. More specifically, 3DVC aims at enabling the synthesis of many high-quality views with a limited bit rate. For example, with a representation based on stereoscopic video and respective depth maps, the bit rate is decoupled from the number of views.

It is expected that 3DVC enhances the 3D rendering capabilities, when compared to the limited depth range supported with frame-compatible formats. Moreover, 3DVC does not trade-off resolution. When compared to MVC, 3DVC is expected to significantly reduce the bit rate needed to generate the views required for display.

Technology requirements for 3DVC are detailed in [10]. Regarding data format, the uncompressed data format shall support stereo video as well as other configurations beyond stereo. To enable efficient and high quality view synthesis, supplementary data, including depth maps, segmentation information, transparency or specular reflection, and occlusion data, shall be supported.

In terms of compression, the bit rate for video and supplementary data should not exceed twice the bit rate of state-of-the-art compressed 2D video. Moreover, 3DVC should outperform state-of-the-art multiview coding. Compression should not adversely impact the visual quality of synthesized views. The compressed data format shall include forward compatibility with AVC and the forthcoming High Efficiency Video Coding (HEVC). It shall also include a mode that enables simple stereo and mono compatibility.

With respect to rendering, 3DVC should support superior rendering capability when compared to current state-of-the-art representations. In addition, it shall be display-independent and support various types and sizes of displays. Finally, 3DVC shall support a variable stereo baseline and an appropriate depth range.

A Draft Call for Proposals has been released in January 2011 [5]. The Final Call for Proposals is scheduled to be issued at the 96th MPEG meeting in March 2011. Technologies are sought for efficient 3D video compression as well as high quality view synthesis. Proponents should make available coded test material before October 2011, and submit documents describing the proposals by November 2011. Subjective assessment starts in October and evaluation will take place at the 98th MPEG meeting in November-December 2011.

In the development of 3D video technologies, quality assessment is a significant challenge [11]. While subjective evaluation of 2D video quality has reached some maturity with several methodologies recommended by ITU, subjective assessment of 3D video raises new issues. In particular, the viewing experience becomes multi-dimensional and involves not only visual quality, but also depth perception and viewing comfort. Finally, the specific 3D display technology also has a significant impact. To evaluate 3DVC submissions, MPEG intends to carry out subjective tests on both stereoscopic and auto-stereoscopic displays.

4. Concluding remarks

3D video is a very hot topic nowadays. In this article, we first reviewed existing standardized 3D formats. We then discussed the forthcoming standard for 3D stereo and multiview video compression currently under development in MPEG, with the objective to go beyond the capabilities of current solutions.

On a parallel path, MPEG is also exploring a Frame-Compatible Enhancement (FCE) for AVC, which targets compatible transmission of stereoscopic video at high definition resolution.

References

[1] ISO/IEC JTC1/SC29/WG11, “Vision on 3D Video”, N10357, Lausanne, Switzerland, Feb. 2009.

[2] A. Vetro, S. Yea, A. Smolic, “Towards a 3D Video Format for Auto-Stereoscopic Displays”, Proc. SPIE Applications of Digital Image Processing XXXI, San Diego, CA, August 2008.

[3] K. Müller, P. Merkle, G. Tech, and T. Wiegand, “3D Video Formats and Coding Methods”, Proc. IEEE Int. Conf. in Image Processing, Hong Kong, Sept. 2010.

[4] M. Tanimoto, “Overview of Free Viewpoint Television”, Signal Processing: Image Communication, vol. 21, no. 6, July 2006.

[5] ISO/IEC JTC1/SC29/WG11, “Draft Call for Proposals on 3D Video Coding Technology”, N11830, Daegu, Korea, Jan. 2011.

[6] A. Vetro, “Frame Compatible Formats for 3D Video Distribution”, Proc. IEEE Int. Conf. in Image Processing, Hong Kong, Sept. 2010.

[7] ISO/IEC JTC1/SC29/WG11, “Text of ISO/IEC 14496-10:200X/FDAM 1 Multiview Video Coding”, N9978, Hannover, Germany, July 2008.

[8] T. Wiegand, G.J. Sullivan, G. Bjøntegaard, and A. Luthra, “Overview of the H.264/AVC Video Coding Standard”, IEEE Trans. on Circuits and Systems for Video Technology, vol. 13, no. 7, July 2003.

[9] P. Merkle, A. Smolic, K. Mueller, and T. Wiegand, “Efficient Prediction Structures for Multiview Video Coding”, IEEE Trans. Circuits and Systems for Video Technology, vol. 17, no. 11, Nov. 2007.

[10] ISO/IEC JTC1/SC29/WG11, “Applications and Requirements on 3D Video Coding”, N11829, Daegu, Korea, Jan. 2011.

[11] Q. Huynh-Thu, P. Le Callet, M. Barkowsky, “Video Quality Assessment: from 2D to 3D - Challenges and Future Trends”, Proc. IEEE Int. Conf. in Image Processing, Hong Kong, Sept. 2010.

Frederic Dufaux is a CNRS Research Director at Telecom ParisTech. He is also Editor-in-Chief of Signal Processing: Image Communication.

Frederic received his M.Sc. in physics and Ph.D. in electrical engineering from the Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland, in 1990 and 1994 respectively. He has held research positions at EPFL, Emitall Surveillance, Genimedia, Compaq, Digital Equipment, MIT, and AT&T Bell Labs.

Frederic has been involved in the standardization of digital video and imaging technologies for more than 15 years, participating both in the MPEG and JPEG committees. He is currently co-chairman of JPWL and JPSearch. He is the recipient of two ISO awards for his contributions.

His research interests include image and video coding, 3D video, visual quality assessment, video surveillance, privacy protection, image and video analysis, multimedia content search and retrieval, and video transmission. He is the author or co-author of more than 100 research publications and holds 17 patents issued or pending.

A Discussion on the Accuracy and Performance of Local Stereo Algorithms for Stereo Correspondence

Stefano Mattoccia, University of Bologna, Bologna, Italy
Leonardo De-Maeztu, Public University of Navarre, Spain
stefano.mattoccia@unibo.it, leonardo.demaeztu@unavarra.es

1. Introduction

Stereo vision aims at inferring 3D information from two images of the same scene simultaneously acquired from two different viewpoints. Due to the large number of application scenarios that can take advantage of 3D information, this topic received a lot of attentions in the last decades and an extensive review can be found in [1, 2].

Given two images, referred to as reference (R) and target (T), if we are able to find corresponding points (i.e. projections of the same scene point in R and T) stereo vision allows obtaining depth by means of a simple triangulation [1]. However, finding corresponding points in the two images is a challenging task and many algorithms have been proposed. According to [1, 2] most approaches perform four steps (*cost computation*, *cost aggregation*, *disparity optimization* and *refinement*) and algorithms can be roughly classified in *local* approaches and *global* approaches.

The former class mainly relies on cost aggregation and in most cases ignores disparity optimization deploying, on a point basis, a simple *Winner Takes All* (WTA) strategy. Although the simplest approach aggregates cost on a fixed area (referred to as *support*) centered in the points under examination, more sophisticated and effective methods that aggregates costs according to image content have been proposed. In general, local algorithms have a very simple computational structure and a small memory footprint. Nevertheless, most accurate algorithms are computationally expensive.

On the other hand, global algorithms perform a disparity optimization on the whole image by means of an energy function that jointly enforces photometric consistency between images and a smoothness term that models the evidence that scenes are piecewise smooth. Most global algorithms do not perform cost aggregation focusing only on disparity optimization. Despite their effectiveness these algorithms are in most cases computationally expensive and have a large memory footprint. These drawbacks render these algorithms not suited to most practical applications.

In the next section we'll focus our attention on recent local state of the art algorithms that adapt their supports to image content discussing their accuracy and performance.

2. Local algorithms for accurate stereo correspondence based on adapting weights

Local algorithms that shape their support by means of an *adapting weight* strategy outperform other local approaches (see [4, 5] for a recent evaluations) enabling to obtain results comparable to global methods.

Algorithms based on the adapting weight strategy, originally proposed in [6], aggregate costs on a fixed squared support and assign to each point a weight computed, with respect to the central point of each support, according to the image content.

The Adaptive Weight (AW) approach [6], inspired by Gestalt theory, encodes the relevance of each matching cost according to a proximity and color distances. The former assigns higher scores to points closer to the central point while the latter assigns higher confidence to points with similar colors with respect to the central point. This basic principle, depicted in Figure 1, is applied to each point of reference and target supports (and corresponding weights are then multiplied). Although very effective [4, 5, 6] this method is computationally expensive (it requires several minutes with standard test images [3]).

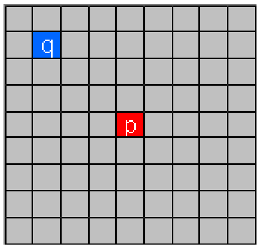


Figure 1 *Adapting weights* strategy: the matching cost of point q is weighted according to its spatial distance and color distance wrt to the central point p.

The Segment Support (SS) [7] approach improves AW deploying segmentation as additional cue. This method discards the proximity constraint (i.e. spatial distance) and computes weights with the following strategy (for reference and target support): if point q belong to the same segment of p q receives the highest weight, otherwise point q is weighted according to the color distance wrt to p. Unfortunately, despite its effectiveness, this method [4] doubles the execution time of the AW approach.

A further improvement based on the adapting weight strategy was proposed in [8]. Similarly to SS this approach discards the proximity constraint and computes weight (for points belonging to the support in reference and target images) according to the geodesic distance between p and q. The geodesic distance for point q is defined as the path with the minimum cost between q and the central point p. The cost between two adjacent points is computed as the Euclidean distance in the RGB color space. The execution time of this method is high and comparable to those of the previous approach.

Despite their effectiveness, methods based on the adapting weight strategy are computationally expensive and not suited to most practical applications. Nevertheless, a computational framework that combines the effectiveness of the adapting weight strategy with the efficiency of traditional correlation based approaches was proposed in [9]. This method, referred to as Fast Bilateral Stereo (FBS), computes approximated weights for reference and target images on a block basis assigning to each point within a block a single value assuming as reference for the block the central point q. The weight assigned to each block is computed according to the spatial and color distance between the center of the block and the central point p. On the other hand, matching costs are computed precisely on a block basis by means of incremental calculation schemes such as *box-filtering* [10] or *integral images* [11]. The block-based strategy for weight computation deployed by FBS is depicted in the following figure.

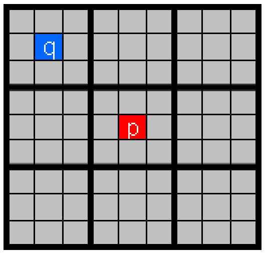


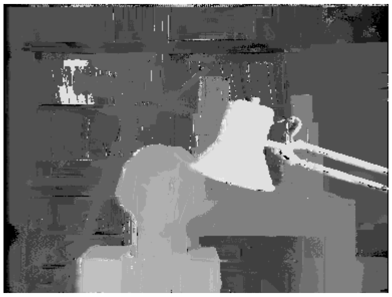
Figure 1 Block-based adapting weight strategy deployed by the FBS approach: the matching cost of point q is weighted according to the spatial distance and color distance of the block containing q wrt to the central point p.

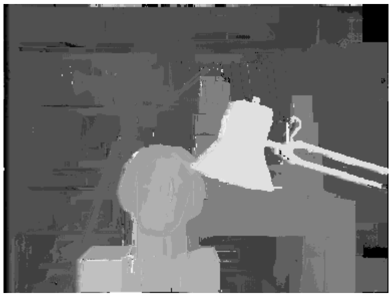
The strategy deployed by FBS enables to obtain results comparable to [6] and [7] at a fraction of the time required by these algorithms. For example, on standard stereo pairs, the execution time drops from minutes to seconds [9].

3. Experimental results

According to the metric defined in [4] (i.e. overall sum of errors in non occluded and discontinuity regions) the overall error computed on the 4 stereo pairs of the Middlebury dataset [2, 3] for AW is 83.32, for SS is 67.03 and for FBS is 67.55. The measured execution time according to [4] for AW and SS is, respectively, 20 minutes and 39 minutes. On the other hand, the measured execution time for FBS [9] is 29 seconds.

The following figures report the raw disparity maps computed on the Tsukuba stereo pair [2, 3] by AW, SS and FBS with 3x3 blocks. A qualitative analysis of these disparity maps shows that the FBS approach enables to obtain results comparable, and even better in some circumstances, to AW and SS.





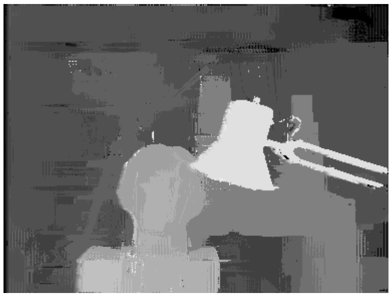


Figure 3: From top to bottom: left image of the Tsukuba stereo pair, disparity map computed by AW, disparity map computed by SS and disparity maps computed by FBS. Disparity maps and detailed experimental results are available on the web site accompanying paper [4].

In FBS the optimal size of the blocks was found to be 3x3; however, by increasing this parameter a further speed-up can be obtained thus enabling to trade efficiency for accuracy.

3. Conclusions

Recent local algorithms based on the adapting weight strategy enable to obtain very accurate disparity maps. Unfortunately, due to their high execution time, these approaches are often not suited to most practical applications. Nevertheless, a framework that computes weight on a block basis and matching cost on a point basis by means of incremental calculation schemes enables to obtain equivalent accuracy dramatically reducing the execution time.

References

[1] R. Szeliski, Computer Vision: algorithms and applications, Springer, 2010

[2] D. Scharstein and R. Szeliski, A taxonomy and evaluation of dense two-frame stereo correspondence algorithms, Int. Journal of Computer Vision, 47(1/2/3):7-42, April-June 2002

[3] D. Scharstein and R. Szeliski, Middlebury Stereo Vision Page, http://vision.middlebury.edu/stereo/

[4] F. Tombari, S. Mattoccia, L. Di Stefano, E. Addimanda, Classification and evaluation of cost aggregation methods for stereo correspondence, Int. Conference on Computer Vision and Pattern Recognition (CVPR 2008) http://www.vision.deis.unibo.it/spe/SPEHome.aspx

[5] M. Gong, R.G. Yang, W. Liang, W., M.W. Gong, A performance study on different cost aggregation approaches used in real-time stereo matching. Int. Journal Computer Vision 75(2), 283–296 (2007)

[6] K.J. Yoon and I.S. Kweon, Adaptive support weight for correspondence search, PAMI 28(4), pp 650-656,

[7] F. Tombari, S. Mattoccia, L. Di Stefano, Segmentation-based adaptive support for accurate stereo correspondence, Pacific-Rim Symposium on Image and Video Technology (PSIVT 2007)

[8] A. Hosni, M. Bleyer, M. Gelautz, and C. Rhemann, Local stereo matching using geodesic support weights, Int. Conference on Image Processing (ICIP2009)

[9] S. Mattoccia, S. Giardino, A. Gambini, Accurate and efficient cost aggregation strategy for stereo correspondence based on approximated joint bilateral filtering, Asian Conference on Computer Vision (ACCV2009), 2009

[10] M.J. McDonnell, Box-filtering techniques, Computer Graphics and Image Processing, 17(1), 65-70, Sept. 1981

[11] F. Crow, Summed-area tables for texture mapping, 11th annual conference on Computer graphics and interactive techniques, 207–212, 1984

Stefano Mattoccia is a Research Associate at the Faculty of Engineering of the University of Bologna, DEIS (Department of Electronics, Computer Science and Systems). He received a Msc in Electronic Engineering and a PhD in Electronic Engineering and Computer Science from the University of Bologna His research interests include fast and accurate stereo vision algorithms, 3D sensor fusion, image matching and registration, mapping of computer vision algorithms on FPGA, GPU and embedded devices.

Leonardo De-Maeztu is a PhD student at the Public University of Navarre. He is currently involved in local stereo matching and real-time GPU implementation of this type of algorithms. Prior to this, he was at NEC Electronics. He has a Master's degree from Public University of Navarre.

TECHNOLOGY ADVANCES

Mobile Multimedia Networks

Guest Editor: Hongbo Jiang, Huazhong University of Science and Technology, China
hongobjiang2004@gmail.com

In the past few years, we have seen a global flurry of the Internet in the rapid roll-out of multimedia - the commercial products such as PPlive, YouTube, and Skype have occupied a large portion of Internet bandwidth. Meanwhile, advances in wireless communication technologies have contributed to an explosive growth of video applications over wireless mobile networks in recent years. However, subject to the lower bandwidth, higher latency, a higher burst error rate, and user's mobility in wireless networks compared to wired networks, the end-devices in wireless networks are more heterogeneous than in wired networks. This E-letter presents five papers, providing some technologies in designing efficient mobile multimedia networks.

In the first paper, titled 'Resource Utilization in Internet Mobile Streaming', we shows a measurement study on the power efficiency in receiving streaming services under different streaming delivery architectures. One observation in this article is that the existing P2P streaming service still lacks power-efficient design and device heterogeneity handling capabilities, and deserves further optimization.

The second paper, titled 'Scalable Video Coding with Compressive Sensing for Wireless Videocast', focuses a newly technology of compressive sensing (CS) on how it facilitates video coding . We discuss many open issues that worth further investigation, e.g., how to optimize the quantization level and the bit allocation for each layer; how to reduce the decoding; how to enhance network protocols to support SVCCS with even lower cost and better performance.

In the third paper, titled 'Video Coding Platforms for Mobile Multimedia Networks', we study video coding platform. We briefly survey video coding requirements for mobile applications, as well as some of the commonly adopted solutions to comply with such requirements in this paper. Finally, we discuss the future directions on the design of the future multimedia embedded systems for mobile networks.

In the fourth paper, titled 'Mobile Vision: Opportunities and Challenges', we summarize several research opportunities or topics, as well as three major challenges, in an emerging research area, named *mobile vision*. This emerging area will be evolving with the interaction among the mobile ecosystem, the internet ecosystem, and the computing cloud.

The last paper is to overview the key issues and research trends in satellite networking, a next generation network.

Hongbo Jiang received the B.S. and M.S. degrees from Huazhong University of Science and Technology, China. He received his Ph.D. from Case Western Reserve University in 2008. After that he joined the faculty of Huazhong University of Science and Technology as an associate professor. His research concerns computer networking, especially algorithms and architectures for high-performance networks and wireless networks.

Resource Utilization in Internet Mobile Streaming

Yao Liu, Fei Li, Songqing Chen, George Mason University
Lei Guo, Microsoft Corporation
{yliud,lifei,sqchen}@cs.gmu.edu, leguo@microsoft.com

1. Internet Mobile Streaming

Recent years have witnessed a quickly increasing demand for Internet streaming to mobile devices. For example, both iOS and Android have native support for Youtube [1]. More and more content providers today also allow their customers to access multimedia content on their mobile devices via wireless connections. However, delivering high quality Internet streaming to mobile devices faces several challenges due to inherent constraints of mobile devices.

First, mobile devices are very heterogeneous, differing from each other in screen sizes, color depth, etc. Thus, streaming content must be customized to appropriate resolution, size, frame rate, bit rate, and encoding format for different types of mobile devices. Such customization could either be done in advance or at runtime. For example, Youtube transcodes the contents into several versions while they are uploaded. On the other hand, Vuclip [2] performs the transcoding upon request. Placeshifting systems like Orb [3] and AirVideo [4] also transcode the content stored at home computers and allow users to access via their mobile devices.

Second, mobile devices have limited resources, including slower CPU speed, smaller memory and storage sizes, and limited battery power. For Internet streaming, the battery power poses a fundamental constraint, as it often demands continuous operations of the wireless network interface card (WNIC) to receive the streaming data, CPU to decode the data, and screen to display scenes.

Today there are mainly three architectures being used to deliver Internet mobile streaming services, namely (1) the Client-Server (C/S) architecture, where a mobile device requests streaming data from a dedicated server, (2) the Client-Proxy-Server (C/P/S) architecture, where a transcoding proxy is deployed to customize the content at runtime, and helps deliver the content to a mobile device, and (3) the Peer-to-Peer (P2P) architecture, where a mobile device shares its uploading bandwidth.

To address the heterogeneous challenge, services using the C/S architecture would pre-code the content into several formats, and supply the appropriate format based on the supported formats of mobile devices; on the other hand, in the C/P/S architecture, the online transcoding is performed with the help of the intermediate proxy; and the P2P architecture lacks efficient transcoding in the design.

To study battery power consumption under these different architectures for Internet mobile streaming, we have conducted measurements with iPod Touch, focusing on the two major power consumption sources: CPU and the WNIC.
We studied three representative streaming services: SPBtv [5] (C/S), Orb [3] (C/P/S), and TVUPlayer [6] (P2P).

2. Battery Power Consumption by CPU

The CPU cycles in a streaming session are mainly used for two purposes: decoding the received data and transcoding for the mobile device. Our measurement results show that, for SPBtv, the video is encoded in H.264 standard, which has native support on iPhone and iOS, and can be decoded by hardware. As a result, decoding the streaming content leaves about 80% idle CPU cycles.

Orb, however, transcodes and delivers the streaming content in flv format, which requires the mobile device to resort to software for decoding. And about 40% more CPU cycles are spent compared to SPBtv.

For P2P based TVUPlayer, which delivers the same streaming content to different platforms including Windows, Mac OS, and iOS, we found that the streaming content is encoded in ASF format. This requires TVUPlayer to perform online transcoding from ASF to the supported codec at client side. And our measurement results show that TVUPlayer leaves less than 10% CPU cycles idle.

Our results show that while all three architectures can handle device heterogeneity, the CPU usage by hardware- and software-based decoding and transcoding is significantly different, which would result in different battery power consumption.

3. Battery Power Consumption by Streaming Data Transmission

Receiving streaming data on mobile devices requires the wireless network interface card (WINC) to be work continuously. This is believed to be another large source of battery power drain [7]. To save the power consumed by data transmission, commodity WNICs today all have power saving mode (PSM) supported.

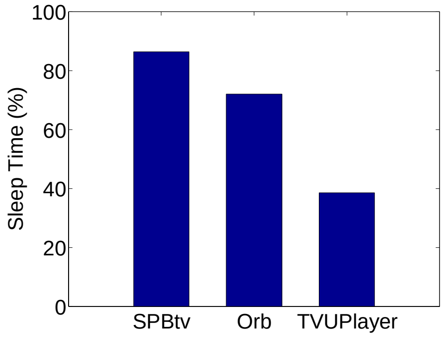


Figure 1: Sleep Time (%) of the WNIC

Figure 1 gives an overview of the percentage of time the WNIC spends in PSM (sleep mode) during 1-hour streaming sessions. This confirms that PSM does take effect in all these streaming services. However, the sleep time percentage also differs significantly. For example, P2P based TVUPlayer can sleep for less than 40% time, while C/S based SPBtv can sleep for more than 80% time, which further lead to different total power consumption.

In 802.11 PSM, if there's no network activity for a pre-defined time period, the WNIC would switch to PSM to save power. Thus, the inter-packet delay plays an important role in saving power during streaming data transmissions.

Figure 2 shows a snapshot of traffic pattern of receiving data from SPBtv. Because of the traffic shaping technique used in SPBtv, the streaming data are sent in burst, the WNIC can switch to PSM before the next burst arrives. This results that SPBtv allows the WNIC to sleep for over 80% time for power saving.

For TVUPlayer, on the other hand, Figure 3 shows that traffic shaping is not used as the inter packet delay does not show any bi-modal pattern, and only about 2% packets have an inter-packet delay larger than 100 ms, resulting in much less battery power saving during streaming.

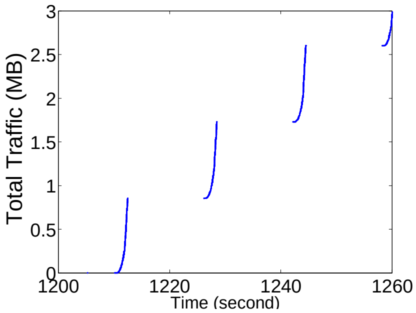


Figure 2: Bursty Traffic Pattern of SPBtv

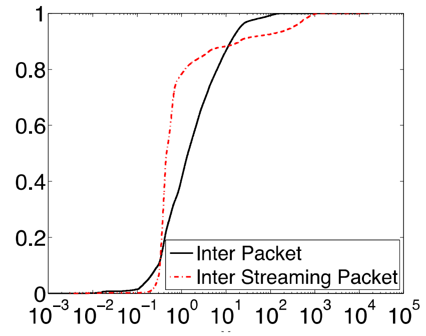


Figure 3: Inter Packet Delay (ms) of TVUPlayer (CDF)

Furthermore, for TVUPlayer, the inter-packet delay deviates significantly from the inter-streaming-packet delay as shown in Figure 3. Our examination reveals that in addition to streaming data, the mobile device also needs to exchange highly frequent control packets (buffermaps and fine-grained data chunk requests) to neighbors. It also needs to share its uploading bandwidth. The frequent control packets (twice more frequent than streaming data packets) and the additional uploading packets significantly change the traffic pattern, and further aggravate the power consumption on a mobile device.

Because of the high CPU usage due to transcoding, and less WNIC sleep time, P2P based TVUPlayer consumes much more power than C/S based SPBtv. Our stress test results show that, starting with a fully charged battery, the streaming can last for 6.4 hours when watching SPBtv, and only 3 hours when watching TVUPlayer.

4. Conclusion

Our study shows the different power efficiency in receiving streaming services under different streaming delivery architectures. While C/S and C/P/S based Internet mobile streaming have adopted techniques for power saving and heterogeneity handling, the existing P2P streaming service still lacks power-efficient design and device heterogeneity handling capabilities, and deserves further optimization.

References

[1] “YouTube,” http://www.youtube.com.
[2] “Vuclip,” http://m.vuclip.com.
[3] “OrbLive,” http://itunes.apple.com/app/id290195003.
[4] “Air Video,” http://itunes.apple.com/app/id306550020.
[5]“SPBtv,” http://itunes.apple.com/app/id356830174.
[6] “TVUPlayer,” http://itunes.apple.com/app/id323640984.
[7] T. Pering, Y. Agarwal, R. Gupta, and R. Want, “Coolspots: Reducing the power consumption of wireless mobile devices with multiple radio interfaces,” in Proc. of MobiSys, 2006.

Yao Liu is a PhD student in Computer Science at George Mason University. She received the B.S. degree in Computer Science from Nanjing University, China in June 2007. Her research interests include multimedia systems and peer-to-peer networks.

Fei Li received the B.S. degree in Computer Science from Jilin University, Changchun, China, in 1997, the M.S., M.Phil., and Ph.D. degrees in Computer Science from Columbia University, New York, NY, in 2002, 2007, and 2008, respectively. He joined the Department of Computer Science at George Mason University as an Assistant Professor in 2007. His research interests include online and approximation algorithm design and analysis, combinatorial optimization, and scheduling algorithms.

Lei Guo received the BS degree in space physics and the MS degree in computer science from the University of Science and Technology of China in 1996 and 2002, respectively. He received the PhD degree in computer science and engineering from the Ohio State University in 2007. He is currently a senior member of technical staff at the Social Search Sciences division of Microsoft Bing Search. His research interests include social search systems and algorithms, distributed information systems, social networks and semantic web, P2P systems, multimedia systems, wireless networks, and Internet measurement and modeling.

Songqing Chen received the PhD degree in computer science from the College of William and Mary in 2004. He received the M.S. and B.S. degrees in Computer Science from Huazhong University of Science and Technology in 1999 and 1997, respectively. He is currently an associate professor of computer science at George Mason University. His research interests include the Internet content delivery systems, Internet measurement and modeling, operating systems and system security, and distributed systems and high performance computing. He is a recipient of the US NSF CAREER Award and the AFOSR YIP Award.

Scalable Video Coding with Compressive Sensing for Wireless Videocast

Siyuan Xiang and Lin Cai, University of Victoria, BC, Canada
{ siyxiang,cai}@ece.uvic.ca

1. Introduction

Channel coding such as Reed-Solomon (RS) and convolutional codes has been widely used to protect video transmission in wireless networks where the communication channel has inherent impairments due to fading, shadowing, and interference, etc. However, this type of channel coding is not flexible. It can correct the bit errors only if the error rate is smaller than a given threshold. Therefore, it is hard to find a single channel code suitable for unknown or varying wireless channels. Can we find a flexible channel coding? That is, for a wide range of channel error rate, the effectiveness of channel coding degrades gracefully when the channel condition becomes worse.

Thanks to the recent advance in signal processing, the newly developed compressive sensing (CS) technologies can help to achieve the above goals. Compressive sensing or compressive sampling [3, 1] has been proposed as a new data acquisition framework which can sample and compress sparse or compressible signals in a single operation. Besides, in the research community, people become more and more interested in the characteristics of the acquired measurements [2, 6]. With CS, random linear projection of signals not only makes the encoder very simple but also makes the acquired measurements *democratic* [5], i.e., they are equally important.

If we only treat compressive sensing as an image compression method, there is a huge gap in terms of coding efficiency between compressive sensing and conventional coding methods [4]. Although compressive sensing has the advantage of being a joint source and channel coding, its coding efficiency needs to be improved, since minimizing bandwidth consumption is one of the most important goals in codec design, particularly for wireless transmissions.

A low-complex, scalable video coding architecture based on compressive sensing (SVCCS) for wireless unicast and multicast transmissions has been proposed [7]. SVCCS achieves good scalability, error resilience and coding efficiency. A SVCCS encoded bitstream is divided into a base and an enhancement layer. The layered structure provides quality and temporal scalability. The base layer is composed of a small portion of discrete cosine transform (DCT) coefficients. The enhancement layer consists of compressive sensed measurements. While in the enhancement layer, the CS measurements provide fine granular quality scalability. Then, we study the performance of SVCCS and the contribution of each component of the codec. We also compare the performance of SVCCS and MJPEG in wireless video multicast.

2. Background

Suppose that a signal $x \in R^n$ can be transformed to a coefficient vector θ with some basis Ψ, i.e., $x = \Psi\theta$. Ψ can be any representing basis such as DCT or wavelet. The measurements of compressive sensing, $y \in R^m$, are obtained by multiplying signal x with a measurement matrix $\Phi \in R^{m\times n}$, i.e., $y = \Phi x$. Since $m < n$, (1) is an under-determined system with infinite solutions. Using the reverse operation in (1) to recover x is infeasible.

$$y = \Phi\tilde{x} \qquad (1)$$

However, [6, 2] have shown that $\ell_1$minimization may recover the original signal with high probability, which can be formulated as

$$\min \quad \|\tilde{\theta}\|_{\ell_1}$$
$$\text{subject to } \|y - A\tilde{\theta}\|_{\ell_2} \leq \epsilon \ , \qquad (2)$$

where $A = \Phi\Psi$ and ϵ is the noise energy in the measurements. Then the recovered signal $\hat{x}=\Psi^*\hat{\theta}$, where Ψ^* is the ajoint of Ψ and $\hat{\theta}$ is the solution to (2). In order to make recovery stable and accurate, sensing matrix A must satisfy the restricted isometry property (RIP).

Reference [1] showed the methods of generating sensing matrix holding RIP. One of them is to randomly select m rows from the Fourier matrix. When condition

$$m \geq CS(\log n)^4 \qquad (3)$$

is satisfied, the sensing matrix A obeys RIP with overwhelming probability, where C is a constant. There are four important observations which are exploited in this paper. 1) The sufficient condition (3) for RIP only cares about the number of rows of Fourier matrix instead of which row is selected. In other words, CS measurements are equally important. The property is also called democracy [8]. 2) Compressive sensing is scalable. The more measurements, the smaller recovery error is. 3) Given a fixed number of measurements, the faster the signal decays, the smaller the recovery error is. 4) The recovery error is proportional to noise energy ε. The noise may include quantization and transmission errors, which should be carefully managed.

3. Layered Coding Architecture and Performance

Fig 1 and 2 illustrate the proposed video encoder and decoder architecture, respectively. As shown in 1, video frames are divided into two categories, i.e., I frames and P frames. I frames are DCT transformed and coefficients are extracted in a zigzag order, then uniformly quantized and entropy encoded. Although the number of these coefficients is small, they contain the majority energy of the image. Therefore, after these coefficients are inversely quantized and inversely DCT transformed, the resultant image provides moderate image quality and can be used as a reference frame. Then the difference between the reference frame and the I or P frame, called the difference frame, is fed into the compressive sensing block.

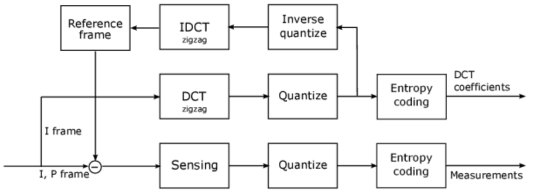


Fig. 1 Encoder

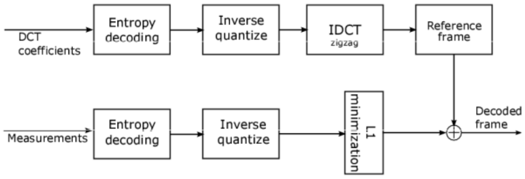


Fig. 2 Decoder

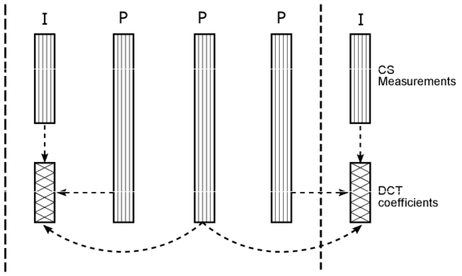


Fig. 3 GOP Structure

Fig. 3 shows the layered structure of frames. The GOP size is four. The arrowed dashed lines indicate the dependence between frames. From the figure, we can see that P frames are dependent on the closest I frame(s)' reference frame(s). The P frame in the middle is dependent on the average of the two reference frames to better exploit temporal redundancy.

We study the performance of wireless multicast with SVCCS. We compare the convolutional code protected MJPEG bitstream and SVCCS. 50 frames are encoded with MJPEG and SVCCS, respectively. The SVCCS encoded bitstream is obtained from the joint source and channel coding approach, so we do not apply channel coding. For MJPEG coded bitstream, we apply convolutional code (code rate is 1/2). After channel coding, the doubled average frame size of MJPEG is even larger than that of SVCCS; thus, SVCCS can take less channel bandwidth. We assume that the communication channel is AWGN and modulation scheme is DBPSK. Assume that the base layer of SVCCS can be correctly received. This assumption is acceptable as the base layer only counts for 2% of the coded bitstream, which can be protected with very low cost. The average PSNR of the base layer is 21.45 dB. Fig. 4 shows the advantage of SVCCS which is strongly adaptive to channel conditions.

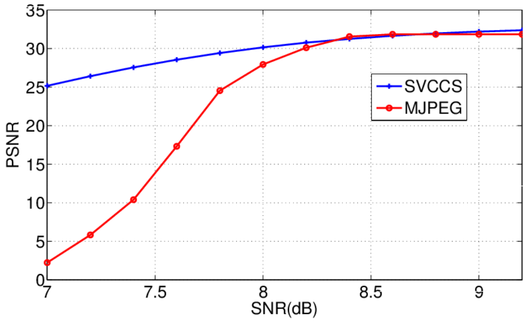


Fig. 4 PSNR vs. SNR

4. Conclusions

CS based video coding is overall a promising direction with many open issues that worth further investigation, e.g., how to optimize the quantization level and the bit allocation for each layer; how to reduce the decoding; how to enhance network protocols to support SVCCS with even lower cost and better performance.

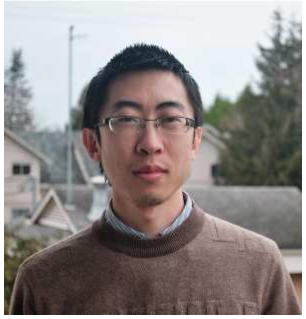
References

[1] E.J. Candes and T. Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? Information Theory, IEEE Transactions on, 52(12):5406–5425, Dec. 2006.

[2] Z. Charbiwala, S. Chakraborty, S. Zahedi, Younghun Kim, M.B. Srivastava, Ting He, and C. Bisdikian. Compressive oversampling for robust data transmission in sensor networks. In INFOCOM, 2010 Proceedings IEEE, pages 1–9, Mar. 2010.

[3] D.L. Donoho. Compressed sensing. Information Theory, IEEE Transactions on, 52(4):1289–1306, April 2006.

[4] V.K. Goyal, A.K. Fletcher, and S. Rangan. Compressive sampling and lossy compression. Signal Processing Magazine, IEEE, 25(2):48–56, Mar. 2008.
[5] J.N. Laska, P. Boufounos, M.A. Davenport, and R.G. Baraniuk. Democracy in action: Quantization, saturation, and compressive sensing. preprint, 2009.
[6] S. Pudlewski, T. Melodia, and A. Prasanna. C-dmrc: Compressive distortion-minimizing rate control for wireless multimedia sensor networks. In Sensor Mesh and Ad Hoc Communications and Networks (SECON), 2010 7th Annual IEEE Communications Society Conference on, pages 1–9, Jun. 2010.
[7] S. Xiang and L. Cai, "Scalable Video Coding with Compressive Sensing for Wireless Videocast," IEEE ICC'11, Kyoto, Japan, June 2011.



Siyuan Xiang is a Ph.D. Student with the Department of Electrical & Computer Engineering at the University of Victoria. He received his M.Eng. degree in Tongji University, Shanghai, China, in 2008. He is the recepient of China Scholarship Council (CSC) Scholarship. His research interest is multimedia communication.

Lin Cai is an Associate Professor with the Department of Electrical & Computer Engineering at the University of Victoria. She received her M.A.Sc. and PhD degrees (awarded Outstanding Achievement in Graduate Studies) in electrical and computer engineering from the University of Waterloo, Waterloo, Canada, in 2002 and 2005, respectively. Dr. Cai has been awarded the NSERC Discovery Accelerator Supplement Grant in 2010. She is a recipient of the best paper award in IEEE ICC'08 and the best academic paper award in IEEE WCNC'11, respectively. She has served as associate editor for IEEE Trans. on Wireless Communications, IEEE Trans. on Vehicular Technology, Journal of Communications and Networks, International Journal of Sensor Networks, and EURASIP Journal on Wireless Communications. She is a senior member of IEEE and a member of ACM.

Video Coding Platforms for Mobile Multimedia Networks

Tiago Dias, INESC-ID / IST-TU Lisbon / ISEL-PI Lisbon, Portugal
Nuno Roma, INESC-ID / IST-TU Lisbon, Portugal
Leonel Sousa, INESC-ID / IST-TU Lisbon, Portugal
{ Tiago.Dias, Nuno.Roma, Leonel.Sousa }@inesc-id.pt

1. Background and Motivation

Recently, mobile networking infrastructures have experienced profound changes owing not only to the expansion of the Internet to this specialized domain, but also to an ever increasing user demand for more innovative and better quality interactive multimedia services. As a result, several difficult challenges have been posed to network and computer architects, owing to the huge amounts of data that are processed in such class of services, as well as to its quite restrictive constraints in terms of processing rate, latency and QoS. Nonetheless, applications based on digital video services, like video telephony, Internet video streaming, or 3GPP IMS mobile multimedia telephony, are nowadays already supported up to some extent by existing mobile networks, in order to comply with the latest user requirements.

This important breakthrough results not only from all the technological innovations and novel techniques that have been successfully applied in multimedia networks over the last few years, but also from the proposal of newer video standards capable of providing high coding efficiency (e.g., H.264/AVC, AVS, VC-1) [1]. Moreover, several other advances have been achieved in the computer architecture domain (e.g., VLIW, SIMD and multi-threading organizations), which allowed to design state-of-the-art processors capable of fulfilling the huge computational requirements of these high complexity and data intensive coding algorithms. Even though, several different challenges still urge to be tackled in the design of current and future embedded systems for mobile multimedia applications, so that such portable and handheld battery supplied products are capable of supporting the next generation of high definition, and interactive video applications [2].

Video coding requirements for mobile applications are briefly presented in the following sections, as well as some of the commonly adopted solutions to comply with such requirements. Finally, future trends for the design of the next generation of network oriented multimedia embedded systems are also discussed.

2. Video Coding in Mobile Devices

Modern video standards (e.g., H.264/AVC [3]) are structured in a Video Coding Layer (VCL) and a Network Adaptation Layer (NAL). Such approach not only allows achieving high compression efficiency in the VLC, which defines the video representation, but also optimal “network friendliness” in the NAL. While the NAL is of crucial importance to transmit video over a large variety of networks, it is the VCL that poses most challenges to mobile system designers, especially for real-time operation [4].

These challenges mostly result from the limited computing power and reduced memory capacity of such embedded systems, which hardly comply with the computational requirements of the highly complex algorithms adopted in the block-based hybrid video coding scheme of the VLC (see Fig. 1). In such predictive coding approach, an image is processed in groups of pixels (macroblocks) using motion compensated temporal prediction based on Motion Estimation (ME). Then, they are transformed to the frequency domain, quantized and entropy encoded to reduce the amount of data to be transmitted.

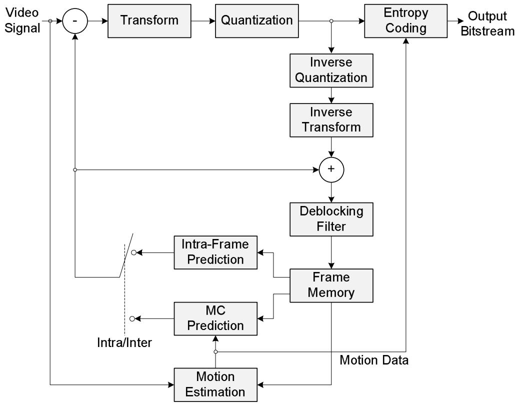


Fig. 1 – Generic block diagram of a video encoder.

Among these operations, the ME, the deblocking filtering, the intra prediction and transform coding processes are the ones that more severely constrain the implementation of video encoders and decoders in mobile systems [4]. As an example, ME requires about 80% of the total computational power of a video encoder, while the deblocking filtering contributes in about 32% to the complexity of the decoder. Hence, processors adopted in mobile embedded systems often include specialized or dedicated hardware structures to fulfill the processing of these operations.

3. Multimedia Architectures for Mobile Embedded Systems

Most current processors for mobile embedded systems consist of highly efficient Systems-on-Chip (SoCs), that not only offer several processing elements to implement the general purpose and several multimedia algorithms commonly used in mobile applications, but also to do the interface with the system peripherals (e.g., touch screen LCDs, video cameras, etc). Typically, these processing structures are of different classes, so that the flexibility and the performance levels required by such diverse algorithms can be efficiently met.

As it can be seen from Fig. 2, a General Purpose Processor (GPP) is always available in most multimedia mobile architectures to support the generic operations, as well as the tasks involving non-regular processing algorithms. While older designs included only a single processing core (i.e, the TI OMAP2 series, the Apple A4 or the Samsung Hummingbird processors), the most recent ones are already shared-memory multi-core architectures with either two (e.g., the Samsung Exynos, the Apple A5 and the Intel Moorestown processors to be used in the next generation of mobile phones) or four cores (e.g., the TI OMAP5 series processors). Additionally, such cores now operate at relatively higher clock frequencies (all above 1 GHZ), due to the increased multi-tasking processing and the more elaborate and compute intensive interactive applications that have been introduced.

Conversely, these SoCs often include several specialized processing cores, in order to efficiently support the realization of the most time and performance critical tasks. Among such operations are ME, intra prediction and filtering in video coding, whose underlying algorithms present different degrees of regularity. Consequently, dedicated processors with algorithm specific hardware architectures are often preferred for the implementation of the tasks with more regular data flow processing [5], while Application Specific Instruction set Processors (ASIPs) are generally chosen to support the remaining tasks [6]. Nonetheless, Digital Signal Processors (DSPs) are also often included in this class of heterogeneous structures, thus providing increased performance in a broader range of applications. For example, the TI OMAP series of processors includes a C64x DSP that allows it to efficiently support multiple video standards.

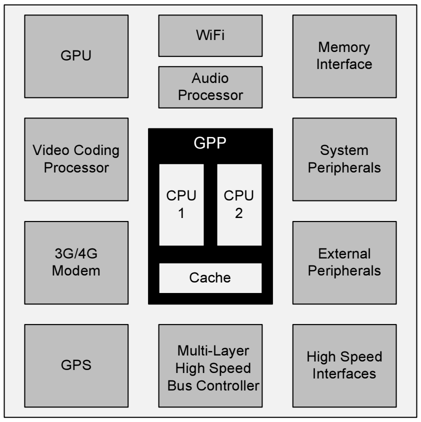


Fig. 2 – Block diagram of a typical mobile multimedia embedded system.

4. Future Trends

The forthcoming multimedia architectures for mobile embedded systems promise impressive performance levels, to account for all the new and even more demanding computational and bandwidth requirements of the next generation of multimedia applications that will be soon made available in mobile multimedia networks (e.g., HD video streaming, video conferencing, on-line gaming, and so on). However, other challenges still will also arise. Namely, the diversity and heterogeneity of the involved applications should require the dynamic adaptation of the hardware structures to the specific properties of the algorithms to be implemented at any given time instant. For example, the hardware structure of the embedded system can be reconfigured to support the playback of video sequences coded using different standards.

In this scope, silicon areas comprising reconfigurable logic are expected to become a very important feature of future embedded systems for mobile applications. Such processing structures shall still be based on heterogeneous multi-core architectures using multi-core GPPs and bus centric interconnection topologies based on multiple hierarchical buses. This approach shall not only significantly increase the flexibility of multimedia embedded systems, but also greatly improve its hardware efficiency and power consumption levels. These are very important factors in the design of portable and battery supplied multimedia devices for the electronic consumer market, since they directly influence both the cost and the usability of such products.

References

[1] Rao, K. R. and Do N. K., “Current video coding standards: H.264/AVC, Dirac, AVS China and VC-1,” *Proc. 42nd Southeastern Symp. System Theory (SSST)*, pp. 1-8, Mar. 2010.

[2] Minoru E., “Insights into future mobile multimedia applications,” Proc. 15th Int. Conf. Multimedia (MULTIMEDIA '07), pp. 851-851, Sep. 2007.

[3] T. Wiegand, G. Sullivan, G. Bjntegaard, and A. Luthra, “Overview of the H.264/AVC video coding standard,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 7, pp. 560-576, Jul. 2003.

[4] S. Saponara, C. Blanch, K. Denolf, and J. Bormans, “The JVT advanced video coding standard: complexity and performance analysis on a tool-by-tool basis,” *Proc. Packet Video Workshop (PV'03)*, Apr. 2003.

[5] Dias T., López S., Roma N. and Sousa L., “High Throughput and Scalable Architecture for Unified Transform Coding in Embedded H.264/AVC Video Coding Systems, to appear in *Int. Conf. Embedded Computer Systems: Architectures, Modeling and Simulation (SAMOS XI)*, Samos - Greece, Jul. 2011.

[6] Dias T., Roma N., Sousa L. and Ribeiro M., “Reconfigurable architectures and processors for real-time video motion estimation,” *Journal of Real-Time Image Processing - Special issue on Field-Programmable Technology*, Springer Berlin, vol. 2, n. 4, pp. 191-205, Dec. 2007.

Tiago Dias received the B.Sc. and M.Sc. degrees on Electrical and Computer Engineering in 2004 and 2006 from the Technical University of Lisbon, where he is now also pursuing his PhD degree. His research activities are being performed in the Signal Processing Systems Group (SiPS) of Instituto de Engenharia de Sistemas e Computadores - R&D in Lisbon (INESC-ID), where he has been since 2004. He is also an assistant lecturer at the Electronic and Telecommunications and Computer Engineering Department of the High Institute of Engineering of Lisbon (ISEL), Polytechnic Institute of Lisbon (IPL), where he lectures courses on embedded systems and computer architectures. His current research interests are specialized and reconfigurable architectures, as well as the design of multi-core embedded systems for video coding.

Nuno Roma received the Ph.D. degree in Electrical and Computer Engineering from Instituto Superior Técnico (IST), Technical University of Lisbon, Portugal, in 2008. He is currently an Assistant Professor with the Department of Computer Science and Engineering at IST, and a Senior Researcher of the Signal Processing Systems Group (SiPS) of Instituto de Engenharia de Sistemas e Computadores R&D (INESC-ID). His research interests include specialized computer architectures for digital signal processing (including biological sequences processing and image and video coding/transcoding), embedded systems design and compressed-domain video processing algorithms. He has contributed to more than 40 papers to journals and international conferences. Dr. Roma is a member of the IEEE Circuits and Systems Society and a member of ACM.

Leonel Sousa received the PhD degree in Electrical and Computer Engineering from Instituto Superior Técnico (IST), Technical University of Lisbon, Portugal, in 1996. He is currently a Full Professor of the Electrical and Computer Engineering Department at IST and a senior researcher at INESC-ID. He has contributed to more than 200 papers in journals and international conferences. He is an associate editor of the Eurasip Journal on Embedded Systems and served in the program committee of several conferences, for example as General Chair of ISPDC’09, Topic Co-Chair of SAMOS’10 and Topic Chair of Euro-Par’11. He is a senior member of both the IEEE and the ACM. His research interests include VLSI architectures, parallel and distributed computing and multimedia system.

Mobile Vision: Opportunities and Challenges

Gang Hua, IBM T. J. Watson Research Center
ghua@us.ibm.com

1. Introduction

Ten years ago, it was difficult to imagine that one day digital camera would become a standard peripheral of mobile phones. Nowadays, smart phones have not only been equipped with camera sensors, but also various other sensors such as accelerometer, gyroscope, and even GPS, etc.

The ubiquitous high quality phone cameras and the ever increasing computational capacity they have, along with other rich sensors equipped, provide a lot of new opportunities for accurate location based and user centered services using state of the art computer vision technologies.

This has largely fostered an emerging research area, namely *mobile vision*, in the past several years. Although it is still in its starting stage, the attention from both industry and academia are enormous. This is not a surprise since we are indeed in the era of transition from PC based computing to mobile and cloud based computing.

In this short paper, we will briefly summarize emerging research opportunities in mobile vision, and discuss about challenges we are facing.

2. Opportunities

Location based service, and mixed/augmented reality have been the buzz words in the past several years, where user centric visual computing is essential. We've observed a series of efforts in industry which is trying to leverage visual image recognition for these applications, including SnapTell [1], Nokia Point & Find [2], and Google Goggles [3].

In essence, all these applications start from a snapshot photo taken by the users using their smart phones. This photo will then be matched against an pre-annotated image database to extract useful information to be provide to the users, be it movie reviews, restaurant menus, and historical description of the buildings, etc..

The image search is often achieved by efficient indexing and matching of modern local image descriptors such as SIFT [4], DAISY[5], and CHoG[6], some of which leveraged efficient indexing and matching scheme such as visual vocabulary tree [7].

Mixed/augmented reality mobile video gaming is another area which has caught a lot of attention. Blair's group [8] has done extensive work in this space and there is a lot of potential to renovate the video game market.

With the increased quality and pixel resolution of smart phone cameras (e.g., Nokia N8 is equipped with a 12-megapixel camera sensor), one would wonder why we would need additional digital cameras. Besides, smart phone usually has much more computational capacity, which enables mobile computational photography on the fly, such as panoramic stitching [9].

Visual motion based gesture interface on mobile devices is another area that has been studied in the past years. The TinyMotion [10] and PEYE [] system are the representatives, where the mobile phone camera is utilized as a input sensor by capture the motion gesture based on light weight computer vision algorithms.

In addition to these four major sub-areas in mobile vision, there are also many other topics that may be of interest to the research community in different mobile and user centered experiences. Here we list some of them, including the four discussed above, without getting into more details:

- Mobile mixed/augmented reality
- Mobile and internet vision
- Mobile video games
- Mobile perceptual interfaces
- Multimodal vision system
- Mobile computational photography
- Mobile visual search
- Mobile visual computing for social networks
- Multiple-view analysis and 3D models
- Interactive visual parsing& annotation
- Mobile Soft-biometrics
- Emotion, gesture, expressiona analysis
- Audio-visual personal identification
- Mobile visual assistant

3. Challenges

Besides the common concern on the battery life of mobile devices in supporting of mobile vision applications, there are three prominent challenges in mobile vision when compared with traditional computer vision applications.

First of all, although the computational capacity of smart phones has become more and more powerful, it is still not sufficient to handle large scale visual computing tasks. For this perspective, migrating major of the computing into the cloud is essential. This links the mobile ecosystem with cloud computing. How to ensure real-time responses to users' interaction will be a major issue to be tackled.

Secondly, most of the mobile vision applications are driven by large amount of annotated visual data. For example, to enable vision based location recognition, large collection of street-view images is a prerequsite. How to acquire such data will be a challenge. One potential way is to harness the massive user generated visual data on the internet, such as those online image/video sharing systems. This links the mobile ecosystem with the internet ecosystem, where a large amount of users activities are driven by the enormous amount of visual data.

Last but not least, as a intrinsic human centered problem, how to leverage the rich contextual information in a visual computational model to make more robust mobile vision system and better satisfy the users' need and intention is a very important problem to research. Essentially the kind of semantic information the users would like to extraplate from the visual data may be evolving with the change of the context information. This is often refered as *emergent semantics*. It is not unreasonable to claim that all semantics in mobile vision applications are emergent – the modeling of which is difficult and hence a great problem to research on.

4. Conclusions

In this paper, we reviewed research opportunities or topics, as well as three major challenges, in an emerging research area, dubbed *mobile vision*. We believe that this emerging area will be evolving with the interaction among the mobile ecosystem, the internet ecosystem, and the computing cloud.

References

[1] http://www.snaptell.com/
[2] http://pointandfind.nokia.com/main_publisher

[3] http://www.google.com/mobile/goggles/#text

[4] D. G. Lowe, "Distinctive Image Features from Scale Invariant Keypoints", *IJCV*, 2004
[5] S. A. J. Winder, G. Hua, and M. A. Brown, "Picking the Best DAISY", *CVPR*, 2009
[6] V. Chandrasekhar, G. Takacs, D. M. Chen, S. S. Tsai, R. Grzeszczuk, and B. Girod, "CHoG: Compressed histogram of gradients A low bit-rate feature descriptor", *CVPR*, 2009
[7] D. Nistér and H. Stewénius, "Scalable Recognition with a Vocabulary Tree". *CVPR*, 2006
[8] http://www.augmentedenvironments.org/lab/
[9] Y. Xiong and K. Pulli, "Fast image stitching and editing for panorama painting on mobile phones," in IEEE Workshop on Mobile Vision, 2010
[10] J. Wang and J. Canny, "TinyMotion: Camera Phone Based Interaction Methods", in ALT.CHI of ACM CHI, Montreal, Canada, 2006
[11] G. Hua, T-Y. Yang, S. Vasireddy: PEYE: Toward a Visual Motion Based Perceptual Interface for Mobile Devices. ICCV-HCI 2007.

Dr. Gang Hua is a Research Staff Member at IBM Watson Research Center. Before that, he was a Senior Researcher at Nokia Research Center, Hollywood (09-10), and a Scientist at Microsoft Live Labs Research from (06-09). He received the Ph.D. degree in Electrical and Computer Engineering from Northwestern University in 2006. He is an Associate Editor of IEEE Trans. on Image Processing, an Associate Editor of IAPR Journal of Machine Vision and Applications, a Guest Editor of IEEE Trans. on Pattern Analysis and Machine Intelligence on *Real-world Face Recognition*, and a Guest Editor of International Journal on Computer Vision on *Mobile Vision*. He has also served as Chairs, TPC Members, or Reviewers for a large number of prestigious international conferences and journals. In particular, he is an Area Chair of IEEE International Conf. on Computer Vision, 2011, an Area Chair of ACM Multimedia 2011, and a Workshops and Proceedings Chair of IEEE Conf. on Face and Gesture Recognition, 2011. He received the Richter Fellowship and the Walter P. Murphy Fellowship from Northwestern University in 2005 and 2002, respectively. He is a member of IEEE and a member of ACM. He is the author of more than 50 peer reviewed publications in prestigious international journals and conferences. As of March, 2011, he holds 2 US patent and has 18 more patents pending.

Satellites Networking: Research Trends and Open Issues

Junfeng Wang and Wei Dong, College of Computer Science, Sichuan University, Chengdu, P.R. China
wangjf@scu.edu.cn

The global Internet has experienced explosive growth in the past decades. Moreover, the proliferation of bandwidth-intensive multimedia services and massive expansion in the number of serviced users impose new challenges to the development of the global Internet. With the capacity of providing ubiquitous connectivity seamlessly and continuously to any part of the world, satellite network will become an integral part of next generation networks (NGNs).

While the first generation of satellite systems (e.g., Iridium, ICO, Globalstar, Odyssey) proposed in the early 1990s only provide voice service and low-speed data traffic [1], the second generation of satellite communications (SATCOM) networks have been proposed with the aim of offering high-speed Internet access and multimedia information services. Astrolink, Cyberstar, Spaceway, SkyBridge, Teledesic, and iSky are among this generation of satellite communication networks [2]. These satellite networks will provide a wide range of services such as video on demand, multimedia data, high-speed Internet access, interactive video, as well as other existing Internet-based applications. However, the need to support these multimedia services places new challenges on satellite systems and networks.

In this letter, we discuss the key issues and research trends in satellite networking.

Topological dynamics

With the merits of shortened propagation latency and smaller signal attenuation needed compared with geostationary (GEO) satellite systems, low earth orbit (LEO) satellite networks have attracted much attention from academic and industrial communities.

However, LEO based networks suffer from the topological dynamics considerably [3][4]. First, topological dynamics is crucial to the design of LEO constellation. With a better understanding of the dynamics, satellite constellations can be designed and optimized with the least topological dynamics [5]. Second, the underlying topological features of satellite networks are of great importance in routing protocol development. Current routing protocols for satellite networks adopt either on board routing scheme or dynamic routing protocol. In the former case, system period is divided into a serial of snapshots in which a static topology is assumed and a routing table is pre-computed corresponding to each snapshot. Consequently, the length of the snapshot, i.e., the dynamics, determines the amount of memory to store the routing table. On the other hand, if dynamic routing protocol is utilized, frequent topological changes will incur significant signaling overhead. Moreover, understanding of LEO satellite topological dynamics is also the basis for predictable mobility management and satellite link handover management for quality of service (QoS) guarantees.

In [5], the dynamical activities of regular LEO satellite network topologies are systematically quantified, and the number and length of network snapshots are formulated concisely. However, the dynamics of random networks (e.g., hybrid multi-layered satellite networks) still deserves further study.

Handover

Handover in satellite systems is due to the asynchronous movement of the satellite relative to Earth. From the point of view of the network layer, handovers can be classified into three categories: link-layer handover, network-layer handover, and transport-layer handover.

While link-layer handover schemes have been investigated in depth (refer to [6] and the references therein for more information), there are few research conducted about the other two handover schemes. Due to the movement of the satellites, the communication endpoints may have to change their IP address, resulting in network handover. Mobile IP [7][8] is an IETF standard to cope with this problem, and it allows nodes to maintain all ongoing communications while moving. However, Mobile IP suffers from a number of drawbacks such as high handover latency, high packet loss rate, requirement for infrastructure change [9][10]. Furthermore, the IP address change may also impose an impact on TCP performance, and [11] presented some results by simulation.

Consequently, future research should focus on network-layer handover schemes and transport layer schemes. More precisely, more intelligent handover schemes should be designed in such way that the handover process has few impacts on the network layer and transport layer.

Routing

Classical terrestrial Internet routing protocols, such as Open Shortest Path First (OSPF) and Routing Information Protocol (RIP), update the routing table based on the exchanged topological dynamics between neighbor routers. Therefore, these protocols are not suitable for satellite networks because some satellite systems, such as LEO constellations, experience frequent topological changes. Applying such protocols in satellite network will incur substantial overhead.

The usual solution is that the periodic and predictable nature of the constellation topology is utilized when routing protocols are designed. Based on the periodicity, network state can be seen as a finite state machine (FSM) and a routing table corresponding to each state is calculated beforehand. When the network shifts from one state to a new state, the corresponding routing table is loaded accordingly. However, this approach is not scalable. Moreover, a fixed routing table is utilized and current traffic status is seldom considered when such routing protocols are designed. More flexible routing protocol should be developed to cope with above-mentioned problems.

Reliable multicast routing

Many applications take advantage of multicast delivery, such as video on demand, interactive video, tele-education. Considering the merits of the broadcast nature of satellite links, supporting multicast service in satellite networks are more natural and attractive.

However, there are several challenges to be faced when multicast protocols are designed for satellite networks. *Scalability* is one of the issues to be considered. While a simple multicast application sends data to only a group of receivers, in some application there may be hundreds of receivers. Moreover, with explosive multimedia traffic, the demand for supporting tens of millions of receiver in a single will be necessary.

Recently, applying network coding to satellite networks has been attracted a lot of interests both from the academic and industry. The advantages of Network Coding include increasing the network's throughput, saving bandwidth, and providing load balancing to the networks, reliability improvement over lossy link. With an emphasis on multicast application, these merits have been demonstrated in a lot of research papers [12][13][14]. However, most of the researches assume specific network topologies. How to apply network coding to satellite networks deserves future research.

Enhanced TCP

While TCP performs quite well in wired networks, it suffers from significant performance degradation due to the characteristics of satellite links. To improve the performance of TCP in satellite networks, several problems must be solved.

First, a mechanism should be developed to differentiate between the reasons of packet loss. TCP interprets all packet drops as signals of congestion and reduces its transmission rate to mitigate the congestion , and this situation can be even worse because satellite channels exhibit a higher bit error rate (BER) than typical terrestrial networks.

Second, TCP source increments its window too slowly in slow start phase, and decrements it too drastically responding to congestion. If the window in the slow phase is too slow, the bandwidth of satellite link is not fully utilized. On the other hand, because the window reduction is too drastic when responding to congestion, it will take a long time to recover from a single packet loss. These inefficiencies are aggregated because satellite network has a large bandwidth-delay product.

Third, the RTT dynamics of satellite network maybe affect the performance of TCP. TCP throughput is limited by the following formula:

$$Throughput_{\max} = window\ \ size / RTT \quad (1)$$

In some satellite environments the RTT varies over time due to topological dynamics of satellite network. For instance, in LEO constellation the propagation delay to and from the satellite varies over time, resulting in variable RTT. From Equation (1), we can see that the dynamics of RTT may have a major impact on throughput performance of TCP.

The first two problems of TCP in satellite networks are fully understood and many schemes are proposed to solve these problems [15][16][17]. However, how the RTT dynamics affect the performance of TCP needs further study.

Another research trend is the design of multi-path transmission protocol. Since in satellite networks, especially in non-GEO constellations, there are typically several equivalent routing path of next hop when a packet is forwarding, multi-path transmission is attractive. However, there are some challenges to develop multi-path transmission protocol. First of all, the compatibility problem must be taken into account. Since TCP is the dominant transmission protocol, new multi-path transmission protocol must be compatible with TCP. In addition, since multi-path transmission protocol may rely on routing ability of lower layer (i.e., network layer), cross-layer design methods are probably employed to design more efficient protocol. As for the multi-path protocol design itself, the focus should be on the resilience of the protocol. For instance, the protocol should be aware of routing situations where one routing path becomes unavailable. Furthermore, with the current traffic status of each paths considered, the protocol should distribute the traffic on all available paths evenly (e.g., the amount of distributed traffic on each path is proportional to the available bandwidth). At last, since RTT delay of each path may be different, the SACK option should be enabled to cope with out of order problem.

Acknowledgements

This work is supported by the National Basic Research Program of China (2007CB310702); the National High-Tech Research and Development Plan of China (2008AA01Z208); the Doctoral Fund of Ministry of Education of China (20090181110053); the Applied Basic Research Program of Sichuan Province of China (Grant No. 2010JY0013); the Youth Foundation of Sichuan Province (09ZQ026-028).

References

[1] Y. Hu and V.O.K. Li, “Satellite-based internet: A tutorial,” IEEE Communications Magazine, vol. 39, no. 3, pp. 154–162, Mar. 2001.

[2] J. Farserotu and R. Prasad, “A survey of future broadband multimedia satellite systems, issues and trends,” IEEE Communications Magazine, vol. 38, no.6 pp. 128–133, Jun. 2000.

[3] A. Ferreira, J. Galtier, P. Penna, “Topological design, routing and hand-over in satellite networks,” Handbook of Wireless Networks and Mobile Computing, pp. 473-507, Feb. 2002.

[4] J. Sun, E. Modiano, “Routing strategies for maximizing throughput in LEO satellite networks”, IEEE Journal on Selected Areas in Communications, vol. 22, no. 2, pp. 273–286, Feb. 2004.

[5] J. Wang, L. Li, M. Zhou, “Topological dynamics characterization for LEO satellite networks,” Computer Networks, Apr. 2006.

[6] P. K. Chowdhury, M. Atiquzzaman, andW. Ivancic, “Handover Schemes in Satellite Networks: State-of-the-Art and Future Research Directions,” IEEE Communications Surveys, vol. 8, no. 4, 4th Quarter 2006.

[7] D. Johnson, C. Perkins, J. Arkko, “Mobility Support in IPv6,” RFC 3775, 2004.

[8] C.E. Perkins, “IP Mobility Support.” IETF RFC 3344, Aug. 2002.

[9] I.W. Wu, W.S. Chen, H.E. Liao, and F.F. Young, “A seamless handoff approach of Mobile IP protocol for mobile wireless data networks,” IEEE Transactions on Consumer Electronics, vol. 48, no. 2, pp. 335–344, May 2002.

[10] S. Fu et al., “Architecture and Performance of SIGMA: A Seamless Handover Scheme for Data Networks,” IEEE ICC, Seoul, South Korea, pp. 3249–3253, May 2005.

[11] H. Huang and J. Cai, "Improving TCP performance during soft vertical handoff," Proc. IEEE AINA'05, Taipei, Taiwan, Mar 2005.

[12] R. Ahlswede, N. Cai, S.-Y. R. Li, and R. W. Yeung, “Network information flow,” IEEE Trans. on Info. Theory, vol. 46, no. 4, pp. 1204–1216, Jul 2000.

[13] Y.Wu, P.A. Chou, and S.-Y.Kung, “Information exchange in wireless networks with network coding and physical-layer broadcast,” Microsoft Research, Tech. Rep. MSR-TR-2004-78, Aug. 2004.

[14] C.Fragouli, J.Widmer, and J.-Y.L.Boudec, “On the benefits of network coding for wireless applications,” Proc. WiOpt, Boston, USA, pp. 1–6, Apr. 2006.

[15] J. Chen, L. Liu, X. Hu, F. Xu, “Hop-by-hop transport for satellite networks”, Proceedings of IEEE Aerospace conference, Big Sky, Montana, USA, pp. 1-7, 2009.

[16] Akyildiz I. F., Morabito, G., Palazzo, S.., “TCP-Peach: a new congestion control scheme for satellite IP networks”, IEEE/ACM Transactions on Networking, vol. 9, no. 3, pp. 307-321, 2001.

[17] Gerla, M., Sanadidi, M.Y., Ren Wang et al., “TCP Westwood: congestion window control using bandwidth estimation”, Proceedings of IEEE Global Telecommunications Conference, San Antonio, TX, USA, pp. 1698-1702, 2001.

Junfeng Wang received the M.S. degree in Computer Application Technology from Chongqing University of Posts & Telecommunications, Chongqing in 2001 and Ph.D. degree in Computer Science from University of Electronic Science and Technology of China, Chengdu in 2004. From July 2004 to August 2006, he held a postdoctoral position in Institute of Software, Chinese Academy of Sciences. From August 2006, Dr Wang is with the College of Computer Science, Sichuan

University as a professor. His recent research interests include spatial information networks, network and information security, and intelligent transportation system.

Wei Dong received the M.S degree in communication and systems from Electronic Science and Technology of China, Chengdu. He is currently a Ph.D student at the College of Computer Science, Sichuan University. His research interests cover a wide variety of topics in wireless, and satellite networks, with emphasis on seamless handover schemes, and design of transport layer protocols for wireless networks.

CALL FOR PAPERS

IEEE Multimedia Communications Workshop (MMCOM) 2011

Date: December 5 or 9, 2011

Place: Houston, Texas, USA (to be held at IEEE Globecom 2011)

Web Link: **http://committees.comsoc.org/mmc/CFP_MMCOM%202011.pdf**

Organizers:
Prof. Thomas Magedanz, TU Berlin, Germany (thomas.magedanz@tu-berlin.de)
Prof. Jiangtao (Gene) Wen, Tsinghua University, China (jtwen@tsinghua.edu.cn)
Dr. Xiaoli Chu, King's College London, UK (xiaoli.chu@kcl.ac.uk)
Prof. Yung-Hisang Lu, Purdue University, USA (yunglu@purdue.edu)