Mobile Media Communication, Processing, and Analysis: a review of recent advances

Wen Gao, Ling-Yu Duan*, Jun Sun
The Institute of Digital Media, the School of EE &CS,
Peking University, Beijing, China
{wgao, lingyu, sunjun}@pku.edu.cn

Junsong Yuan, Yonggang Wen,
Yap-Peng Tan, Jianfei Cai, Alex C. Kot
Nanyang Technological University, Singapore
{jsyuan, ygwen, EYPTan, ASJFCai, EACKOT}@ntu.edu.sg

***Abstract*—In this paper, we review recent advances in mobile media communication, processing, and analysis. To identify the opportunities and challenges in fast growing mobile media computing, we discuss several emerging topics including mobile visual search, retargeting, mobile video streaming, and cloud based mobile media computing. According to the infrastructure of mobile devices vs. servers, we come up with essential concerns in mobile media computing such as wireless bandwidth consumption, mobile energy saving, media adaptation for better quality of services, the computational load shift from mobiles to servers, etc. With booming mobile Apps on diverse media consumption, it is envisioned that mobile media research and development is bringing about significant achievements in traditional topics of communication, processing, and analytics.**

I. INTRODUCTION

The rapid progress in mobile devices (smart phone, tablets), wireless communication (3G/WIFI) and multimedia processing in the past decade has unleashed a broad spectrum of emerging mobile applications such as visual search and augmented reality, scalable video coding, image/video retargeting, etc. Beyond great benefits from mobile computing, it is envision that we need to address quite a few challenges in mobile media communication, processing and even analysis to facilitate the pervasive creation and consumption of mobile media content seamlessly. In this paper, we will review recent advances and topics in mobile media research and development as illustrated in Figure 1. First, we discuss mobile visual search (MVS) and challenging issues in wireless environment. Distinct from traditional visual analytics, mobile visual search attempts to combine visual features extraction and compression to tackle the issues of wireless network constraints and mobile battery consumption. Second, we present research efforts in image/video retargeting and scalable video coding, which are to alleviate the adaptation and streaming issues of mobile media. To further support large scale image/video processing and mobile computing infrastructure, a cloud computing platform is important, which will be discussed as well.

II. Mobile Visual Search

Smart phones and Tablet PCs have evolved into powerful image and video processing devices, due to high resolution color cameras and application specific integrated circuits (ASIC) embedded units. With GPS assisted location based services and broadband wireless networks connection, mobile devices have shown great potentials in

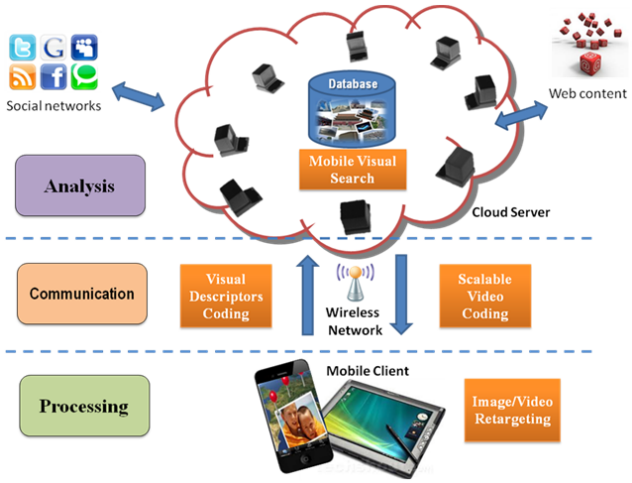

Figure 1. Mobile Media Communication, Processing, and Analysis

visual search and augmented reality applications, such as CD/Book cover search, location recognition, scene retrieval, product search, etc.

Existing mobile visual search systems like Google Goggles [1], Nokia Point & Find [2], Kooaba [3], Ricoh iCandy [5] and Amazon Snap2tell [7] follow the client-server architecture. For the mobile end, the retrieval pipeline needs to reduce memory usage (less than 1MB), power consumption, and in particular the upstream query delivery latency. It is easy to imagine the significantly reduced amount of data transmitted over wireless network may reduce network latency and improve user experience. At the sever end, the retrieval process shall be very efficient in order to scale up to large image datasets.

In 3G wireless environments, the visual query delivery is subject to relatively slow or unstable mobile network. The quality of user experience heavily depends on how much information to transmit. This issue becomes even more crucial in streaming augmented reality applications. Distinct from traditional visual communication, the time consuming delivery of query images is unnecessary, as MVS aims to perform visual search rather than precise reconstruction of query images at the server end. With the ever growing processing power in mobile devices, recent works have proposed to extract compact visual descriptors of query images directly at the mobile end [8] [9] [10] [11] [12], and then send such descriptors over a wireless link at low bit rates (say 512B, 1KB, 2KB per query). The descriptors are expected to be compact, discriminative, and meanwhile efficient in extraction to reduce the overall query delivery latency as much as possible. In particular, the ongoing MPEG standardization of compact descriptors

* Ling-Yu Duan is the corresponding author.

for visual search (CDVS) have involved big industry and academia efforts from STMicroelectronics, NEC, Nvidia, Samsung, Nokia, Qualcomm, Stanford Univ., and Peking Univ. etc. [13].

Aside from low latency query delivery, such on-device extraction of compact visual descriptors can significantly lighten the overload of network communication at the server end. Different from text based search engine, online visual search systems have to deal with similar challenges from bandwidth consumption at the server end, when receiving tens of thousands of concurrent image queries. How to optimize the network bandwidth usage in scalable visual search remains unexploited in depth. More importantly, in mobile media communication and processing, battery consumption would probably become an important bottleneck of advanced mobile computing. Power saving is a critical and challenging issue as well. Undoubtedly, upstream sending an entire image or high-dimensional signature brings about serious energy consumption in 3G wireless network.

State-of-the-art local descriptors (e.g. SIFT [14], SURF [15], and PCA-SIFT [16]) are over size, as sending hundreds of these local descriptors costs even more budget than original images. On the other hand, for zero latency query delivery, state-of-the-art compact local descriptors [9] [10] are not compact enough. For instance, CHoG descriptor [9] costs 50 times n bits, where n is the number of local features per image. Intuitively, an extremely "compact" solution was proposed to compress the local descriptors into a Bag-of-Words histogram [8] [43]. However, local descriptor information was completely ignored, which degrades retrieval performance as local descriptors based geometric verification cannot be accomplished. Recent development has shown that an effective solution is to leverage compact global signature and local descriptors [44].

III. IMAGE/VIDEO RETARGETING

Like low bit rate visual search, mobile media consumption puts more emphasis on improving user experiences. From the media processing point of view, image/video retargeting is another typical application. For example, users shrink or stretch the display size of image/video to adapt to the screen of mobile devices. When the size or aspect ratio of target display is quite different from the original one, simply adding black bars or uniform scaling would bring about unpleasant viewing experiences. Adding black bars cannot use the entire display, while uniform scaling may distort the salient visual content. To maximize the viewing quality, content-aware retargeting techniques emerge, aiming to preserve important visual contents at the expense of distorting less important content. Below we review retargeting techniques for adapting image/ video to different display.

Retargeting relies on content analysis and processing techniques. Users are sensitive to any noticeable distortion of retargeted images, so image retargeting needs to preserve the shape of salient objects and prevent the discontinuity of unimportant regions. To maximize viewing experience, content-aware image retargeting addresses two challenging issues: (1) determining importance of regions or pixels in an image, i.e. saliency models; (2) efficient and effective resizing of non-uniformly distorted regions or pixels.

State-of-the-art content-aware image retargeting methods include cropping [18], seam carving [19], multi-operator [20], mesh-based retargeting [21]. Cropping [18] employs an attention model to detect important regions and crop out the region for display. Seam carving [18] tries to carve a group of optimal seams iteratively based on the energy map of images. Rubinstein et al. [20] proposes to combine different retargeting methods including uniform scaling, cropping and seam carving. In addition, mesh-based methods [21] partition source images by meshes, where deformation is allowed by adjusting the shape of meshes. Important regions' mesh based shape is kept well.

Comparing to image retargeting, video retargeting is much more challenging, as video involves rich motions. Besides the performance in spatial domain, video retargeting shall consider the performance in temporal domain. For instance, we need to maintain consistent shape transformation of an object within consecutive frames, but have to avoid motion artifacts like shaking and flickering in temporal domain. Beyond image retargeting, video retargeting needs to address more challenging issues: (1) modeling temporal constraints, which align regions/pixels across neighboring frames and elegantly constrain the transformation of temporally aligned regions/pixels; (2) modeling temporal saliency to determine persistent important visual content across frames; (3) seeking a performance tradeoff between spatial and temporal dimensions; and (4) improving computational efficiency.

Research efforts have been devoted to those issues. For example, video cropping [22] [23] employ temporal constraints to smooth the trajectory of copped rectangular windows. Unfortunately, scene context and salient objects may be spatially discarded. Moreover, cropping introduces virtual camera motion. References [24] [25] [28] introduce temporal constraints by restricting the transformation of temporally adjacent pixels/regions. Rubinstein [25] extends the seam carving method [19] to videos by imposing temporal constraints.

To alleviate temporal artifacts [24] [25] from complex or strong motions, pixel-based [26] [27] and mesh-based video warping [28] [29] [34] have been proposed to estimate motions, by which saliency map or constraints may be constructed. For example, Kraehenbuhl et al. [26] detects moving areas by optical flow and averages the moving areas' importance maps of consecutive frames to form the temporal saliency map. However, when the temporal window could not exactly align with the movement duration of objects, temporal artifacts would appear. Yen et al. [27] employs panorama mosaic to model temporal saliency map and constraints.

On the other hand, Wang et al. [28] propose a mesh-based video warping to model temporal constraints, which aligns the meshes within neighboring frames by estimating object and camera motions. However, strong temporal constraints in the cases of intense camera or object motion would lead to missing unimportant regions, thereby producing distortion of salient objects. As a result, Wang et al [29] proposed a combination of cropping and mesh-based video warping.

State-of-the-arts content aware video retargeting methods are computationally expensive. Hardware acceleration like GPU has been tried out [26] [29]. In [30], a novel FPGA based architecture was proposed to deal with computational complexity.

In summary, on-device image/video adaptation has posed grand challenges in terms of low complexity image/video analysis and processing. A feasible strategy is to move the higher computational complexity from the mobile end to the server end. That is, media is actually stored and processed on the server. With protocols, mobile users take advantage of the remote super computing infrastructure to accomplish complicated image/video retargeting and subsequently download the resulting image/video to mobile devices.

IV. SCALABLE VIDEO CODING

Recently, fast growing 3G or WIFI network has witnessed the big impact of mobile video streaming on traditional video application scenarios such as video broadcast/unicast, video conferencing, video surveillance, etc. To extensively deploy such services, video content may be delivered to different terminals through various and variable transmission channels. Nowadays video streaming has attracted lots of research and development in mobile media communication. Mobile video streaming needs to address challenging issues inherent to mobile devices and mobile network. Anyway, mobile streaming allow consumers to watch video anywhere and anytime, and is becoming a more and more popular way to consume video content.

Compared to traditional video streaming, mobile streaming has proposed big challenges, arising from rich and expanded application scenarios. One big challenge comes from the wide range of user environments. First, portable devices (e.g., laptops, tablets and smart phones) are with different screen resolutions or CPU processing power. Clearly, it is infeasible to achieve the best quality of service (QoS) for the users of different devices through deploying a common video stream. Second, the transmission of mobile streaming data is subject to unstable wireless networks and even the switch of channels (for example, from Wi-Fi to 3G network), so that the bandwidth fluctuation is inevitable. When video is streamed at a constant bit rate, it would be difficult, if not impossible, for the client to maintain an uninterrupted and smooth video playing. In summary, to improve viewing experiences for different users, mobile video stream has to be adaptable with a terminal's capability and connection status.

To address the challenging issues in mobile streaming, scalable Video Coding (SVC) [31] has provided a workable solution. As an extension of H.264/AVC [32], SVC allows mobile users to implement efficient and standard compatible scalability in terms of video frame rate, spatial resolution, and SNR quality. SVC encodes video streams with multiple layers, namely, one base layer for basic video service and several enhancement layers for improved video quality. By discarding the packets of enhancement layers, the bit stream can be generated as requested to deploy video streaming services according to the client capability and network environment.

To allow SVC elegantly work in mobile video streaming [33] [34], we need to address several practical issues. First, optimizing SVC coding efficiency and complexity is a top priority. To support scalability, SVC coding efficiency is lower than H.264/AVC single layer encoding while the computational complexity is much higher. It is necessary to further improve the coding efficiency and reduce the complexity for mobile applications. Second, optimizing SVC bit extraction is necessary. When the enhancement packets are discarded, the bit selection is supposed to be optimal in terms of rate-distortion (R-D). To this end, the concept of Quality Layer is proposed and implemented in JSVM [35]. The basic idea is to assign a priority value to every packet according to its R-D impact. In bit extraction, those packets with lower priorities are discarded first. Similar work has been done in [36] [37] [38], focusing on distortion modeling and bit extraction of SVC. At the streaming server, one remaining issue lies in the SVC scheduling strategy or quality control algorithm. The server needs to make decisions on how to adjust video quality dynamically according to available bandwidth and client conditions. Relevant research work can be found in [39] [40].

Generally speaking, the priority assignment and quality control are performed at the server side, so no more computational burden is incurred at the client side. This allows for lightweight client terminals, which exactly reflects the trend of "cloud computing". To fulfill mobile streaming, the shift of computational complexity from mobile devices to powerful servers is crucial. Under SVC framework, streaming servers play a active role in customizing video services for mobile clients according to their individual capabilities and condition, hence the overall user satisfaction is reached

V. CLOUD COMPUTING PLATFORM

As discussed above, mobile media cloud computing is becoming an important computing paradigm to support mobile media services

Mobile media cloud computing can be defined as a novel computing model whereby the media analysis, processing and storage are shifted from mobile devices to centralized computing platforms in the cloud [41]. The communication between a mobile device and the cloud are normally based on wireless channels (e.g., 3G, WLAN). Terminals on mobile devices can be thin native client or web browser.

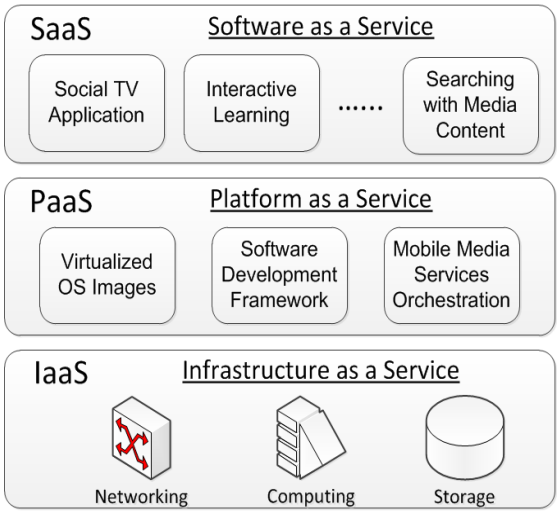

Figure 2. The layered model of mobile media cloud computing

As a natural extension of cloud computing, mobile media cloud computing can be interpreted in a layered service model originating from the cloud computing [42]. Basically, this platform consists of three layers in a bottom up order, IaaS (Infrastructure as a Service) layer, PaaS (Platform as a service) layer, and SaaS (Software as a Service) layer as shown in figure 1. In this new paradigm, the operational cost of rich media services over mobile networks can be significantly reduced; meanwhile, the user experience of mobile media consumption can be substantially improved.

At the IaaS layer, all of physical resources including storage, networking and computing resources are virtualized as a unified resource pool. These resources can be exposed to media applications with specified service level agreement (SLA) or user demand. The key challenge in this layer is the resource management problem, namely, how to properly schedule the available resources in the cloud infrastructure, and to efficiently allocate appropriate portions for individual users or applications as requested.

The PaaS layer offers an integrated platform to build up and deploy mobile media applications and services based on the cloud infrastructure. This layer typically consists of a set of APIs, which provide interfaces to dramatically utilize the virtualized resources, and orchestrates mobile media services. The key challenges in this layer are three-fold, i.e., service discovery, service orchestration and security issues. Service discovery is to map requested media services to available resources in an optimal fashion. Service orchestration is expected to coordinate different media services from the perspectives of both internal system and external service entry points. Finally, data protection mechanism is required to protect the user privacy data and solve other security issues.

The SaaS layer supports mobile media consumption directly from end users on their mobile devices such as social TV applications etc. This layer contains a wide variety of software deployments that work under the infrastructure of mobile media cloud computing. The key challenge in this layer may be attributed to the guarantee of desirable quality of experience (QoE) at user end. In particular, significantly reducing the energy consumption of mobile devices as well as the latency from application execution deserves more research efforts to enhance the QoE.

VI. CONCLUSIONS

As mobile devices proliferate, media consumption is frequently performed on these devices nowadays. Recent advances have shown that mobile media communication, processing and analytics will be

built upon cloud services. Hence, the research and development is of upmost importance in boosting the competitiveness of traditional media techniques in cloud computing. Exemplar mobile applications like low bit rate mobile visual search, scalable video coding, retargeting are well aligned with the goal of mobile media platform upon cloud services. However, for concrete domains, we usually have to spot domain specific challenges in depth. For example, the recently initiated **ROSE** (Rapid-Rich Object SEarch Lab) project between Peking Univ. (China) and Nanyang Technological Univ. (Singapore) identifies key thrusts in the near future of mobile visual search, with a vision to build the largest collection of structured domain object database in Asia and to develop a rapid and rich object mobile search.

ACKNOWLEDGEMENT

This work was supported in part by the Chinese Natural Science Foundation under Contract No. 61271311 and No. 61210005, and in part by the Rich and Rapid Object Search (ROSE) Project.

REFERENCES

[1] Google Goggles, http://www.google.com/mobile/goggles/.

[2] Nokia Point and Find, http://www.pointandfind.nokia.com.

[3] Kooaba, http://www.kooaba.com.

[4] B. Erol, E. Ant´unez, and J. Hull, "Hotpaper: multimedia interaction with paper using mobile phones," in Proc. of the 16th ACM Multimedia Conference, New York, NY, USA, 2008.

[5] J. Graham and J. J. Hull, "Icandy: a tangible user interface for itunes," in Proc. of CHI '08: Extended abstracts on human factors in computing systems, Florence, Italy, 2008.

[6] J. J. Hull, B. Erol, J. Graham, Q. Ke, H. Kishi, J. Moraleda, and D. G. Van Olst, "Paper-based augmented reality," in Proc. of the 17th International Conference on Artificial Reality and Telexistence, 2007.

[7] SnapTell, http://www.snaptell.com.

[8] D. Chen, S. Tsai, V. Chandrasekhar, G. Takacs, J. Singh, and B. Girod. Tree histogram coding for mobile image matching. In DCC, 2009.

[9] V. Chandrasekhar, G. Takacs, D. Chen, S. Tsai, R. Grzeszczuk, and B. Girod. Chog: Compressed histogram of gradients a low bit-rate feature descriptor. In CVPR, 2009.

[10] V. Chandrasekhar, G. Takacs, D. Chen, S. Tsai, J. Singh, and B. Girod. Transform coding of image feature descriptors. In VCIP, 2009.

[11] M. Makar, C. Chang, D. Chen, S. Tsai, and B. Girod. Compression of image patches for local feature extraction. In ICASSP, 2009.

[12] B. Girod, V. Chandrasekhar, D. Chen, N.-M. Cheung, R. Grzeszczuk, Y. Reznik, G. Takacs, S. Tsai, and R. Vedantham. Mobile visual search. In IEEE Signal Processing Magazine, 2011.

[13] CDVS1."Call for proposals for compact descriptors for visual search," N12201. Turin, Italy: ISO/IEC JTC1/SC29/WG11, 2011.

[14] D. Lowe. Distinctive image features from scale invariant keypoints. In IJCV, 2004.

[15] H. Bay, T. Tuytelaars, and L. Van Gool. SURF: Speeded up robust features. In ECCV, 2006.

[16] Y. Ke and R. Sukthankar. PCA-SIFT: A more distinctive representation for local image descriptors. In CVPR, 2004.

[17] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In CVPR, 2006.

[18] Santella A., Agrawala M., DeCarlo D., Salesin D., and Cohen M. Gaze-based interaction for semi-automatic photo cropping. In SIGCHI, 2006.

[19] Avidan S. and Shamir A. Seam carving for content-aware image resizing. Acm Transactions on Graphics, 2007.

[20] Rubinstein M., Shamir A., and Avidan S. Multi-operator media retargeting. AcmTransactions on Graphics, 2009.

[21] Y.-S. Wang, C.-L. Tai, O. Sorkine, and T.-Y. Lee, "Optimized scaleand- stretch for image resizing," ACM Trans. Graph., vol. 27, no. 5, pp. 118:1–118:8, Dec. 2008.

[22] F. Liu and M. Gleicher, "Video retargeting: Automating pan and scan," in Proc. ACMInt. Conf. Multimedia, Santa Barbara, CA, Oct. 2006,.

[23] Ye Luo, Junsong Yuan, Ping Xue and Qi Tian, Salient Region Detection and Its Application to Video Retargeting . ICME11, 2011

[24] L.Wolf, M. Guttmann, and D. Cohen-Or, "Non-homogeneous contentdriven video-retargeting," in Proc. IEEE Int. Conf. Comput. Vis., Rio de Janeiro, Brazil, 2007, pp. 1–6.

[25] M. Rubinstein, A. Shamir, and S. Avidan, "Improved seam carving for video retargeting," ACM Trans. Graph., vol. 27, no. 3, 2008.

[26] P. Kraehenbuhl, M. Lang, A. Hornung, and M. Gross, "A system for retargeting of streaming video," ACM Trans. Graph., vol. 28, no. 5, pp. 126:1–126:10, 2009.

[27] Tzu-Chieh Yen, Chia-Ming Tsai, Chia-Wen Lin: Maintaining Temporal Coherence in Video Retargeting Using Mosaic-Guided Scaling. TIP. 20(8):2339-2351 (2011)

[28] Y.-S.Wang, H. Fu, O. Sorkine, T.-Y. Lee, and H.-P. Seidel, "Motion-aware temporal coherence for video resizing," ACM Trans. Graph., vol. 28, no. 5, 2009.

[29] Y.-S. Wang, H.-C. Lin, O. Sorkine, and T.-Y. Lee, "Motion-based video retargeting with optimized crop-and-warp," ACM Transactions on Graphics, vol. 29, no. 4, 2010.

[30] Pierre Greisen and Manuel Lang and Simon Heinzle and Aljosa Smolic. Algorithm and VLSI Architecture for Real-Time 1080p60 Video Retargeting. In HPG 2012.

[31] H. Schwarz, D. Marpe, and T. Wieg and, "Overview of the scalablevideo codingextension of H.264/AVC," IEEE Trans. Circuits Syst.Video Technol., vol. 17, no. 9,pp. 1103–1120, Sep. 2007

[32] T. Wieg and, G. J. Sullivan, G. Bjontegaard, 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

[33] M. Wien, R. Cazoulat, A.Graffunder, A. Hutter, P. Amon, Real-time system for adaptive video streaming based on SVC, IEEE Trans. Circuits Syst. Video Technol., 17 (9) (September 2007), pp. 1227–1237

[34] T. Schierl, T. Stockhammer, T. Wiegand, Mobile video transmission using scalable video coding, IEEE Trans. CSVT, 17 (9) (2007),

[35] I. Amonou, N. Cammas, S. Kervadec, S. Pateux, "Optimized Ratedistortion Extraction with Quality Layers in the Scalable Extension ofH.264/AVC" in IEEETrans. Circuits Syst. Video Technol., vol. 17, no. 9,pp. 1186–1193, Sep. 2007.

[36] J. Sun, W. Gao, D. Zhao, W. Li, "On Rate-distortion Modeling andExtraction of H.264/SVC Fine-Granular Scalable Video" in IEEE Trans.Circuits Syst. Video Technol., vol. 19, pp. 323–336, Mar. 2009.

[37] E. Maani, A. K. Katsaggelos, "Optimized Bit Extraction Using DistortionModeling in the Scalable Extension of H.264/AVC" in IEEE Trans. ImageProcessing, vol. 18, no. 9, pp. 2022–2029, Sep. 2009.

[38] W. Zhang, J. Sun, J. Liu, and Z.Guo, "Optimized bit extraction of SVC exploiting linear error model." In Circuits and Systems (ISCAS), 2012 IEEE International Symposium on, pp. 1887-1890. IEEE, 2012.

[39] K. Gao, J. Zhai, J. Li and C. Wang, "Real-Time scheduling for scalable video coding streaming system," IEEE Sarnoff Symposium, Princeton, NJ, pp. 1-4, Mar. 2006.

[40] R. Fortuna, L. A. Grieco, G. Boggia, and P. Camarda, "A scheduling strategy for P2P-TV systems using scalable video coding," IEEE GLOBECOM Workshops, Miami, Florida, pp. 949 – 953, Dec. 2010.

[41] Hoang T. Dinh, Chonho Lee, Dusit Niyato, and Ping Wang, "A survey of Mobile Cloud Computing: Architecture, Applications and Approaches", Wireless Communications and Mobile Computing, 2011.

[42] M. Armbrust, A. Fox et. al.,Above the Clouds: A Berkeley View of Cloud Computing. Technical Report, No. UCB/EECS-2009-28, Uiversity of California at Berkley, USA, Feb. 2009

[43] R. Ji, et al."Location discriminative vocabualry coding for mobile landmark search," Int. Journal of Computer Vision, vol. 96, no. 3, 2012.

[44] CDVS2. "Test Model 4: Compact descriptor for visual search," W12929, ISO/IEC JTC1/SC29/WG11, Shanghai, China, 2012.