Recent Progress in 3D Multimedia Research

Dinei Florencio

Microsoft Research
One Microsoft Way, Redmond, WA, USA
dinei@microsoft.com

***Abstract*—The first International Workshop on Hot Topics in 3D multimedia (Hot3D), was help in Singapore, in July 2010, in conjunction with the IEEE International Conference on Multimedia and Expo (ICME). Sponsored by four different IEEE societies, ICME is the IEEE flagship conference on Multimedia. Hot3D was a vibrant workshop, very well attended, with an impressive collection of papers covering several aspects of 3D multimedia. This report summarizes some of the results presented at the workshop.**

I. Introduction

Research in 3D multimedia is at a potentially revolutionary point. One of the driving forces of this revolution is 3D display technology. With the latest technology progress in that front, 3D displays are at the verge of becoming widespread and reasonably priced, and that may include even autostereoscopic displays. From another direction, increases in computational power - including powerful GPUs - has allowed an ever-increasing realism in 3D scene generation. 3D audio in now often tightly integrated with 3D environments, including 5.1 (and higher) and even 3D soundfield reproduction. Haptic systems are also being tightly integrated within 3D systems. Finally, new depth cameras, coupled with new 3D analysis and synthesis algorithms is close to enable commercial-quality 3D rendering of real scenes, instead of being restricted to synthetic scenes as in the past.

All these factors together create the ”Perfect Storm”: an environment prone for an explosion of related technology and applications, with a speed of development that does not fit in the (slower) cycle of traditional conferences and journals. In other words, while appropriate venues for presenting research at advanced stages is plentiful, the 3D Multimedia community lacked an appropriate venue for receiving feedback during early or initial stages of the development of radical or potentially disruptive technologies.

This is exactly the void that Hot3D tries to fill, providing an environment for lively discussion of early-stage, potentially disruptive research. The first Hot3D workshop was help in Singapore, on July 23rd, 2010, in conjunction with the IEEE International Conference on Multimedia & Expo.

Hot3D'2010, July 23, 2010, Singapore. (in Conjunction with ICME'2010)
978-1-4244-4464-9/09/$25.00 ©2009 IEEE.

The event attracted a large number of researchers, working of a variety of topics relating to 3D multimedia. It was a unique opportunity to interact with other researchers working on 3D Multimedia, under an environment designed to facilitate discussion and feedback in early stage research, as well as forge new collaboration.

This paper summarizes the research presented at the conference.

II. 3D Content, Quality, and Haptics

The keynote speech ”Next Generation 3D Video Representation, Processing and Coding” was presented by Aljoscha Smolic, from Disney Research. It discussed a number of aspects, from the current state of 3D content production, from technologies under development, to expected limitations of technologies, and research still needed to fully exploit the capabilities of 3D consumer equipment in the near future.

In a very lively section, a position paper [1] was presented by Eckhard Steinbach calling on the need for additional research on 3D Haptics. Another position paper [2]was presented by Touradj Ebrahimi, discussing the need and status of 3D quality evaluation.

III. Depth Estimation and 3D Formats

Jachalsky et. al. [3], proposed a confidence map which combines the consistency and quality of a match. It explicitly models the reliability of each disparity estimate. Such a confidence map represents valuable, additional information that can be leveraged in subsequent steps of the 3D processing chain. In this regard, this paper presents an extension of a cross bilateral filter that leverages this reliability information during a fast-converging refinement step in order to create robust and reliable disparity maps.

Kang and Ho [4]propose a high-quality multi-view depth generation method using multiple color and depth cameras. After capturing low-resolution depth maps by three TOF cameras, the depth information is warped into color image positions and used as the initial disparity value. By applying the stereo matching using belief propagation with the initial disparity information, they obtain more accurate and stable multi-view disparity maps, compared to those results without the initial disparity information.

In shape from focus (SFF), noise, illumination variation and oriented features degrade the performance of focus measure operator. Mendapara et.al. introduce the use of complex wavelets due to shift-invariance and directionality of the transformation suitable for detecting various types of features which plays a pivotal role in depth estimation of a scene [5]. A quadrature pair of steerable filters is employed to measure focus by calculating the local oriented energy of the detected features. Experimental examples are provided to illustrate the effectiveness of the approach and the results compare favorably to well-documented methods in literature.

Many 3D formats exist and will co-exist for a long time since there is no 3D standard that defines a generally accepted 3D format. The support for multiple 3D formats will be important for bringing 3D into home. In [6] Zhang proposes a fast and effective method to detect whether an image is a 2D image or a 3D image encoded with a pair of stereo images, and to further identify the exact 3D format in the latter case. The method computes an edge map resulted from image differences between the left and right view images; uses the statistics from the distribution of edge widths combined with structure similarity analysis to detect the existence of a 3D format and to identify the format.

IV. 3D DISPLAYS, IMAGING AND 3D AUDIO

Plenoptic cameras1are able to acquire multi-view content that can be displayed on auto-stereoscopic displays. Depth maps can be generated from the set of multiple views. As it is a single lens system, very often the question arises whether this system is suitable for 3D or depth measurement. The underlying thought is that the precision with which it is able to generate depth maps is limited by the aperture size of the main lens. In [7] Drazic explores the depth discrimination capabilities of plenoptic cameras. A simple formula quantifying the depth resolution is given and used to drive the principal design choices for a good depth measuring single lens system.

Multi-user autostereoscopic displays have been developed within the European Union-funded MUTED and HELIUM3D projects. These utilize head tracking in order to provide images that are displayed in regions referred to as exit pupils that follow the users eye positions. In the MUTED displays images are produced on a direct-view liquid crystal display (LCD) with novel optics controlled by the head tracker replacing the conventional backlight. Surman et. al. [8] describe the design and construction of the displays along with evaluation results and future developments, describing principle of operation, current status, and the multimodal potential of the HELIUM3D display.

Traditional 3D audio systems often have a limited sweet spot for the user to perceive 3D effects successfully. In [9] Song et. al. present a personal 3D audio system with loudspeakers that has unlimited sweet spots. The idea is to have a camera track the users head movement, and recompute the crosstalk canceler filters accordingly. The system is the first non-intrusive 3D audio system that adapts to both the head position and orientation with six degrees of freedom. This work has later been extended to incorporate room modeling [10] further improving the results [11].

V. 3D SYNTHESIS AND 3D VIDEO CODING

Research on error resilience in multi-view coding is currently receiving considerable interest. While there is a multitude of literature concerning error recovery in 2D video, due to the statistical difference in motion compensation among temporal frames and disparity compensation among view points, such methods are inadequate to cater to the requirements of multiview video transmission. in [12], Dissanayake et. al. address the above issue by transmission of redundant disparity vectors for error recovery purposes. The system, which is implemented using the Joint Scalable Video Model (JSVM) codec and tested using a simulated Internet Protocol (IP) packet network environment, can be used along with a suitable error concealment scheme to provide robust multiview video transmission. The experimental results suggest that the proposed algorithm experiences a slight degradation of quality in error free environments due to the inclusion of redundant data. However, it improves the reconstructed picture quality significantly in error prone environments, specifically for Packet Loss Rates (PLRs) greater than 7%.

Interest in 3D video visualization systems is a ever growing field. Such areas include the provision of 3D content to users thus opening the exploration of 3D video communication and transmission. To address communication and transmission one must consider error resilience. Multiple Description Coding (MDC) can provide a robust video communication over wireless networks. However it can introduce high levels of redundancy. In [13], Adedoyin et. al. propose a scalable MDC architecture using motion vector (MV) encoding for 3D video. Experimental results show that the algorithm can improve the frame quality by up to 2dB over a pixel based interpolation scheme with residual coding while significantly reducing the bit rate compared to a pixel and motion interpolation schemes.

In the multiview video plus depth (MVD) representation for 3D video, a depth map sequence is coded for each view. In the decoding end, a view synthesis algorithm is used to generate virtual views from depth map sequences. Many of the known view synthesis algorithms introduce rendering artifacts especially at object boundaries. In [14], Chen et. al. present a depth-level-adaptive view synthesis algorithm that reduces the amount of artifacts and improves the quality of the synthesized images. The algorithm introduces awareness of the depth level so that no pixel value in the synthesized image is derived from pixels of more than one depth level. Improvements on objective quality of the synthesized views are achieved in five out of eight test cases, while the subjective quality of the proposed method was similar to or better than that of the view synthesis method used by Moving Picture Experts Group (MPEG).

Luat Do et. al. [15] present their ongoing research on view synthesis of free-viewpoint 3D multi-view video for 3DTV. With the emerging breakthrough of stereoscopic 3DTV, they have extended a reference free-viewpoint rendering algorithm to generate stereoscopic views. Two similar solutions for converting free-viewpoint 3D multi-view video into a stereoscopic vision have been developed. These solutions take into account the complexity of the algorithms by exploiting the redundancy in stereo images, since they aim at a real-time hardware implementation. Both solutions are based on applying a horizontal shift instead of double execution of the reference free-viewpoint rendering algorithm for stereo generation (FVP stereo generation), so that the rendering time can be reduced by as much as 3040 %. The trade-off however, is that the rendering quality is 0.50.9 dB lower than when applying FVP stereo generation. Results show that stereoscopic views can be efficiently gen- erated from 3D multi-view video by using unique properties in stereoscopic views, such as identical orientation, similarities in textures and small baseline.

VI. Other recent developments in 3D Multimedia

The 3D mutimedia are is extremely active right now. At ICME in Singapore, besides the Hot3D workshop, there was a panel on 3D multimedia, and a number of other papers addressed aspects related to 3D. The panel,entitled "3D Multimedia: Research Status and Opportunities" was moderated by Dinei Florencio, from Microsoft Research, and counted with a number of specialist in the field. These included Ebroul Izquierdo, from the University of London, Aljoscha Smolic, from Disney Research, Eckehard Steinbach, from TU Mnchen, Phil Surman, from De Montfort University, Ruigang Yang, from University of Kentucky, and Cha Zhang, from Microsoft Research.

Additionally, a number of papers from the conference were reporting relevant results on 3D multimedia. Trocan et all [16] investigates the use of compressed sensing in multiview imaging, proposing an algorithm that drives image recovery using the projection-domain residual between the random measurements of the image in question and a disparity-based prediction created from adjacent, high-quality images.

One of the papers in 3D audio won one of the ICME best paper awards[17], and has recently been extended to analyze parameter sensitivity [18]. It improves the accuracy of 3D sound source localization (SSL) by up to one order of magnitude compared to the state of the art [19], by using the additional information available on early reflections. It builds on a sparsity-motivated room modeling technique [10] and on recent results on maximum likelihood SSL [20], [21]. Another interesting paper on 3D sound [22] places sounds on a 3D virtual environment for user navigation through a scene.

One of the important applications of 3D is the area of remote collaboration. That area has some recent interesting results relating to pseudo-3D [23] and multiview coding [24] and it was well represented at ICME with an interesting paper from HP labs [25] evaluating eye contact.

VII. Conclusion

Research in 3D Multimedia is advancing at a huge pace. Progress in several fronts are been presented, and significant enhancements are been achieved in areas like 3D displays, 3D audio, free viewpoint synthesis, 3D virtual environments, 3D collaboration and many others. We expect this area to be very dynamic for the next several years, culminating with a pervasive 3D displays and environments. We have some exciting years ahead.

References

[1] E. Steinbach, R. Chaudhari, and J. Kammerl, "Current status and future directions of haptics in 3d multimedia," in *Hot3D (ICME'2010)*, 2010.

[2] L. Goldmann and T. Ebrahimi, "3d quality is more than just the sum of 2d and depth," in *Hot3D (ICME'2010)*, 2010.

[3] J. Jachalsky, M. Schlosser, and D. Gandolph, "Confidence evaluation for robust, fast-converging disparity map refinement," in *Hot3D (ICME'2010)*, 2010.

[4] Y.-S. H. Yun-Suk Kang, "High-quality multi-view depth generation using multiple color and depth cameras," in *Hot3D (ICME'2010)*, 2010.

[5] P. Mendapara, A. Baradarani, and Q. M. J. Wu, "An efficient depth map estimation technique using complex wavelets," in *Hot3D (ICME'2010)*, 2010.

[6] T. Zhang, "3d image format identification by image difference," in *Hot3D (ICME'2010)*, 2010.

[7] V. Drazic, "Optimal depth resolution in plenoptic imaging," in *Hot3D (ICME'2010)*, 2010.

[8] P. Surman, R. Brar, I. Sexton, and K. Hopf, "Muted and helium3d autostereoscopic displays," in *Hot3D (ICME'2010)*, 2010.

[9] M.-S. Song, C. Zhang, D. Florencio, and H.-G. Kang, "Personal 3d audio system with loudspeakers," in *Hot3D (ICME'2010)*, 2010.

[10] D. Ba, F. Ribeiro, C. Zhang, and D. Florncio, "L1 regularized room modeling with compact microphone arrays," in *ICASSP*, 2010.

[11] M.-S. Song, C. Zhang, D. Florencio, and H.-G. Kang, "Enhancing loudspeaker-based 3d audio with room modeling," in *MMSP*, 2010.

[12] M. B. Dissanayake, D. D. Silva, S. Worrall, and W. Fernando, "Error resilience technique for multi-view coding using redundant disparity vectors," in *Hot3D (ICME'2010)*, 2010.

[13] S. Adedoyin, W. Fernando, and A. Kondoz, "Scalable mdc for 3d stereoscopic video using motion vector encoding," in *Hot3D (ICME'2010)*, 2010.

[14] Y. Chen, W. Wan, M. Hannuksela, J. Zhang, H. Li, and M. Gabbouj, "Depth-level-adaptive view synthesis for 3d video," in *Hot3D (ICME'2010)*, 2010.

[15] P. H. N. D. W. Luat DO, Svitlana Zinger, "Conversion of free-viewpoint 3d multi-view video for stereoscopic displays," in *Hot3D (ICME'2010)*, 2010.

[16] M. Trocan, T. Maugey, J. Fowler, and B. Pesquet-Popescu, "Disparity-compensated compressed-sensing reconstruction for multiview images," in *ICME*, 2010.

[17] F. Ribeiro, D. Ba, C. Zhang, and D. Florncio, "Turning enemies into friends: Using reflections to improve sound source localization," in *ICME*, 2010.

[18] F. Ribeiro, C. Zhang, D. Florncio, and D. Ba, "Using reverberation to improve range and elevation discrimination for small array sound source localization," *IEEE Trans. on Audio, Speech and Lang. Proc.*, vol. 18, no. 7, pp. 1781–1792, Sept 2010.

[19] C. Zhang, D. Florencio, and Z. Zhang, "Why does phat work well in low noise, reverberative environments?" in *ICASSP*, 2008.

[20] C. Zhang, Z. Zhang, and D. Florêncio, "Maximum likelyhood sound source localization for multiple directional microphones," in *ICASSP*, 2007.

[21] C. Zhang, D. Florêncio, D. Ba, and Z. Zhang, "Maximum likelihood sound source localization and beamforming for directional microphone arrays in distributed meetings," *IEEE Trans. on Multimedia*, vol. 10, no. 3, pp. 538–548, April 2008.

[22] W. Ma, Y. Liu, Y. Wang, Y. Xu, H. Zha, and W. Gao, "Interactive viewpoint-space navigation for visual-audio exhibition of painting," in *ICME*, 2010.

[23] C. Zhang, Z. Yin, and D. Florêncio, "Improving depth perception with motion parallax and its application in teleconferencing," in *MMSP*, 2009.

[24] D. Florencio and C. Zhang, "Multiview video compression and streaming based on predicted viewer position," in *ICASSP*, 2009.

[25] K.-H. Tan, I. N. Robinson, B. Culbertson, and J. Apostolopoulos, "Enabling genuine eye contact and accurate gaze in remote collaboration," in *ICME*, 2010.