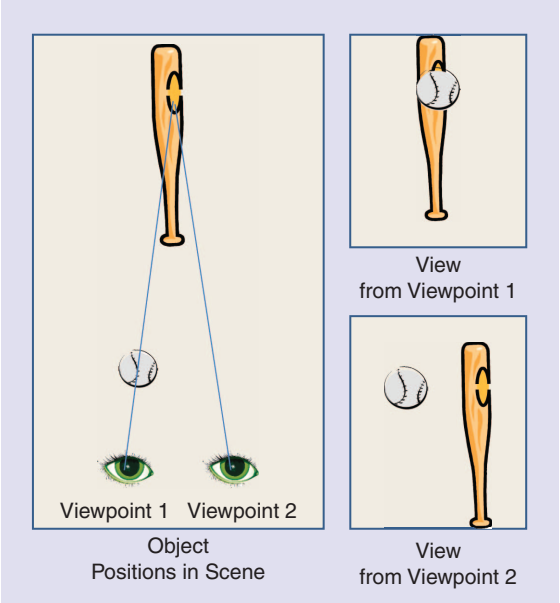
Cha Zhang, Dinei Florêncio, and Zhengyou Zhang

Improving Immersive Experiences in Telecommunication with Motion Parallax

Human sensorial perception of the surrounding environment is very intricate and only partially understood. Visual depth perception, for example, is often attributed to stereo vision. However, if we close one eye, it becomes immediately obvious how much depth information we acquire from other mechanisms. One of these mechanisms is motion parallax, i.e., the fact that the relative apparent positions of objects change when we move our viewpoint. Traditional three-dimensional television (3DTV) systems provide carefully edited stereo video but often lack the capability of rendering any motion parallax. Similar to the visual experience, our audio experience is also affected by our body motion. Indeed, in addition to information about the environment captured by two ears with a fixed head position, we also capture (often, unconsciously) a richness of information about the environment by moving our heads to sample the sound field around us.

To produce a truly immersive experience, telecommunication systems will have to exploit the richness of this perception. Video should supply not only stereo vision, but motion parallax as well. Audio should reproduce the full richness of sound fields. In other words, the reproduced environment (be it a natural or synthetic environment) must be consistent between the audio and video and provide the parallax sensorial perception of real-world environments. Signal processing plays a fundamental role in providing this experience. Indeed, many existing techniques can be used to improve signal capture, head tracking, beamforming, audio spatialization, and free viewpoint synthesis. In the following, we will discuss some of these techniques, many of which are active areas of research.



[FIG1] The observed image depends on the observation point.

MOTION PARALLAX

The term parallax is derived from the Greek παράλλαξις (parallaxis), which means "alteration." Parallax can be defined as the change in the received signal that occurs as the consequence of the change in sensor position. This term is commonly used in reference to vision: the relative apparent positions of objects will vary depending on the observation position. Figure 1 illustrates the phenomenon. From Viewpoint 1, the ball is aligned with the bat, as shown by the straight line passing through the ball between the eye and the bat. From Viewpoint 2, however, the ball is located to the left of the bat. This variation is parallax. When the variation in viewpoint being considered is the difference between left and right eye positions, it is referred to as stereo parallax, or stereopsis. When the variation in the viewpoint being considered is due to head movement, it is referred to as motion parallax. Studies have shown that humans experience depth perception from either or both parallaxes [1].

Note that the distinction between stereopsis and motion parallax may be subtle, but they are very different in their implications for communication systems. It may be reasonably simple to capture and send two (fixed) views to generate the stereo parallax, such as the case of 3-D movies. In contrast, to generate motion parallax, the user's head motion has to be known before the image can be rendered. This makes the problem much more complex and demanding, as each frame has to be generated in real time [2]. As such, techniques for acquiring multiple signals, compressing multiple views, estimating head position, and generating the views, all in real time, may all be needed.

Although not mentioned as often, audio is also subject to parallax. More specifically, head movements induce small variations in the transfer functions between the sound sources and the ears. Humans use these variations to aid, for instance, in sound source localization (SSL). Additionally, the relative delay between the direct path and

Digital Object Identifier 10.1109/MSP.2010.939037
Date of publication: 17 December 2010

the reflected sounds will vary, giving information about the room. This interaction of sound waves propagating in different directions creates a complex spatially and time-varying pressure function, normally referred to as the sound field. This sound field carries rich information about the sound sources, their location, and the environment. Reproducing and manipulating the sound field adequately will significantly enhance the feeling of immersion.

MOTION TRACKING

Motion tracking is an essential piece of technology to enable motion parallax rendering, and has been extensively studied in the virtual reality (VR) and augmented reality (AR) literature. According to Welch and Foxlin [3], motion tracking systems usually derive pose estimations from electrical measurements of mechanical, inertial, acoustic, magnetic, optical, and/or radio frequency sensors, and each approach has its advantages and limitations. Lately, many systems in the VR community apply hybrid approaches, fusing multiple sensor inputs to achieve very high accuracy and very low delay. For instance, with the user wearing a device of the size of a tennis ball, the six degrees of freedom (DoF) HiBall-3100 system from 3rdTech Inc. can track with the accuracy of a few tenths of a millimeter of position resolution, hundredths of a degree of rotation accuracy, and within 1 ms of latency.

In immersive telecommunication applications, a nonintrusive motion tracking scheme would be ideal, as these systems strive to maintain the most natural and immersive way for communication. Considering that cameras and microphones are abundant in telecommunication systems, it is natural to favor visual or acoustic tracking schemes.

ACOUSTIC TRACKING

Acoustic tracking in immersive telecommunication applications is often conducted with microphone arrays. A typical microphone array can be a compact linear or circular unit with four to eight microphones at 5–20 cm apart, but more generally may consist of a set of microphones distributed strategically in the room. Consider an array of P microphones. Given a frequency domain source signal $S(\omega)$, the signals received at these microphones can be modeled as

$$X_i(\omega) = \alpha_i(\omega) S(\omega) e^{-j\omega\tau_i} + N_i(\omega), \quad (1)$$

where $X_i(\omega)$ is the received signal; $i = 1, \ldots, P$ is the index of each microphone; τ_i is the propagation delay from the source location to the ith microphone location; $\alpha_i(\omega)$ is the gain factor (including the effects of the propagation energy decay, the gain of the corresponding microphone, and the directionality of the source and the microphone), and $N_i(\omega)$ is the noise sensed by the ith microphone. Depending on the application, this noise term could include a room reverberation term to increase the robustness of the derived algorithm.

The goal of acoustic tracking, or sound source localization, is to derive the sound source from estimating optimal propagation delays τ_i. One of the most popular algorithms for multiple microphone SSL is the so-called steered response power—phase transform (SRP-PHAT) [4], which derives the source location by finding the maximum value of

$$\mathcal{R}(\mathbf{s}) = \int \left| \sum_{i=1}^{P} \frac{X_i(\omega) e^{j\omega\tau_i}}{|X_i(\omega)|} \right|^2 d\omega, \quad (2)$$

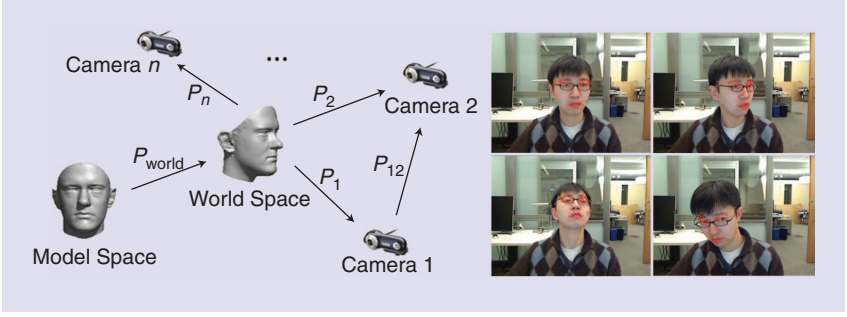
where $\mathbf{s}$ represents the source location. The time delays τ_i depend on $\mathbf{s}$, and the above function is usually maximized through hypothesis testing. Recently, a maximum likelihood (ML) solution for SSL was proposed by Zhang et al. [5], which achieves higher performance by modeling the noise component, and degenerates to SRP-PHAT if the noise is negligible.

VISUAL TRACKING

Acoustic tracking can cover a wide range and even tolerate a certain degree of occlusions, but the accuracy is usually not high enough for driving motion parallax based audio/video rendering. A more accurate scheme of tracking is through one or multiple cameras.

Face tracking has been explored extensively in the computer vision community. Early methods often relied on color histograms, background subtraction, motion estimation, etc. Such methods work well for applications such as surveillance or smart camera systems, where the requirement on the precision of the tracking result is relatively low. To drive motion parallax, a model-based scheme would be more appropriate.

Figure 2 shows the basic principle of a model-based multicamera face-tracking algorithm [6]. A generic rigid 3-D face model is used to model the face region. Given a video sequence or multiple sequences from multiple cameras, the tracking algorithm estimates the rotation and translation of the face in the video, together with the relative rotation and translations between the observing cameras. With multiple cameras, the tracking is more robust to occlusions and large head pose changes. Another advantage of the algorithm is that the calibration information



[FIG2] Model-based face tracking from one or multiple cameras. P_{world} is the transformation from model space to world space; P_n is the projection of the face model to the nth camera; P_{ij} is the transformation between cameras. All these parameters are estimated automatically in the algorithm.

between the cameras is estimated on the fly, which avoids tedious calibration work and handles the case where cameras are moved during tracking. The visual tracking system can achieve the precision of 4–5° in angular error, and a few tenths of pixels in translation error (which translates to less than 1 mm in the physical world). This precision is sufficient for many motion parallax driven applications, as will be presented in the next two sections.

On an additional note, if the person being tracked is also speaking, it is possible to fuse both audio and visual cues for accurate tracking [7].

IMMERSIVE AUDIO

Human auditory perception is very rich. Indeed, our auditory system gives us detailed information about the environment, precise location of a sound source, its nature or construction, and the path from the source to the ear. Our awareness of that perception, however, is often less than that relating to visual clues. For example, when watching a movie on a TV screen, we are fully aware that the image is coming from the constrained screen location. Yet, we do not seem to have the same level of consciousness regarding sound. Only when we experience surround sound in its fullness do we realize how limited stereo sound is. To be fully immersive, a system has to fully exploit the richness of human sound perception, including parallax effects. That does not require, however, acquiring and faithfully reproducing a sound field remotely. Instead, signal processing techniques can be used to provide similar perception at a much smaller cost and complexity. For example, when capturing sound, beamforming techniques can be used in conjunction with microphone arrays to reduce reverberation and environmental noise. In other words, to capture the originally produced sound, without the complexities and nuances introduced by the room reverberation and noise. Audio spatialization techniques can then be used to synthetically place that source at a desired location, and to convey information about the desired synthetic room or environment. Furthermore, by doing that synthetically, we can add those for the particular location of the listener, providing the parallax effects, and thus the rich sound experience that comes with it. Indeed, by choosing appropriate signal processing techniques for capture, synthesis, and reproduction of sound, we can provide a rich sound experience, significantly improving the immersive experience in telecommunication applications.

BEAMFORMING

Microphone arrays play an important role in speech capture for immersive communication. More specifically, they enable beamforming techniques that focus the array to a particular spatial location, thus reducing environmental noise, as well as ambient reverberation. Besides improving intelligibility and subjective quality (which are probably important in any application), removing the environmental noise and ambient reverberation is also an important factor for immersive communication, as the desired virtual environment may not be the same as the actual environment the speaker is in. After capturing the desired signal as cleanly as possible, the environmental noise and/or reverberation of the virtual environment can be added, if desired.

The simplest beamforming technique is known as delay and sum. It is based on the fact that the distance between a sound source and each microphone will vary, depending on the particular location of the microphone. Thus, the phase of each frequency component will also vary among microphones. A delay and sum beamformer simply delays each microphone signal by the amount needed to align the phases for the desired source location, and adds up all the signals. Summing these appropriately delayed signals will induce constructive interference for signals coming from the location for which the phases were aligned. For other locations, interference will be positive or negative, but the gain will always be smaller than at the target location. Thus, the sound coming from the target location is amplified in relation to other locations.

More elaborate techniques for beamforming are available. In particular, null steering can be very effective in reducing noise from localized sources. In a way, null steering can be seen as the complement of delay and sum: while delay and sum places a beam at the source, null steering places a null at the interfering source. A common way of implementing null steering is by algorithms based on minimum variance distortionless response (MVDR). The basic MVDR algorithm works by minimizing the output variance of the array, with the constraint of zero distortion at the desired source location. This works well if the desired source and the interference are uncorrelated. However, in real environments reverberation is often significant, and MVDR can lead to signal cancelation. Much research has been done in making MVDR more robust [8]. Other directions of research in microphone arrays include speech model-based arrays, blind source separation, and optimum subjective filtering [9].

AUDIO SPATIALIZATION

One of the key elements in providing immersive audio is the ability to spatialize sound, i.e., to make a sound source appear to come from a certain location in space, without the need to place an actual loudspeaker at that location. The most common application in telecommunication is to render the sound at the same location as each remote party video is displayed. The same technique can also be used in audio-only systems, providing the user with a much better spatial sensation, and thus reducing the user's cognitive load to identify distinct remote participants.

Humans perceive direction based on a number of clues. The two most dominant are interaural time difference (ITD) and interaural level difference (ILD). These two are mostly responsible for horizontal direction discrimination. For elevation, the directional response of each ear (also known as head-related transfer function (HRTF) plays an important role. Finally, distance perception relies heavily on direct to reverberation

ratio and other indirect cues. Since these can all be estimated (at least to some degree), elaborate spatialization is possible if the user wears headphones [10]. In that case, we compute or estimate the signal that would arrive at each ear based on the desired location of the source, and play that sound directly into the user's ear by using a headphone. This has been successfully done for years in gaming and other applications. Nevertheless, for immersive communication applications, the use of headphones is often undesirable. Thus, spatialization using loudspeakers is very important. The most elementary way of doing loudspeaker spatialization is known as amplitude panning. It consists of playing the same sound at each of the two loudspeakers, setting the amplitude of each to produce the desired location perception. This produces reasonable results for placing the source in the line segment between the loudspeakers. Of course, by increasing the number of loudspeakers, a more diverse area can be covered. For more flexible spatialization, and in particular to place the source outside the line segments defined by the loudspeakers, more elaborate techniques have been used. In particular, and most relevant to the discussion on parallax, it has recently been shown that spatialization performance can be significantly improved by head position and orientation tracking. This can be used to cancel the HRTF associated with the true loudspeaker location. In addition, one can add the HRTF of the desired virtual location to provide a rich experience without the use of headphones [11]. Further, as techniques based on tracking become more accurate, it can be expected that the full sound field experience can be reproduced. That would include reproducing the variations in the environment response based on small head movements that are necessary to provide the sound parallax sensation.

IMMERSIVE VIDEO

Motion parallax can also be used to improve the immersive rendering of videos for telecommunication. The basic idea is to render the remote party differently depending on the local user's head position, therefore creating a viewing experience similar to the real world. Such techniques are applicable to either regular two-dimensional displays or 3-D displays.

Assuming the head pose can be reliably tracked, the main challenge in creating the motion parallax effect in video is the content creation and delivery. Specifically, how to capture and render the view-dependent content and how to send the captured videos to the remote party across the network. In the following, we present a number of schemes to address these issues.

SINGLE VIEW MOTION PARALLAX

In low-end telecommunication systems, there is usually only a single video camera for each meeting attendee. Since reconstructing the 3-D geometry from a single video is an ill-posed problem, how to create the motion parallax effect becomes a very interesting problem.

One popular technique is billboarding, which has been used in the computer graphics literature to save the number of polygons in a 3-D scene but still achieve certain degree of photo realism. For telecommunication, one can take the remote party's video, and perspectively warp it to mimic viewpoint change. The video is treated as a planar object in billboarding. This is incorrect considering that the scene usually contains large depth variations (foreground person and background environment). Nevertheless, the billboarding effect is still appealing and effective as an approximation of motion parallax.

In our recent work [2], we presented two additional effects for single view motion parallax, specifically, box framing and layered video. In box framing, a virtual box is rendered around the edges of the monitor. The remote party's video is pasted on a plane behind the virtual box, creating an interesting effect that the remote party is behind the display rather than on the display in regular video communication. In layered video, the remote party's video is first segmented into foreground and background layers. These two layers are then set at different depths for 3-D rendering. Certainly, the challenge is how to perform the layer segmentation. In this regard, there has been much work in the literature, including using stereo cameras, regular Web cams, and time-of-flight cameras.

MULTIVIEW MOTION PARALLAX

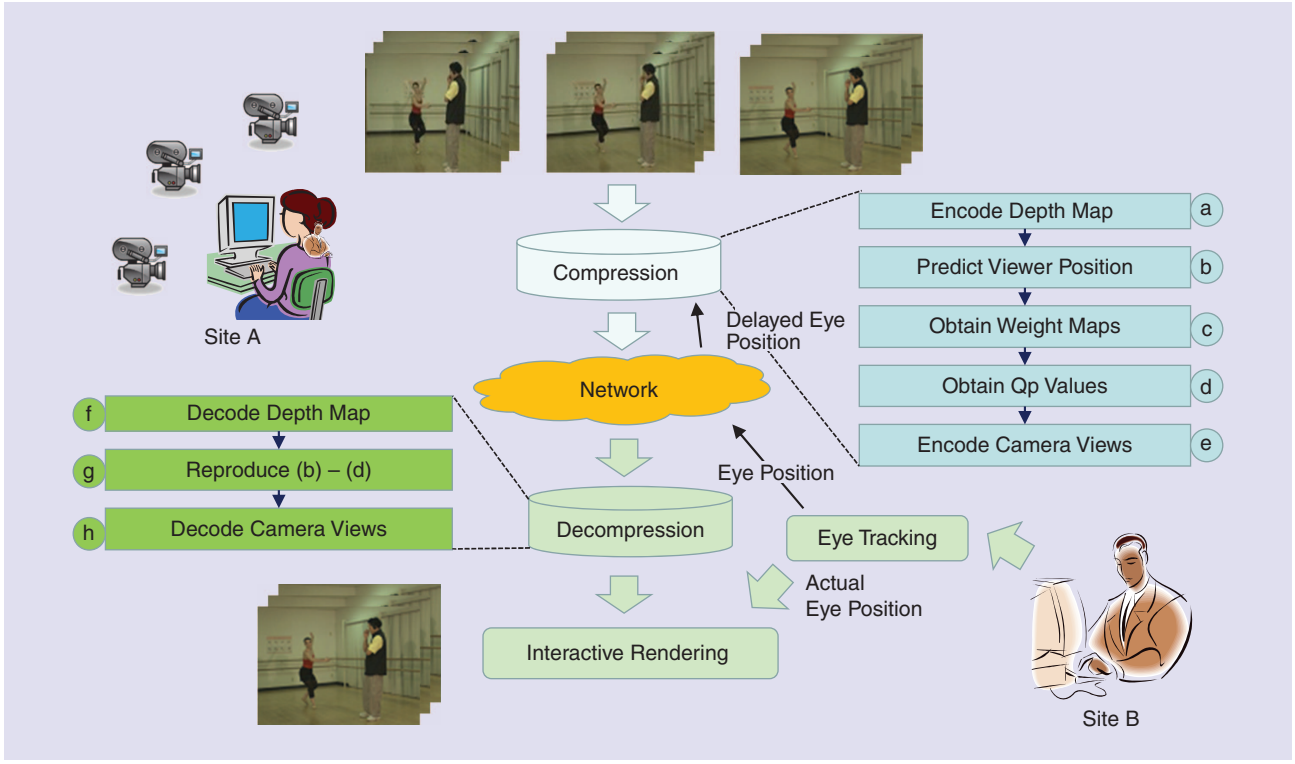
Even with layered video, the motion parallax effect is still flawed, because planar objects are rare in the real world. Ideally, one would utilize accurate scene geometry to ensure faithful motion parallax rendering. This can be achieved by multiview imaging [12]. Multiview imaging uses multiple video cameras to capture the scene from a few different viewpoints. It then applies various image-based or model-based rendering methods to render the scene at an arbitrarily new viewpoint. Such a task is nontrivial and has attracted a lot of research interest in the past two decades.

In its simplest form, assuming some rough geometry of the scene is available (through depth reconstruction, structured light, and time-of-flight cameras), one can render the scene at a new viewpoint by first splitting the to-be-rendered images into many light rays. Each light ray can be traced back to the scene geometry, and then projected back to the nearby captured views. The color of the light ray is thus the weighted average of the projected ones, where the weights depend on angular difference, distance, and occlusions. [13].

Much literature in multiview imaging has also focused on reconstructing the scene geometry through the images captured and then rendering the scene accordingly. This is usually referred as model-based rendering. Popular methods include silhouettes-based visual hulls, plane sweeping, and graph cut. Some of these methods can be accelerated with the graphics processing unit (GPU), and are thus feasible for real-time telecommunication applications.

MULTIVIEW VIDEO COMPRESSION FOR MOTION PARALLAX

When multiview imaging is used for motion parallax, how to effectively compress and stream the captured data



[FIG3] **Multiview video compression with motion parallax.**

remains a challenging problem due to the huge amount of video data involved. While the multiview video compression problem has been studied for many years [14], the goal of most existing algorithms is to use the least amount of bits to faithfully reproduce the original multiview video. In the application of motion parallax enabled telecommunication, however, additional redundancy can be explored due to the interactive viewing nature.

More specifically, since the user only sees the remote party from a particular position determined by the motion tracker, only the light rays that are required to render the corresponding view are needed for compression and streaming. The most efficient solution is thus to send the user's current head position to the remote party to guide the encoding process. Nevertheless, due to network delays, such a scheme can cause a problematic lag in the interactive viewing experience. An alternative solution is to have the encoding party actively compensating the network delay by predicting the viewer's future positions, as shown in Figure 3. At the encoding site *A*, a multiview video sequence is captured, together with a depth map (optional). Based on the viewer position predicted from the latest received eye position, the videos are encoded with adaptive quality. At the decoding site *B*, the videos are decoded, but the actual viewer position is used for interactive rendering. There could certainly be some discrepancy between the predicted and actual viewer positions, hence a probabilistic prediction scheme can often outperform deterministic schemes [15].

OUTLOOK

Motion parallax is critical to immersive audio/visual experience. With the development of high-precision, nonintrusive motion tracking schemes such as those based on multiple cameras, we expect that it will be widely adopted for future personal telecommunication systems. As briefly described in this column, there are many interesting signal processing problems related to motion parallax. These include ways of providing a richer experience, by incorporating parallax into audio and video reproduction. This achieves a better sense of immersion, as the experience is closer to human perception in natural environments. Many of the existing techniques are already at a point where they can be incorporated into products. At the same time, however, many of these algorithms still require enhancements, as current solutions may lack accuracy, robustness, and/or computational efficiency. As such, researchers will find this field full of interesting research topics, with challenging problems that have real application and significant impact.

AUTHORS

Cha Zhang (chazhang@microsoft.com) is a researcher in the Communication and Collaboration Systems Group at Microsoft Research.

Dinei Florêncio (dinei@microsoft.com) is a researcher in the Communication and Collaboration Systems Group at Microsoft Research.

Zhengyou Zhang (zhang@microsoft.com) is a principal researcher in the Communication and Collaboration Systems Group at Microsoft Research.

REFERENCES

[1] B. Rogers and M. Graham, "Similarities between motion parallax and stereopsis in human depth perception," *Vis. Res.*, vol. 22, no. 2, pp. 261–270, 1982.

[2] C. Zhang, Z. Yin, and D. Florêncio, "Improving depth perception with motion parallax and its application in teleconferencing," in *Proc. IEEE Int. Workshop Multimedia Signal Processing*, 2009, pp. 1–6.

[3] G. Welch and E. Foxlin, "Motion tracking: No silver bullet, but a respectable arsenal," *IEEE Comput. Graph. Appl.*, vol. 22, no. 6, pp. 24–38, 2002.

[4] M. Brandstein and H. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 1997, vol. 1, pp. 375–378.

[5] C. Zhang, Z. Zhang, and D. Florêncio, "Maximum likelihood sound source localization for multiple directional microphones," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 2007, vol. I, pp. 125–128.

[6] Q. Cai, A. Sankaranarayanan, Q. Zhang, Z. Zhang, and Z. Liu, "Real-time head pose tracking from multiple cameras with a generic model," in *Proc. IEEE Workshop Analysis and Modeling of Faces and Gestures*, 2010, pp. 25–32.

[7] Y. Chen and Y. Rui, "Real-time speaker tracking using particle filter sensor fusion," *Proc. IEEE*, vol. 92, no. 3, pp. 485–494, 2004.

[8] M. Brandstein and D. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*. New York: Springer-Verlag, 2001.

[9] D. Florêncio and H. Malvar, "Multichannel filtering for optimum noise reduction in microphone arrays," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 2001, vol. 1, pp. 197–200.

[10] W.-G. Chen and Z. Zhang, "Highly realistic audio spatialization for multiparty conferencing using headphones," in *Proc. IEEE Int. Workshop Multimedia Signal Processing*, 2009, pp. 1–6.

[11] M. Song, C. Zhang, D. Florêncio, and H. Kang, "Personal 3D audio system with loudspeakers," in *Proc. 2010 IEEE Int. Workshop Hot Topics in 3D (Hot3D)*, 2010, pp. 1600–1605.

[12] A. Kubota, A. Smolic, M. Magnor, M. Tanimoto, T. Chen, and C. Zhang, "Multiview imaging and 3DTV," *IEEE Signal Processing Mag.*, vol. 24, no. 6, pp. 10–21, 2007.

[13] C. Buehler, M. Bosse, L. McMillan, S. Gortler, and M. Cohen, "Unstructured lumigraph rendering," in *Proc. ACM SIGGRAPH*, 2001, pp. 425–432.

[14] M. Flierl and B. Girod, "Multiview video compression," *IEEE Signal Processing Mag.*, vol. 24, no. 6, pp. 66–76, 2007.

[15] C. Zhang and D. Florêncio, "Joint tracking and multiview video compression," *Proc. SPIE Conf. Visual Communications and Image Processing*, 2010. [SP]

special REPORTS continued from page 17

system is based on Arbor Networks Inc. technology.

Webroot, another Internet security provider, recently acquired BrightCloud, a Web content classification and security service provider. Webroot plans to integrate BrightCloud's technology into its own proprietary malware detection and SaaS technologies.

One of the problems still facing the adoption of cloud computing product and services, says the CSA's Jim Reavis, is that many government agencies and companies (even industries), while quick to adopt private clouds, are doing their own thing. The CSA hopes to help by bringing guidance to frameworks and data security standards and plans to hold several educational events in 2011.

Recognizing that the interests of several of its technical societies overlap, the IEEE Technical Activities Board (TAB) expects to make some adjustments of its own—asking the Societies to more clearly define their field of interest (FOI). That could be a challenge for the IEEE's Computer, Communications, and Consumer Electronics Societies, which have already adopted the cloud computing sector and one of their FOIs. TAB has given the Societies until 2015 to define their FOI. [SP]