Wenjun Zeng
University of Missouri, zengw@missouri.edu

Microsoft Kinect Sensor and Its Effect

Zhengyou Zhang
Microsoft Research

Recent advances in 3D depth cameras such as Microsoft Kinect sensors (www.xbox.com/en-US/kinect) have created many opportunities for multimedia computing. Kinect was built to revolutionize the way people play games and how they experience entertainment. With Kinect, people are able to interact with the games with their body in a natural way. The key enabling technology is human body-language understanding; the computer must first understand what a user is doing before it can respond. This has always been an active research field in computer vision, but it has proven formidably difficult with video cameras. The Kinect sensor lets the computer directly sense the third dimension (depth) of the players and the environment, making the task much easier. It also understands when users talk, knows who they are when they walk up to it, and can interpret their movements and translate them into a format that developers can use to build new experiences.

Kinect's impact has extended far beyond the gaming industry. With its wide availability and low cost, many researchers and practitioners in computer science, electronic engineering, and robotics are leveraging the sensing technology to develop creative new ways to interact with machines and to perform other tasks, from helping children with autism to assisting doctors in operating rooms. Microsoft calls this the Kinect Effect. On 1 February 2012, Microsoft released the Kinect Software Development Kit (SDK) for Windows (www.microsoft.com/en-us/kinectforwindows), which will undoubtedly amplify the Kinect Effect. The SDK will potentially transform human-computer interaction in multiple industries—education, healthcare, retail, transportation, and beyond.

The activity on the news site and discussion community KinectHacks.net helps illustrate the excitement behind the Microsoft Kinect technology. Kinect was launched on 4 November 2010. A month later there were already nine pages containing brief descriptions of approximately 90 projects, and the number of projects posted on KinectHacks.net has grown steadily. Based on my notes, there were 24 pages on 10 February 2011, 55 pages on 2 August 2011, 63 pages on 12 January 2012, and 65 pages on 18 February while I was writing this article. This comment from KinectHacks.net nicely summarizes the enthusiasm about Kinect: ''Every few hours new applications are emerging for the Kinect and creating new phenomenon that is nothing short of revolutionary.''

Editor's Note

Sales of Microsoft's controller-free gaming system Kinect topped 10 million during the first three months after its launch, setting a new Guinness World Record for the Fastest-Selling Consumer Electronics Device. What drove this phenomenal success? This article unravels the enabling technologies behind Kinect and discusses the Kinect Effect that potentially will transform human-computer interaction in multiple industries.

Kinect Sensor

The Kinect sensor incorporates several advanced sensing hardware. Most notably, it contains a depth sensor, a color camera, and a four-microphone array that provide full-body 3D motion capture, facial recognition, and voice recognition capabilities (see Figure 1). A detailed report of the components in the Kinect sensor is available at www.waybeta.com/news/58230/microsoft-kinect-somatosensory-game-device-full-disassembly-report-_microsoft-xbox. This article focuses on the vision aspect of the Kinect sensor. (See related work for details on the audio component.[1])

1070-986X/12/$31.00 © 2012 IEEE

Published by the IEEE Computer Society

Figure 1b shows the arrangement of the infrared (IR) projector, the color camera, and the IR camera. The depth sensor consists of the IR projector combined with the IR camera, which is a monochrome complementary metal-oxide semiconductor (CMOS) sensor. The depth-sensing technology is licensed from the Israeli company PrimeSense (www.primesense.com). Although the exact technology is not disclosed, it is based on the structured light principle. The IR projector is an IR laser that passes through a diffraction grating and turns into a set of IR dots. Figure 2 shows the IR dots seen by the IR camera.

The relative geometry between the IR projector and the IR camera as well as the projected IR dot pattern are known. If we can match a dot observed in an image with a dot in the projector pattern, we can reconstruct it in 3D using triangulation. Because the dot pattern is relatively random, the matching between the IR image and the projector pattern can be done in a straightforward way by comparing small neighborhoods using, for example, normalized cross correlation.

Figure 3 shows the depth map produced by the Kinect sensor for the IR image in Figure 2. The depth value is encoded with gray values; the darker a pixel, the closer the point is to the camera in space. The black pixels indicate that no depth values are available for those pixels. This might happen if the points are too far (and the depth values cannot be computed accurately), are too close (there is a blind region due to limited fields of view for the projector and the camera), are in the cast shadow of the projector (there are no IR dots), or reflect poor IR lights (such as hairs or specular surfaces).

The depth values produced by the Kinect sensor are sometimes inaccurate because the calibration between the IR projector and the IR camera becomes invalid. This could be caused by heat or vibration during transportation or a drift in the IR laser. To address this problem, together with the Kinect team, I developed a recalibration technique using the card in Figure 4 that is shipped with the Kinect sensor. If users find that the Kinect is not responding accurately to their actions, they can recalibrate the Kinect sensor by showing it the card. The idea is an adaptation of my earlier camera calibration technique.[2]

The depth value produced by the Kinect sensor is assumed to be an affine transformation

Figure 1. Microsoft Kinect sensor. (a) The Kinect sensor for Xbox 360. (b) The infrared (IR) projector, IR camera, and RGB camera inside a Kinect sensor.

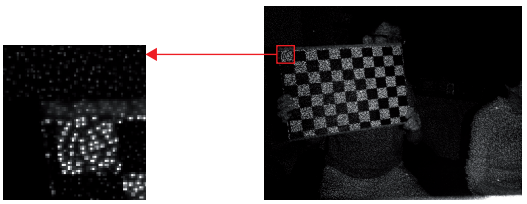

Figure 2. The infrared (IR) dots seen by the IR camera. The image on the left shows a close-up of the red boxed area.

Figure 3. Kinect sensor depth image. The sensor produced this depth image from the infrared (IR) dot image in Figure 2.

Figure 4. Kinect calibration card. To recalibrate the Kinect sensor, the RGB camera's coordinate system determines the 3D coordinates of the feature points on the calibration card, which are considered to be the true values.

of the true depth value—that is, $Z_{\text{measured}} = \alpha Z_{\text{true}} + \beta$—which we found to be a reasonably good model. The goal of recalibration is to determine α and β. (We could also use a more complex distortion model that applies the same technique.) Using the RGB camera, the recalibration technique determines the 3D coordinates of the feature points on the calibration card in the RGB camera's coordinate system, which are considered to be the true values. At the same time, the Kinect sensor also produces the measured 3D coordinates of those feature points in the IR camera's coordinate system. Minimizing the distances between the two point sets, the Kinect sensor can estimate the values of α and β and the rigid transformation between the RGB camera and the IR camera.

Kinect Skeletal Tracking

The innovation behind Kinect hinges on advances in skeletal tracking. The operational envelope demands for commercially viable skeletal tracking are enormous. Simply put, skeletal tracking must ideally work for every person on the planet, in every household, without any calibration. A dauntingly high number of dimensions describe this envelope, such as the distance from the Kinect sensor and the sensor tilt angle. Entire sets of dimensions are necessary to describe unique individuals, including size, shape, hair, clothing, motions, and poses. Household environment dimensions are also necessary for lighting, furniture and other household furnishings, and pets.

In skeletal tracking, a human body is represented by a number of joints representing body parts such as head, neck, shoulders, and arms (see Figure 5a). Each joint is represented by its 3D coordinates. The goal is to determine all the 3D parameters of these joints in real time to allow fluent interactivity and with limited computation resources allocated on the Xbox 360 so as not to impact gaming performance. Rather than trying to determine directly the body pose in this high-dimensional space, Jamie Shotton and his team met the

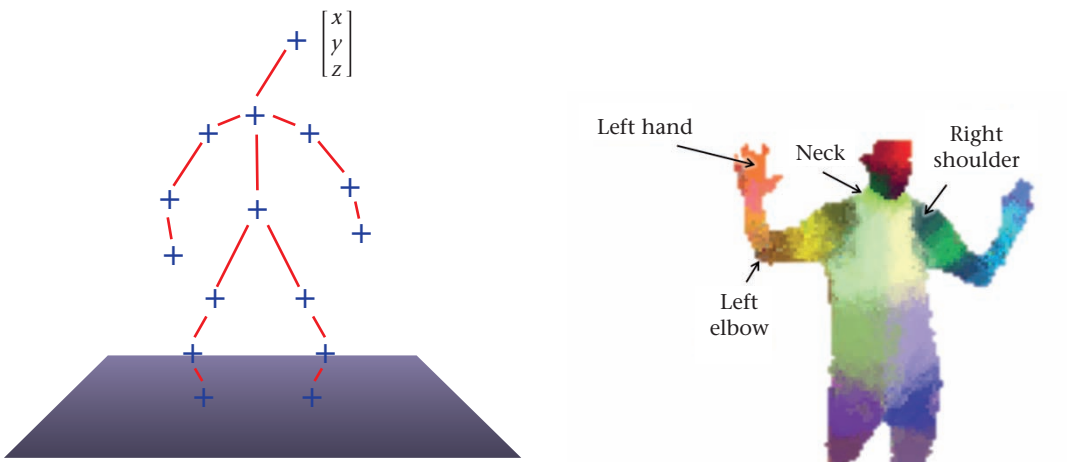

Figure 5. Skeletal tracking. (a) Using a skeletal representation of various body parts, (b) Kinect uses per-pixel, body-part recognition as an intermediate step to avoid a combinatorial search over the different body joints.

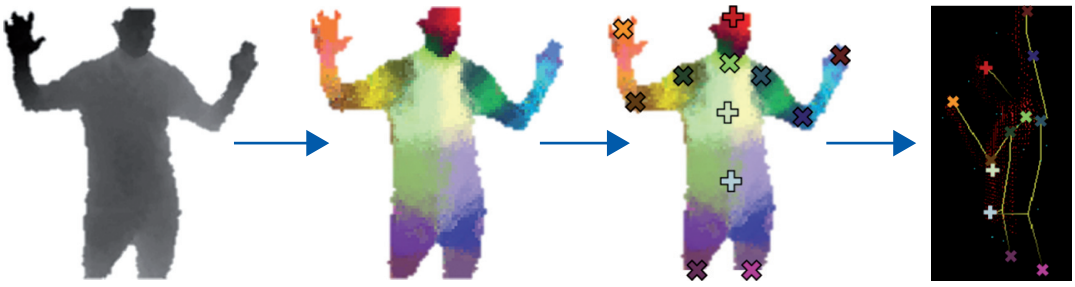

Figure 6. The Kinect skeletal tracking pipeline. After performing per-pixel, body-part classification, the system hypothesizes the body joints by finding a global centroid of probability mass and then maps these joints to a skeleton using temporal continuity and prior knowledge.

challenge by proposing per-pixel, body-part recognition as an intermediate step (see Figure 5b).[3] Due to their innovative work, Microsoft honored the Kinect Skeletal Tracking team members with the 2012 Outstanding Technical Achievement Award (www.microsoft.com/about/technicalrecognition/Kinect-Skeletal-Tracking.aspx).

Shotton's team treats the segmentation of a depth image as a per-pixel classification task (no pairwise terms or conditional random field are necessary). Evaluating each pixel separately avoids a combinatorial search over the different body joints. For training data, we generate realistic synthetic depth images of humans of many shapes and sizes in highly varied poses sampled from a large motion-capture database. We train a deep randomized decision forest classifier, which avoids overfitting by using hundreds of thousands of training images. Simple, discriminative depth comparison image features yield 3D translation invariance while maintaining high computational efficiency.

For further speedup, the classifier can be run in parallel on each pixel on a graphics processing unit (GPU). Finally, spatial modes of the inferred per-pixel distributions are computed using mean shift resulting in the 3D joint proposals. An optimized implementation of our algorithm runs in under 5 ms per frame (200 frames per second) on the Xbox 360 GPU. It works frame by frame across dramatically differing body shapes and sizes, and the learned discriminative approach naturally handles self-occlusions and poses cropped by the image frame.

Figure 6 illustrates the whole pipeline of Kinect skeletal tracking. The first step is to perform per-pixel, body-part classification. The second step is to hypothesize the body joints by finding a global centroid of probability mass (local modes of density) through mean shift. The final stage is to map hypothesized joints to the skeletal joints and fit a skeleton by considering both temporal continuity and prior knowledge from skeletal train data.

Head-Pose and Facial-Expression Tracking

Head-pose and facial-expression tracking has been an active research area in computer vision for several decades. It has many applications including human-computer interaction, performance-driven facial animation, and face recognition. Most previous approaches focus on 2D images, so they must exploit some appearance and shape models because there are few distinct facial features. They might still suffer from lighting and texture variations, occlusion of profile poses, and so forth.

Related research has also focused on fitting morphable models to 3D facial scans. These 3D scans are usually obtained by high-quality laser scanners or structured light systems. Fitting these high-quality range data with a morphable face model usually involves the well-known iterative closest point (ICP) algorithm and its variants. The results are generally good, but these capturing systems are expensive to acquire or operate and the capture process is long.

A Kinect sensor produces both 2D color video and depth images at 30 fps, combining the best of both worlds. However, the Kinect's depth information is not very accurate. Figure 7 shows an example of the data captured by Kinect. Figure 7c, a close-up of the face region rendered from a different viewpoint, shows that the depth information is much noisier than laser-scanned data.

Figure 7. An example of a human face captured by the Kinect sensor. (a) Video frame (texture), (b) depth image, and (c) close up of the facial surface.

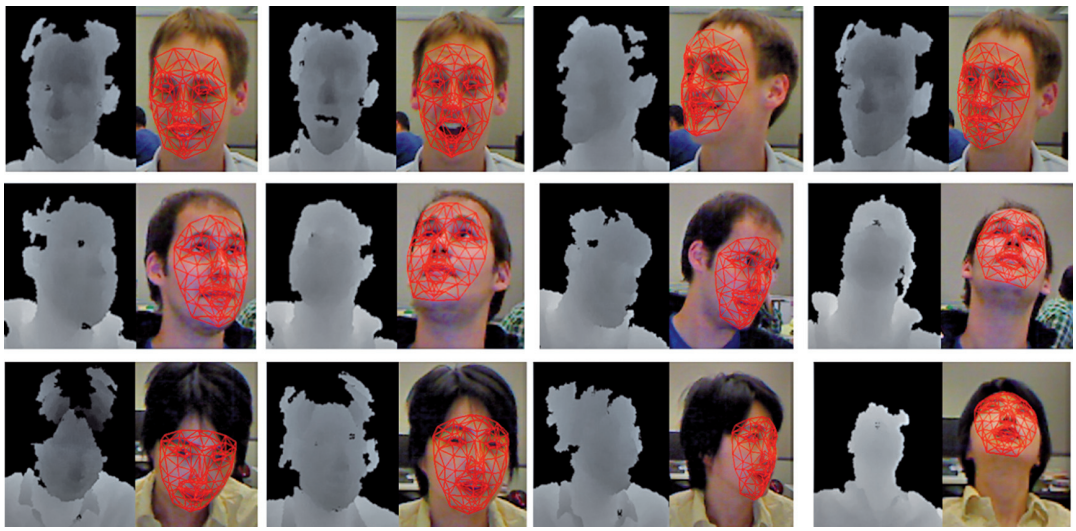

Figure 8. Facial expression tracking. These sample images show the results of Kinect tracking 2D feature points in video frames using a projected face mesh overlay.

We developed a regularized maximum-likelihood deformable model fitting (DMF) algorithm for 3D face tracking with Kinect.[4] We use a linear deformable head model with a linear combination of a neutral face, a set of shape basis units with coefficients that represent a particular person and are static over time, and a set of action basis units with coefficients that represent a person's facial expression and are dynamic overtime. Because a face cannot perform all facial expressions simultaneously, we believe in general the set of coefficients for the action basis units should be sparse, and thus we impose a L_1 regularization.

The depth values from Kinect do not have the same accuracy. Depth is determined through triangulation, similar to stereovision. The depth error increases with the distance squared. Thus, in formulating the distance between the face model and the depth map, although we still use the ICP concept, each point from the depth map has its proper covariance matrix to model its uncertainty, and the distance is actually the Mahalanobis distance. Furthermore, the 2D feature points in the video frames are tracked across frames and integrated into the DMF framework seamlessly. In our formulation, the 2D feature points do not necessarily need to correspond to any vertices or to semantic facial features such as eye corners and lip contours in the deformable face model. The sequence of images in Figure 8 demonstrates the effectiveness of the proposed method.

Microsoft Avatar Kinect has adopted similar technology (www.xbox.com/en-us/kinect/avatar-kinect). With Avatar Kinect, you can control your avatar's facial expression and head through facial-expression tracking and its arm movements through skeletal tracking (see Figure 9). As you talk, frown, smile, or scowl, your voice and facial expressions are enacted by your avatar, bringing it to life. Avatar Kinect offers 15 unique virtual environments to reflect your mood and to inspire creative conversations and performances. In a virtual environment you choose, you can invite up to seven friends to join you for a discussion or have them join you at the performance stage where you can put on a show. Thus, you can see your friends' actual expressions in real time through their avatars.

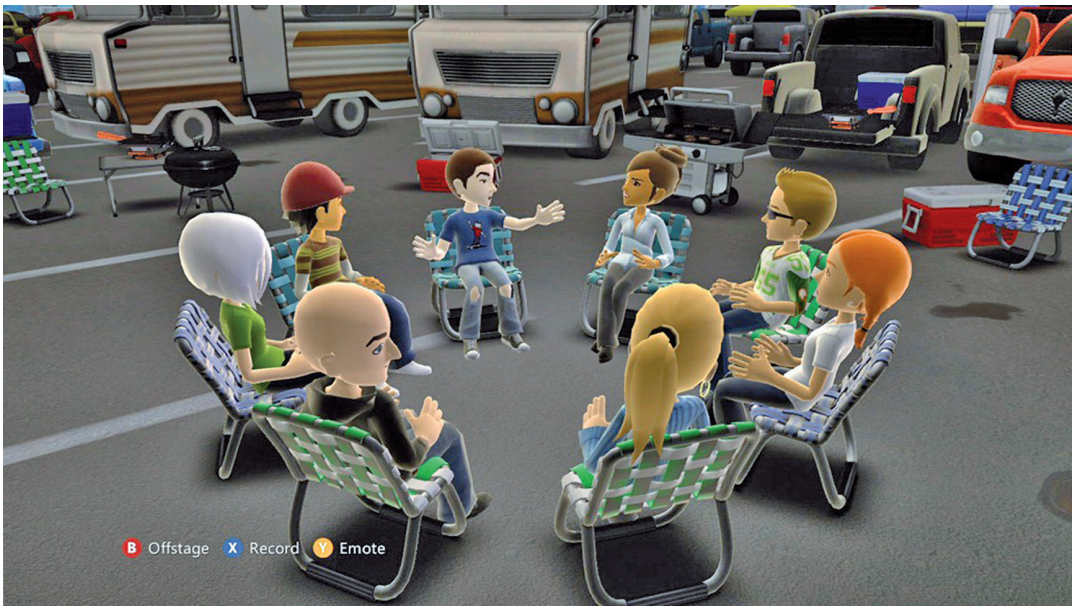

Figure 9. Avatar Kinect virtual environment. Users can control their avatars' facial expressions through facial-expression tracking and body movements through skeletal tracking.

Teleimmersive Conferencing

With increasing economic globalization and workforce mobilization, there is a strong need for immersive experiences that enable people across geographically distributed sites to interact collaboratively. Such advanced infrastructures and tools require a deep understanding of multiple disciplines. In particular, computer vision, graphics, and acoustics are indispensable to capturing and rendering 3D environments that create the illusion that the remote participants are in the same room. Existing videoconferencing systems, whether they are available on desktop and mobile devices or in dedicated conference rooms with built-in furniture and life-sized high-definition video, leave a great deal to be desired—mutual gaze, 3D, motion parallax, spatial audio, to name a few. For the first time, the necessary immersive technologies are emerging and coming together to enable real-time capture, transport, and rendering of 3D holograms, and we are much closer to realizing man's dream reflected in Hollywood movies, from *Star Trek* and *Star Wars* to *The Matrix* and *Avatar*.

The Immersive Telepresence project at Microsoft Research addresses the scenario of a fully distributed team. Figure 10 illustrates three people joining a virtual/synthetic meeting from their own offices in three separate locations. A capture device (one or multiple Kinect sensors) at each location captures users in 3D with high fidelity (in both geometry and appearance). They are then put into a virtual room as if they were seated at the same table. The user's position is tracked by the camera so the virtual room is rendered appropriately at each location from the user's eye perspective, which produces the right motion parallax effect, exactly like what a user would see in the real world if the three people met face to face. Because a consistent geometry is maintained and the user's position is tracked, the mutual gaze between remote users is maintained.

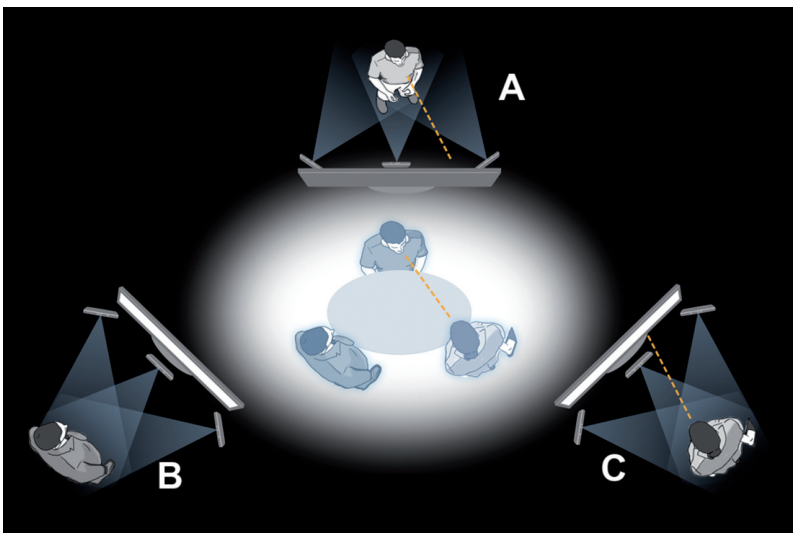

Figure 10. Immersive telepresence. One or multiple Kinect sensors at each location captures users in 3D with high fidelity. The system maintains mutual gaze between remote users and produces spatialized audio to help simulate a more realistic virtual meeting.

In Figure 10, users A and C are looking at each other, and B will see that A and C are

Figure 11. A screen shot of two remote people viewed from a third location. An enhanced 3D capture device runs in real time with multiple infrared (IR) projectors, IR cameras, and RGB cameras.

looking at each other because B only sees their side views. Furthermore, the audio is also spatialized, and the voice of each remote person comes from his location in the virtual room. The display at each location can be 2D or 3D, flat or curved, single or multiple, transparent or opaque, and so forth—the possibilities are numerous. In general, the larger a display is, the more immersive the user's experience.

Because each person must be seen from different angles by remote people, a single Kinect does not provide enough spatial coverage, and the visual quality is insufficient. Cha Zhang at Microsoft Research, with help from others, has developed an enhanced 3D capture device that runs in real time with multiple IR projectors, IR cameras, and RGB cameras. Figure 11 illustrates the quality of the 3D capture we can currently obtain with that device.

A similar system is being developed at the University of North Carolina at Chapel Hill that uses multiple Kinect sensors at each location.[5]

Conclusion

The Kinect sensor offers an unlimited number of opportunities for old and new applications. This article only gives a taste of what is possible. Thus far, additional research areas include hand-gesture recognition,[6] human-activity recognition,[7] body biometrics estimation (such as weight, gender, or height),[8] 3D surface reconstruction,[9] and healthcare applications.[10] Here, I have included just one reference per application area, not trying to be exhaustive. Visit www.xbox.com/en-US/Kinect/Kinect-Effect and www.kinecthacks.net for more examples. **MM**

References

1. I. Tashev, "Recent Advances in Human-Machine Interfaces for Gaming and Entertainment," *Int'l J. Information Technology and Security,* vol. 3, no. 3, 2011, pp. 69–76.
2. Z. Zhang, "A Flexible New Technique for Camera Calibration," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 22, no. 11, 2000, pp. 1330–1334.
3. J. Shotton et al., "Real-Time Human Pose Recognition in Parts from a Single Depth Image," *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (CVPR), IEEE CS Press, 2011, pp. 1297–1304.
4. Q. Cai et al., "3D Deformable Face Tracking with a Commodity Depth Camera," *Proc. 11th European Conf. Computer Vision* (ECCV), vol. III, Springer-Verlag, 2010, pp. 229–242.
5. A. Maimone and H. Fuchs, "Encumbrance-Free Telepresence System with Real-Time 3D Capture and Display Using Commodity Depth Cameras," *Proc. IEEE Int'l Symp. Mixed and Augmented Reality* (ISMAR), IEEE CS Press, 2011, pp. 137–146.
6. Z. Ren, J. Yuan, and Z. Zhang, "Robust Hand Gesture Recognition Based on Finger-Earth Movers Distance with a Commodity Depth Camera," *Proc. 19th ACM Int'l Conf. Multimedia* (ACM MM), ACM Press, 2011, pp. 1093–1096.
7. W. Li, Z. Zhang, and Z. Liu, "Action Recognition Based on A Bag of 3D Points," *Proc. IEEE Int'l Workshop on CVPR for Human Communicative Behavior Analysis* (CVPR4HB), IEEE CS Press, 2010, pp. 9–14.
8. C. Velardo and J.-L. Dugelay, "Real Time Extraction of Body Soft Biometric from 3D Videos," *Proc. ACM Int'l Conf. Multimedia* (ACM MM), ACM Press, 2011, pp. 781–782.
9. S. Izadi et al., "KinectFusion: Real-Time Dynamic 3D Surface Reconstruction and Interaction," *Proc. ACM SIGGRAPH,* 2011.
10. S. Bauer et al., "Multi-modal Surface Registration for Markerless Initial Patient Setup in Radiation Therapy Using Microsoft's Kinect Sensor," *Proc. IEEE Workshop on Consumer Depth Cameras for Computer Vision* (CDC4CV), IEEE Press, 2011, pp. 1175–1181.

Zhengyou Zhang is a principal researcher and research manager of the Multimedia, Interaction, and Communication (MIC) Group at Microsoft Research. His research interests include computer vision, speech signal processing, multisensory fusion, multimedia computing, real-time collaboration, and human-machine interaction. Zhang has PhD and DSc degrees in computer science from the University of Paris XI. He is a fellow of IEEE and the founding editor in chief of the *IEEE Transactions on Autonomous Mental Development*. Contact him at zhang@microsoft.com.