Towards New User Interfaces Based on Gesture and Sound Identification

KRISTJAN KOŠIČ, BOŠTJAN ARZENŠEK AND SAŠA KUHAR, University of Maribor
MATEJ VOGRINČIČ, University Medical Center Maribor

The relatively low price of devices that enable capture of 3D data such as Microsoft Kinect will certainly accelerate the development and popularization of a new generation of user interaction in the business application domain. Although the application interfaces and libraries that make it easier to communicate with these devices continue to be in the process of developing and maturing, they can still be used for the development of business solutions. In addition, gestures and sounds provide more natural and effective ways of human-computer interaction. In this paper we present an overview and a basic comparison of the available sensing devices together with the experience gained during the development of solution ADORA, which main purpose is to assist surgeons with the help of contactless interaction.

Categories and Subject Descriptors: H.5.2 [**Information Interfaces and Presentation**]: User Interfaces—*Evaluation / methodology*

General Terms: Natural user interfaces,Human Computer Interaction

Additional Key Words and Phrases: depth vision, natural user interfaces, healthcare, sensors, kinect

1. INTRODUCTION

The average user communicates with the computer using a keyboard and a computer mouse. The keyboard remains at the core computer interaction since the first commercial computer in 1984. In more than half a century since the invention of the first computer mouse many pointing devices have been introduced, of which best known are a tracking mouse (trackball) and light pen. None of these devices worked out to be better than the keyboard, so the majority of human-computer interaction (HCI) is performed via computer keyboard and mouse, which are still the same as they were at the time of the invention. With the rapid advances of technology other ways of HCI were developed. In the last decade, a noticeable use of touch screen devices and other innovative gaming interfaces were developed and successfully used in practice. At the same time with technology development new challenges emerged, e.g. "How to communicate with computers using complex commands without direct physical contact?" The solution would facilitate and optimize work in many specialized domains. Alexander Shpunt [Dibbell 2011] introduced three-dimensional (3D) computer vision. Simple communication and control of a computer by using the user's movements (gestures) and voice commands was enabled. The sensing device records observed space, takes an image and converts it into a synchronized data stream. Data stream consists of depth data (3D vision) and color data (similar to human vision). Depth vision technology was invented in 2005 by Alexander Shpunt, Zeev Zalevsky, Aviad Maizels and Javier

Author's address: K. Košič, B. Arzenšek, S. Kuhar, Institute of informatics, Faculty of Electrical Engineering and Computer Science, University of Maribor, Smetanova ulica 17, 2000 Maribor, Slovenia, email: kristjan.kosic@um.si, {bostjan.arzensek, sasa.kuhar}@uni-mb.si; M. Vogrinčič, University Medical Center Maribor, Ljubljanska ulica 5, 2000 Maribor, Slovenia, email: matej.vogrincic@ukc-mb.si

In: Z. Budimac (ed.): Proceedings of the 2nd Workshop of Software Quality Analysis, Monitoring, Improvement, and Applications (SQAMIA), Novi Sad, Serbia, 15.-17.9.2013, published at http://ceur-ws.org

Garcia [Zalevsky et al. 2007]. The technology has been established in the world of consumer technology (gaming consoles). Massachusetts Institute of Technology (MIT) ranked gesture interface based technology among the top ten most successful technologies in 2011 [Dibbell 2011]. The major console manufacturers (Microsoft, Sony and Nintendo) upgraded their gaming experience with an advanced motion-sensing interfaces. Sony and Nintendo have developed a wireless controllers (PlayStation Move and Wii MotionPlus), while Microsoft's Xbox console used a completely noncontact approach with the new Kinect sensor. Microsoft Kinect sensor is currently the most visible and easily accessible gaming controller on the market. Microsoft's decision to publish development libraries Kinect for Windows SDK, enabled rapid growth and paved the way for a wide range of potential applications in different domains (medicine, robotics, interactive whiteboards, etc.). Microsoft Kinect is the first commercial composite interface that combines camera to detect body movements and facial and voice recognition. Gartner's hype cycle for HCI technologies [Prentice and Ghubril 2012] mentioned that gesture interface control technology is steadily moving towards greater productivity, a mature market and higher returns of value. Company analytics at Markets & Markets [Marketsandmarkets.com 2013] have evaluated the market value of the hardware and software that allows you to control your computer using gestures and voice commands, such as the Microsoft Kinect, to U.S. $ 200 million in 2010. According to recent market research data, the value of contactless interaction and gesture recognition in 2018 will reach 15 billion U.S. dollars [Marketsandmarkets.com 2013]. HCI defines user experience. Gesture interface control enables the recognition and understanding of human body movement for the purpose of interaction and control of computer systems without direct physical contact [Prentice and Ghubril 2012]. The term "natural user interface" is used to describe systems that communicate with the computer, without any intermediate devices. In the last decade a big leap forward was made from the traditional ways of managing computer software with keyboard and mouse, to non-contact HCI, which was primarily used in the gaming field.

2. SENSING DEVICES FOR HCI

Three types of sensing technologies mainly occur in computer vision research: stereo cameras, time-of-flight (ToF) cameras and structured light [Chen et al. 2013]. Stereo machine vision is biomimetic, where 3D structure is gathered from different viewpoints, similar to human vision. Time of flight cameras estimate distance to an object with the help of light pulses from a single camera. Information such as time to reach the object and speed of light define the distance from the measured point. ToF devices have high precision and are very expensive. Structured light sensors started to develop with the help of PrimeSense technology, which was acquired by Microsoft and built into their Kinect sensor. Main advantage of structured light sensors is the price-performance balance. Microsoft Kinect is priced at consumer level and still obtains sufficient precision level [Gonzalez-Jorge et al. 2013].

2.1 Depth sensing

Vision sensing devices capture distance from the real world, which cannot be obtained directly from an image. Depth images require pre and post-processing. Depth view enables body part detection, pose estimation and action recognition. Body parts and pose detection is a popular topic, while action recognition is starting to receive more research attention [Chen et al. 2013]. A lot of research [Clark et al. 2012; Dutta 2012; Gabel et al. 2012; Khoshelham and Elberink 2012; Stoyanov et al. 2012] is done in the field of body data detection and its transformation to active skeletons. There are still some precision issues as mentioned by [Khoshelham and Elberink 2012], but if there is no need for high precision tracking, currently available devices handle the issues quite successfully.

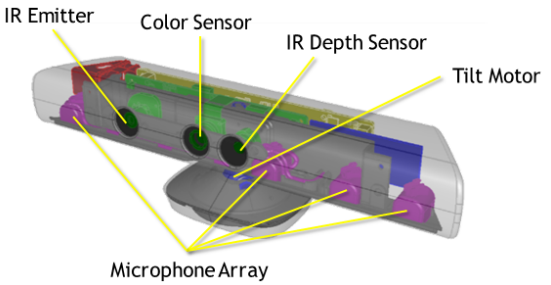

Fig. 1. Kinect sensor structure [Microsoft 2013].

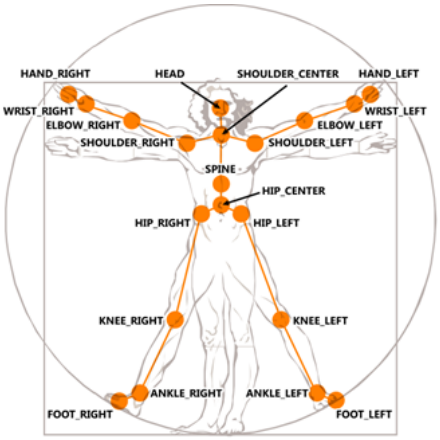

Fig. 2. Skeleton detection points [Microsoft 2013].

2.2 Microsoft Kinect sensor family

Microsoft first announced the Kinect sensor in 2009 under the name ”Project Natal”. In 2012 Kinect for Windows sensor was announced, that enabled use of advanced gesture based functionality in the business application domain. The sensor was accompanied by software development library – Kinect for Windows SDK. Until February 2013 Microsoft sold 24 million units of the Kinect sensor. In the first sixty days on the market eight million units of the Kinect sensor were sold, which gave Kinect the title of fastest selling consumer electronic device and was recorded in the Guinness Book of Records[1]. The sensor has the ability to capture audio, color video and depth data (figure 1). Depth data is detected with the use of infrared light, with which a covered skeleton (figure 2) of a tracking person is formed. Kinect sensor captures depth image by using two separate data streams. The first stream presents the data from the Kinect sensor to the nearest object in millimeters and the second one presents the segmented data from a tracked person. Deep video supports the default resolution of 640x480 pixels, the value can be set to 320x240 or 80x60 pixels. Sensor can recognize up to six different people. According to [Dutta 2012] Kinect sensor was able to capture relative 3D coordinates of markers of 0.0065 m, 0.0109 m, 0.0057 m in the x, y, and z directions. Tests provided such accuracy over the range from 1.0 m to 3.0 m. This means Kinect provides very accurate data in defined ranges.

2.3 Sensors from PrimeSense family

PrimeSense sensor family exists on the HCI market from the very beginning. Inventors [Zalevsky et al. 2007] of depth vision founded their own company called PrimeSense. Their technology expanded with mass production of the Kinect sensor. PrimeSense sensor family consists of Carmine and Capri sensors. Capri 3D sensor is an embedded sensor, which main advantage is in its small size. Sensor aims at the market of mobile devices (mobile phones, tablet computers, smart televisions). Despite its small size it has great potential. At this year’s conference Google IO 2013 a prototype tablet integration with Capri 3D sensor was presented [Crabb 2013]. All sensors are accompanied with software developer kits in the form of open source libraries like OpenNi supported by NITE algorithms.

[1]Guiness World Records http://www.guinessworldrecords.com.

2.4 Asus sensor Xtion

Xtion sensor is available in two versions, Xtion and Xtion Pro Live. It is based on the same depth vision technology as the Kinect sensor family. Xtion is promoted exclusively as a PC sensor and does not require additional power supply [Gonzalez-Jorge et al. 2013]. Asus sensors are used in the same way as Kinect, but for supporting software library OpenNi framework is used. Framework can be used in conjunction with Microsoft Kinect or other sensors based on PrimeSense technology.

2.5 Leap Motion sensor

Leap Motion sensor is a small device with enormous potential and aims to change the way we interact with computers. Sensor is installed next to the keyboard on the office desk and it provides a very accurate individual fingers detection. It is much more accurate than Kinect sensor (up to 200 times) [Hodson 2013]. This makes user interaction very precise and domain specific. The main purpose of Leap Motion sensor is not the depth vision of 3D space, but the exact finger detection and integration of these functionalities with existing applications

2.6 MYO Sensor

MYO is a not a depth vision sensing device but an intelligent armband. It detects motion in two ways: muscle activity and motion sensing [Nuwer 2013]. The MYO uses Bluetooth 4.0 Low Energy to communicate with the paired devices. It features on-board, rechargeable Lithium-Ion batteries and an ARM processor. Sensor is outfitted with proprietary muscle activity sensors. It also features a 9-axis inertial measurement unit. Muscle activity is defined by measuring electric activity in muscles, known as electromyography (EMG). Table I shows a basic comparison between the above-mentioned and other currently available devices.

Table I. Sensing device comparison

DEVICE	CONTROL	DATA SOURCE	VIDEO RESOLUTION	HCI
KINECT XBOX 360	contact-free	A/V/IR	640 x 480 30fps, 320X240 30fps	GS,VC
KINECT FOR WINDOWS	contact-free	A/V/IR	640 x 480 30fps, 320X240 30fps	GS,VC
KINECT ONE	contact-free	A/V/IR	1920 x 1080 60fps	GS,VC
ASUS XTION	contact-free	A/V/IR	1280 x 1200 30fps, 60fps	GS,VC
PRIMESENSE CAPRI	contact-free	A/V/IR	640 x 480	GS
PRIMESENSE CARMINE	contact-free	A/V/IR	640 x 480 30fps	GS,VC
LEAP MOTION	contact-free	-/V/IR	/	GS
SONY MOVE	with controller	A/V/-	640X480 60fps, 320x240 120fps	GS,VC
WII MOTIONPLUS	with controller	-/-/IR	none	GS
MYO	armband	EMG	none	GS

A-audio, V-video, IR-infrared; fps – Frames per Second
GS – Gesture Support, VC – voice control
EMG – Electromyography

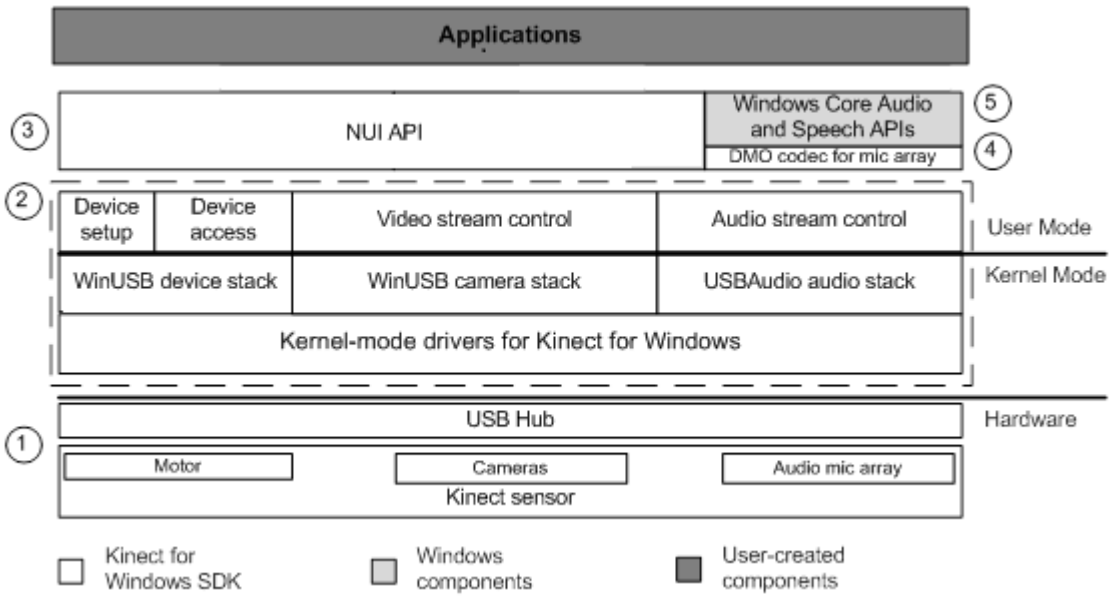

Fig. 3. Kinect SDK architecture [Microsoft 2013].

3. APPLICATION COMMUNICATION INTERFACES

The quick adoption of gesture based devices was enabled with the development of supporting software libraries (SDK). Libraries enable development of new innovative solutions and provide an upgrade to existing applications. Several libraries are publicly available. OpenKinect is an open community that uses Kinect sensor for research. The framework has a very active community, which contributes with a large set of libraries and plug-ins, thereby extending the OpenNi framework. The OpenNi core consists of several components that take care of: (i) analysis of the entire body, (ii) analysis of individual items (finger detection), (iii) recognition of user gestures and (iv) analysis of the environment (objects and people) [OpenNI 2013].

Kinect SDK [Microsoft 2013] enables software developers to develop interactive applications with support for voice command and gestures. Figure 3 shows the architecture components of Kinect for Windows SDK. These components include the following:

(1) Hardware components, including the Kinect sensor and the USB hub
(2) Windows drivers for the Kinect, which are installed as part of the SDK The Kinect drivers support:
    (a) Kinect microphone array as a kernel-mode audio device that you can access through the standard audio APIs in Windows.
    (b) Audio and video streaming controls for streaming audio and video (color, depth, and skeleton).
    (c) Device enumeration functions that enable an application to use more than one Kinect.
(3) Audio and Video Components
(4) DirectX Media Object (DMO) for microphone array beam forming and audio source localization.
(5) Windows standard APIs - audio, speech, and media APIs in Windows as described in the SDK and Microsoft Speech SDK [Microsoft 2013].

Kinect library includes advanced components known as Microsoft.Kinect.Toolkit.Controls plugin. Toolkit controls enable realization of the advanced functionality for desktop applications.

Table II. Software libraries provided with human computer sensors

LIBRARY	CAPTURE	OPEN SOURCE	SENSOR	PLATFORM
KINECT FOR WINDOWS SDK	Yes	No	Kinect	Windows
OPENNI FRAMEWORK	Yes	Yes	PrimeSense family	Windows/Linux/MacOS
OPENKINECT FRAMEWORK	Yes	Yes	Kinect	Windows/Linux/MacOS
POINTCLOUD LIBRARY	Yes	Yes	Kinect, PrimeSense	Windows/Linux/MacOS
LEAP MOTION SDK	Yes	No	Leap Motion	Windows/Linux/MacOS
THALMIC LABS MYO SDK	No	No	MYO armband	Windows/MacOS

The Point Cloud Library (PCL) [PointCloud 2013] is a standalone, large scale, open framework for 2D/3D image and point cloud processing that enables advanced 3D data analysis. It contains numerous state-of-the art algorithms that can be used to filter outliers from noisy data, stitch 3D point clouds together, segment relevant parts of a scene, extract keypoints and compute descriptors to recognize objects. PCL library can create surfaces from point clouds and visualize them [Rusu and Cousins]. The table II below summarizes current software libraries for HCI.

4. LESSONS LEARNED DURING DEVELOPMENT OF A GESTURE BASED SOLUTION ADORA

The objective of any health care institution is to optimize the length of a surgical procedure and increase its quality. ADORA[2] is an interactive physician's assistant enabling a unique presentation of information about a patient before and during surgical procedures. ADORA offers a comprehensive and integrated natural user interface experience for physicians. It is a product of field knowledge, modern information and communication technologies as well as advanced hardware. With its simple use of contact-free interaction it shortens the duration of surgeries and indirectly affects the environmental and economic aspects of healthcare. By using ADORA, physicians are able to actively participate in a surgery also outside the operating theatre. Modern methods of HCI (gestures and voice support) have been integrated into ADORA solution during development. Physicians gained control over patient data with the help of contact-free interaction. Lessons learned and best practices are summarized below and consist of following challenges:

(1) The design of graphical natural user interfaces adapted to support gestures and sound.
(2) Calibration of the sensor and the correct choice of the active detected person.
(3) Proper detection and identification of the sound source.
(4) The correct interpretation of voice command.
(5) Implementation of advanced gestures and functionality that are not supported in the basic development library Kinect (point based rotation, dynamic zoom with feedback, traceable and flexible display of DICOM images [König 2005]).

The first challenge was to design an intuitive and adaptive graphical interface that supported communication with the Kinect sensor. It is very important to include clear feedback to the user (either graphical or voice), especially when designing interactive applications. It is also recommended to include interactive help where the user has the ability to practice applications gestures and voice commands. Interactive help enables a user to learn which gestures and sound commands are required

[2]Advanced Doctor's Operational Research Assistant, http://www.adora-med.com.

to control desired functionalities. During the graphical user interface design, we had to move away from the ”classic” design of user interfaces and address the challenges associated with new ways of interaction that is typical for gesture based applications. Analysis of existing gesture based solutions and Kinect user interface design guidelines helped us in the process of natural user interface design. The user interface had to be adapted to the new user controls. Another challenge was Kinect sensor calibration. Problems arose during user detection from a variety of distances from the sensor and the detection of primary user in the presence of other surgeons. In the operating room there are at least three or more people standing close to each other. Kinect Sensor detects all users as active bodies (skeletons). Kinects correction factors helped us to calibrate the sensor according to the scope and area of usage. Detection and tracing of primary user was solved with the use of voice commands. Kinect sensor detects sound source area, which allowed us to locate the primary person in the room. Primary user is selected by executing a voice command “Follow me”. The user who executed the command becomes active person and a tracked skeleton. Detection and understanding of voice commands represented a special challenge that required knowledge of foreign languages and perfect pronunciation of these. Kinect language support is limited to thirteen languages. Recognition of voice commands takes place in stages. First, the sensor compares the sounds with the selected grammar. Grammar can be determined dynamically or is defined statically in the form of an XML document. Synonyms for each voice commands can also be defined. For example, the voice command “next” can have synonyms such as “forward” and “continue”. Detection of voice command returns sensor detection level called “confidence level ratio”. The ratio has a value from zero to one, where zero means a very weak detection and the value of one very powerful detection. In order to successfully detect voice commands the proper accent, pronunciation and intonation must be satisfied. Sound commands allow navigation through the application, item selection and object manipulation. Special digital medical image format DICOM (Digital Imaging and Communications in Medicine) [Blazona and Koncar 2007] is used in healthcare. One of the challenges was the magnification functionality. In addition to vertical and horizontal axes, depth information from the Z-axis was needed. Z-axis represents the distance from between the person and the sensor. A correct zoom factor had to be calculated. The problem was solved by adapting to the speed of hand movement. This kind of image magnification proved to be very efficient, since the user is able to control the speed and direction of zooming. It was also necessary to determine the center point of the zoom. The problem was solved with coordinate transformation from the current hand position on the widget, to a pixel point on the image itself.

5. CONCLUSION

The revolution that began in the living rooms of innovative researchers continues. Technological challenges of 3D vision have been successfully addressed. Rapid growth of sensors that enable skeleton identification, finger gestures and voice command is expected, especially the rise of applications that will incorporate their functionality. Applications that will enable contactless control of computers and other devices will slowly but surely begin to replace existing ones. Devices are already changing the way we communicate with computers. The internet of things effect is becoming more and more visible. New technologies are replacing the use of devices such as a keyboard and mouse. Touchscreens are already transforming the way we interact with mobile devices. The next step is utilization of gestures and voice commands for computer interaction in our every day. There is still some work needed to be done. Accuracy and precision decrease with range and makes usage of such devices limited to special domains. Accuracy issues can be solved with alternative types of sensing, that provide a different way of gathering motion data that is not based on depth vision (MYO armband). Sensing devices will be built into monitors, laptops, televisions and mobile devices. Given the incredible development of in-

creasingly advanced and intuitive electronic devices we can predict that in the near future systems with similar (if not better) functionalities, as seen in science fiction movies, will be used.

REFERENCES

Bojan Blazona and Miroslav Koncar. 2007. HL7 and DICOM based integration of radiology departments with healthcare enterprise information systems. *International Journal of Medical Informatics* 76, Supplement 3, 0 (2007), S425–S432. DOI:http://dx.doi.org/10.1016/j.ijmedinf.2007.05.001

Lulu Chen, Hong Wei, and James Ferryman. 2013. A survey of human motion analysis using depth imagery. *Pattern Recognition Letters* 34, 15 (2013), 1995–2006. DOI:http://dx.doi.org/10.1016/j.patrec.2013.02.006

Ross A. Clark, Yong-Hao Pua, Karine Fortin, Callan Ritchie, Kate E. Webster, Linda Denehy, and Adam L. Bryant. 2012. Validity of the Microsoft Kinect for assessment of postural control. *Gait and Posture* 36, 3 (2012), 372–377. DOI:http://dx.doi.org/10.1016/j.gaitpost.2012.03.033

Alexandra Crabb. 2013. PrimeSense^{TM} Unveils Capri, World's Smallest 3D Sensing Device at CES 2013. (2013). http://www.primesense.com/news/primesense-unveils-capri/

Julian Dibbell. 2011. *Controlling computers with our bodies*. Journal article. Retrieved July 25, 2013 from http://www2.technologyreview.com/article/423687/gestural-interfaces/

T. Dutta. 2012. Evaluation of the Kinect sensor for 3-D kinematic measurement in the workplace. *Appl Ergon* 43, 4 (2012), 645–9. DOI:http://dx.doi.org/10.1016/j.apergo.2011.09.011 Dutta, Tilak eng Research Support, Non-U.S. Gov't England 2011/10/25 06:00 Appl Ergon. 2012 Jul;43(4):645-9. doi: 10.1016/j.apergo.2011.09.011. Epub 2011 Oct 20.

M. Gabel, R. Gilad-Bachrach, E. Renshaw, and A. Schuster. 2012. Full body gait analysis with Kinect. *Conf Proc IEEE Eng Med Biol Soc* 2012 (2012), 1964–7. DOI:http://dx.doi.org/10.1109/EMBC.2012.6346340 Gabel, Moshe Gilad-Bachrach, Ran Renshaw, Erin Schuster, Assaf eng Clinical Trial 2013/02/01 06:00 Conf Proc IEEE Eng Med Biol Soc. 2012;2012:1964-7. doi: 10.1109/EMBC.2012.6346340.

H. Gonzalez-Jorge, B. Riveiro, E. Vazquez-Fernandez, J. Martínez-Sánchez, and P. Arias. 2013. Metrological evaluation of Microsoft Kinect and Asus Xtion sensors. *Measurement* 46, 6 (2013), 1800–1806. DOI:http://dx.doi.org/10.1016/j.measurement.2013.01.011

Hal Hodson. 2013. Leap Motion hacks show potential of new gesture tech. *New Scientist* 218, 2911 (2013), 21. DOI:http://dx.doi.org/10.1016/S0262-4079(13)60864-7

Kourosh Khoshelham and Sander Oude Elberink. 2012. Accuracy and Resolution of Kinect Depth Data for Indoor Mapping Applications. *Sensors* 12, 2 (2012), 1437–1454. http://www.mdpi.com/1424-8220/12/2/1437

H. König. 2005. Access to persistent health information objects: Exchange of image and document data by the use of DICOM and HL7 standards. *International Congress Series* 1281, 0 (2005), 932–937. DOI:http://dx.doi.org/10.1016/j.ics.2005.03.186

Marketsandmarkets.com. 2013. Gesture Recognition & Touchless Sensing Market (2013 - 2018): By Technology (2D, 3D, Ultrasonic, IR, Capacitive); Product (Biometric, Sanitary Equipment); Application (Healthcare, Consumer Electronics, Automotive); Geography (Americas, EMEA, & APAC). *Market Analysis* (2013). http://www.marketsandmarkets.com/Market-Reports/touchless-sensing-gesturing-market-369.html

Microsoft. 2013. Kinect for Windows SDK. (July 2013). Retrieved August 17, 2013 from www.microsoft.com/kinect

Rachel Nuwer. 2013. Armband adds a twitch to gesture control. *New Scientist* 217, 2906 (2013), 21. DOI:http://dx.doi.org/10.1016/S0262-4079(13)60542-4

OpenNI. 2013. OpenNI framework. (2013). Retrieved July 25, 2013 from www.openni.org

PointCloud. 2013. PointCloud Library documentation. (2013). Retrieved July 2, 2013 from http://docs.pointclouds.org

Stephen Prentice and Adib Carl Ghubril. 2012. Hype Cycle for Human-Computer Interaction. *Gartner* (2012).

Radu Bogdan Rusu and Steve Cousins. 3D is here: Point Cloud Library (PCL). In *IEEE International Conference on Robotics and Automation (ICRA)*. http://www.pointclouds.org/assets/pdf/pcl_icra2011.pdf

Todor Stoyanov, Rasoul Mojtahedzadeh, Henrik Andreasson, and Achim J. Lilienthal. 2012. Comparative evaluation of range sensor accuracy for indoor mobile robotics and automated logistics applications. *Robotics and Autonomous Systems* (2012). DOI:http://dx.doi.org/10.1016/j.robot.2012.08.011

Zeev Zalevsky, Alexander Shpunt, Aviad Maizels, and Javier Garcia. 2007. Method and System for object reconstruction, Patent nr. WO2007043036. (2007). Retrieved April 23, 2013 from http://patentscope.wipo.int/search/en/WO2007043036