The 13th International Symposium on Wireless Personal Multimedia Communications (WPMC 2010)

MACHINE PERCEPTION FOR IDNTIFICATION AND INTERACTION IN THE INTERNET OF THINGS

Zheng-Hua Tan
Department of Electronic Systems, Aalborg University
Aalborg, Denmark
zt@es.aau.dk

ABSTRACT

The most significant step from the Internet towards the Internet of Things (IoT) is to embrace a vast amount of objects in the global system. This first presents a nontrivial challenge to identification due to the limited applicability of Internet Protocol (IP) to these objects. They are therefore identified either via the augmentation of tiny devices such as radio frequency identification (RFID) or via natural feature identification where machine perception is indispensable. Secondly, the value of the IoT consists in the object-object and object-human interaction. Such interaction is considerably different from traditional human-computer interaction because of the large variety of objects and their surroundings. This paper presents the application of machine perception in both identification and interaction, which again are useful in finding information shadows that physical objects cast in the digital world.

I. INTRODUCTION

The *Internet of Things* is a global system interconnecting physical and digital objects where objects and humans may interaction with each other. These objects, each with its own identity, are well beyond only computers and they are our cars, luggage, household applicants, humans and so on. The objects may use sensors to gather information from their surrounding and/or use actuators to interact with it [1]. The growing world of interconnected devices indicates that the Internet of data and people of today is giving way to the Internet of Things of tomorrow [2].

The inclusion of non-computer objects is the most significant step from the Internet towards the IoT, which provides the data sphere with interface to the physical world and opens up tremendous new applications. At the same time, this presents a nontrivial challenge to identification which is not a problem in the Internet due to the use of IP address. In general, these objects can be identified either via augmentation of tiny devices like RFID or via natural feature identification such as biometric-based identification [3]. RFID is considered as the primary means to identify various objects in the physical world. Due to such reasons as cost or form factor [3], however, RFID cannot be used for augmenting every object. Here sensor networks and smart mobile devices come into play. For example, visual tags such as 2D barcodes can be used for tagging objects and high-end mobile phones with cameras are able to read 2D barcodes. This provides an inexpensive solution and moreover, there is no requirement for power supply, usually battery. Other alternatives are natural feature identification techniques, e.g. object recognition [4], which enable tremendously more things to be able to join the IoT.

Further, interacting with daily objects introduces a few new elements, for instance, no keyboard and mouse available, and the user focusing on other tasks in hand and leaving reduced attention for interaction. These together make conventional human computer interaction methods not very suitable and call for new interaction paradigms. Such interaction should be natural, effortless and even invisible. In addition, interaction that analogizes the real-life physical interaction may be favored in order to reduce cognitive load. In [5], a number of Embedded Interaction prototypes are presented where interaction is realized either implicitly or explicitly on the basis of information collected by various sensors embedded in physical objects such as cutting board and knife in a kitchen environment. As another example, egocentric interaction uses as input commands the changes in the spatial relation between user and device (object) [6].

In general, the wide deployment of sensor networks and the growing use of smart mobile devices make signal capturing from the physical world a lot easier. The obtained signals are required to be apprehended by computing machines to make them meaningful. The process of sensing and interpreting sounds, pressure, images or other content, is called machine perception. In this paper we discuss the use of machine perception in identification and interaction which again are useful in finding information shadows that physical objects cast in the digital world.

The paper is organized as follows. Section II describes the concept of information shadow. Machine-perception based identification and interaction are presented in Sections III and IV, respectively. Summary is given in Section V.

II. INFORMATION SHADOW

Our work and life are increasingly dependent on the World Wide Web, which is now beyond a collection of static HTML pages that describe scattered things in the world [7]. Growingly, the Web is the world, meaning everything and everyone in the world casts an *information shadow* in the digital world [7]. In other words, every identified object exists simultaneously in the physical world and in the digital world.

The concept of information shadow has been coined and developed by a number of people. As Greenfield put it in [8], RFID and 2D barcode technologies provide a way to bring physical objects into the data sphere so as to provide them with an informational shadow. Kuniavsky calls the digital representation of an object, accessed through a unique ID, the object's information shadow [9]. Thanks to wired and wireless networking, we can instantaneously see the world of information shadows while or through interacting with the world of objects [9].

Finding information (shadow) has not been easy. We are used to typing URL into a browser or keywords into search engines like Google™ and bing™ to obtain relevant content. This is seen as a convenient and efficient way of getting information. But in the era of the IoT, this might be obsolete as things will change dramatically.

With ubiquitous sensing and computing, the physical world is tightly linked to the digital world. Especially sensors in smart phones are revolutionizing the interface between the two worlds by including sensors like microphones, cameras, motion sensors and location sensors. Finding information shadow goes beyond typing into computers and with the support of tagging and computer perception, the world itself becomes the interface or part of the interface.

An interesting application of such kind is the augmented-reality (AR) browser developed by Layar™ [10]. It combines live video of the real world with computer graphics overlay and other relevant data. In this way, a user of the AR browser can walk down a street and receive instant information based on the location and direction estimated by Global Positioning System (GPS) and compass devices. Basically, pointing the camera of a phone at a specific location will instantly overlay information about the location in the viewfinder [10].

Rather than using positioning methods, machine vision technique is deployed for obtaining the desired information on a physical object by simply taking a picture of it in [4] where GPS can be used optionally.

The development of the Web and sensing technologies is jointly revolutionizing the way how information is gathered and accessed.

III. IDENTIFICATION

While computers are connected to the Internet via IP, the objects in the IoT are connected to the Internet through a variety of means including tag readers and general sensors. Due to the footprint of IP, many objects (e.g. passive RFID, sensors and mobile devices) will not be able to connect to the Internet and be identified through IP.

Objects in the IoT are otherwise identified through two distinct classes of technologies [3]. One is tagging things by augmenting objects with tiny data carriers: visual markers like 2D barcode, magnetically encoded carriers and radio communication devices such as RFID, near field communication, wireless local area network and Bluetooth; the other is sensing and identifying non-augmented objects by natural feature identification, for example, biometrics [3].

Location sensor like GPS combined with compass also provides means for identification of point of interest as demonstrated by the Layar AR browser.

In [4] a computer vision system is developed to enable the use of a camera phone for identifying objects without markers. In their system, object recognition is realized by using local visual features, global geometry and optionally metadata such as location information to boost the performance. The advantages of this technology are no markers required to be attached to the objects and its capability of identifying objects from distance as well as on screens [4].

As compared with machine vision, audition on the other hand has limited use in identification since not every object generates (distinguishable) sounds and further not many objects generate sounds all the time. Speaker recognition (voice biometrics) is one of the few techniques applicable in this context. There are two branches of technology for identifying the person who speaks by using speaker specific information included in speech. One is speaker identification, the process of determining which of the registered speakers a given utterance comes from; the other is speaker verification which verifies whether a speaker is the identity that she/he claims to be. A Gaussian mixture models based framework is commonly used in text-independent speaker recognition applications [11]. The technology appears to be reaching the maturity required for deployment in various applications.

Nowadays camera and microphone sensor networks are widely deployed for monitoring and can be used for identification in traffic and in critical infrastructures like airports. Mobile phones equipped with cameras provide mobile sensors for such purpose in the user's immediate surroundings. Identification of objects provides a means to hyperlink the objects with the digital world, i.e. finding the information shadows. Identification of objects and its use in finding information shadow are illustrated in Figure 1.

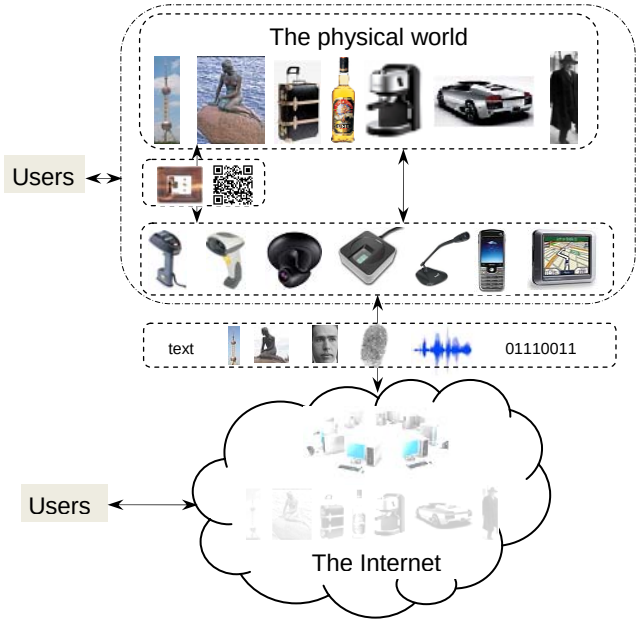


Figure 1: Identification of objects in the IoT.

IV. INTERACTION

Interaction with daily objects differs from interaction with personal computers. It is desirable to interact with objects by using natural communication channels including visual, auditory and tactile and in a natural, effortless and even invisible way.

A. *Natural interaction*

As objects in the IoT are integral parts of our daily work and life, there is a strong desire for a natural interaction with them. Naturalness implies freedom from constraint, formality

or awkwardness, and such interaction should be like interaction between humans, i.e. intuitive, multi-modal and based on context.

Speech is the most natural means of interaction for human beings and it has the unique advantage of no requirement for carrying a device to use it since we have our "device" with us all the time [12]. Speech recognition, in the core of speech interaction, is the process of converting a speech signal to a word sequence. Based on the principles of statistical pattern recognition in particular hidden Markov models, modern speech recognition systems are useful in many applications under controlled environments. Variations such as background noise and reverberation are important issues to be dealt with in the real-world deployment. Further, objects in the IoT are generally characterized as having restricted resources and being interconnected. Speech recognition systems mostly having a high complexity are therefore optimized towards low-resource implementations or adopt a distributed architecture to make use of powerful servers as illustrated in Figure 2 [13].

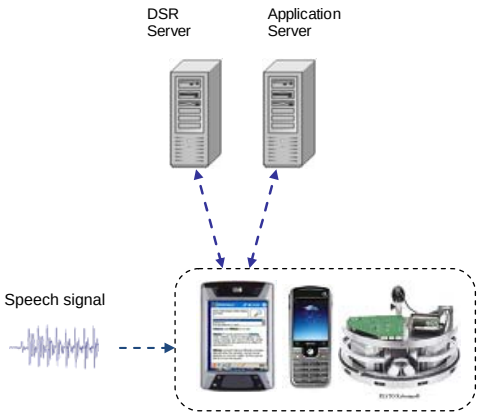
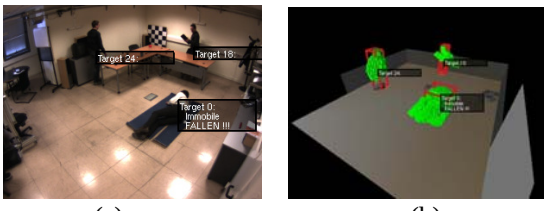


Figure 2: A distributed speech recognition (DSR) system with a recognition server and an application server.

As in ambient intelligence, interaction technology in the IoT should be embedded in the environments and be present only when needed. The realization of such interaction relies on context information, for example, a person's identity, location and emotion, which can be automatically extracted from audio or audio-visual signals [14].

In [15] a set of fixed and calibrated cameras are used to detect and model the bodies of people in three dimensions, as shown in Figure 3.



(a) (b)

Figure 3: Three-dimensional sensing of multiple persons [15]: a) an image with persons and information overlay, and b) detected foreground and information.

Figure 3 demonstrates that the system can detect the location and posture (e.g. falling down) of persons. Sensing the presence and state of people is of importance in many applications such as assistive living.

B. Egocentric interaction

Accelerometers are widely deployed to enable mobile devices to use movements as input commands. With the advances in machine vision, camera-phone motion can also be estimated by using edge detection, region detection, optical flow and so on. Image differencing and correlation of blocks are used in [16] and experimental results show that camera phone can serve as a handwriting capture device performing large-vocabulary, real-time text input, faster than the multi-tap method. This type of interaction is bound to the physical world rather than the user. Instead, an egocentric interaction modality exploits the spatial relation between user and device and uses changes in this relation as input commands [6], [18].

1) Infrared-Diodes Based 3D Interaction

Infrared light and camera techniques are used for tracking head and fingers in [17] in a 3D environment to enhance the immersion factor of computer applications. In this system, the user's fingers and head are infrared-lighted by wearing gloves and glasses embedded with infrared diodes, as shown in Figure 4 (a) and (b).

(a) (b) (c)

Figure 4: Vision-based 3D interaction [17]: (a) glasses, (b) glove and (c) graphical user interface.

The two main factors for infrared based detection are light diffusion angle and light power. To improve the infrared light intensity, boosting is implemented by frequently switching the diode ON (peak current with a high voltage) and OFF (no current flow to cool down and to limit the power consumption) [17]. Nintendo Wiimote and an off-the-shelf webcam are experimentally compared and it is found that the Nintendo Wiimote provides a high accuracy whereas the webcam has a good detection range. To evaluate this interaction method, an application for designing 3D architects has been implemented by using DirectX as shown in Figure 4 (c). Usability tests show that the vision-based tracking system enables the user to smoothly handle objects by using two fingers and view the scene from different angles in a 3D environment simply by moving his head/body around.

2) Face Based Interaction for Mobile Device

To free from wearing intrusive devices like glasses and gloves, face detection and tracking is used to create an

egocentric interactive space between user and device in [18]. The prototype system uses face tracking to determine the spatial relation between user and device. Two dimensions of freedom are deployed: The user can move the mobile device up-and-down for panning, and back-and-forth for zooming in-and-out. A portable music player with a large collection of songs is developed for use case study, where the egocentric interaction modality is used to browse the music library. The study shows that the interaction method is effective, but inefficiency when compared with alternatives such as touch screen [18]. It is also evident that this interaction method is more suitable for panning than for precise selection of specific items.

C. *Embedded interaction*

Networked gadgets with sensors and actuators are embedded into daily objects such as cutting board, knives and tables in a kitchen environment in [5]. These gadgets collect information unobtrusively to support both explicit and implicit interaction. The cutting board acts as a mouse pad by using the load cells underneath to measure the weight change of a finger on it. A knife augmented with sensor is able to detect the type of food being cut. Further the dining table is equipped with sensors for recognizing table-top interactions during meals.

Scratch Input [19] is an input technique that relies on the acoustic signal generated by dragging a fingernail over the surface like wood or wall paint. Example applications demonstrate its potential for interaction and an accuracy of 90% is achieved for six Scratch Input gestures.

D. *Interaction through tagging.*

Tagging can make interaction and finding information shadows much easier by eliminating the need for human inputs or interferences. For example, we can place on an object a 2D barcode that can be a link to a webpage for providing relevant information. In [20], the built-in cameras of mobile phones are used as sensors for reading 2D visual codes attached to physical objects or showing on screens, in order to retrieve object-related information.

Nabaztag:tag and Mir:ror are two interesting examples of interaction through tagging developed by Violet™ [21]. Nabaztag:tag is an Internet-connected mini-robot being able to talk, hear, smell and move. The mini-robot can also detect RFID tags and provide information in a nonintrusive way by using light. The other system, Mir:ror, can detect the objects shown to it and take actions based on the detection. The system can communicate over the Internet, recognize objects with RFID tags, and trigger programmed actions including launching applications, connecting to a website and updating information on social networks.

V. SUMMARY

This paper describes the concept of information shadow and investigates the application of machine perception in the Internet of Things for identifying objects and interacting with them. Several prototype systems presented in the paper show that machine perception is applicable in this context. It is also evident that object identification and interaction can facilitate finding information shadows in the digital world, thus achieving the fusion of the physical and virtual worlds.

REFERENCES

[1] Commission of the European Communities, *Internet of Things – An action plan for Europe*, 2009.

[2] *ITU Internet Reports 2005: The Internet of Things – Executive Summary*. ITU, Geneva, 2005.

[3] *RFID and the Inclusive model for the Internet of Things*. CASAGRAS Final Report, 2009.

[4] T. Quack, H. Bay and L. Van Gool, "Object Recognition for the Internet of Things," in *Proc. Of Internet of Things 2008*, Zurich, Switzerland, March 2008.

[5] M. Kranz, P. Holleis and A. Schmidt, "Embedded Interaction: Interacting with the Internet of Things," *IEEE Internet Computing*, vol. 14, no. 2, March 2010, pp: 46-53.

[6] T. Pederson and D. Surie, "Towards an Activity-Aware Wearable Computing Platform based on an Egocentric Interaction Model," in *Proc. of IFIP UCS 2007 Conference on Ubiquitous Computing Systems*, Springer LNCS 4836, pp. 211-227.

[7] T. O'Reilly and J. Battelle, "Web Squared: Web 2.0 Five Years On," Web 2.0 Summit, San Francisco, CA, USA, October 2009.

[8] A. Greenfield Everyware, *The dawning age of ubiquitous computing*, New Riders Press, March, 2006.

[9] M. Kuniavsky and A. Creech, "Information Shadows: How Ubiquitous Computing Serializes Everyday Things," The Serials Librarian, Vol. 56, No. 1-2, Jan 2009, pp: 65–78.

[10] http://www.layar.com/

[11] R. Saeidi, P. Mowlaee, T. Kinnunen, Z.-H. Tan, M.G. Christensen, S.H. Jensen and P. Fränti, "Signal-to-Signal Ratio Independent Speaker Identification Co-Channel Speech Signals," in *Proc. Of the 20th International Conference on Pattern Recognition (ICPR)*, Istanbul, Turkey, August 2010.

[12] Z.-H. Tan, R. Haeb-Umbach, S. Furui, J.R. Glass and M. Omologo, "Introduction to the Issue on Speech Processing for Natural Interaction with Intelligent Environments," *IEEE Journal of Selected Topics in Signal Processing*, 2010.

[13] Z.-H. Tan and B. Lindberg (eds.), *Automatic speech recognition on mobile devices and over communication networks*, Springer-Verlag, London, Feb. 2008.

[14] J. Schmalenstroeer and R. Haeb-Umbach, "Online Diarization of Streaming Audio-Visual Data for Smart Environments," *IEEE Journal of Selected Topics in Signal Processing*, 2010.

[15] M. Andersen, R. S. Andersen, N. Katsarakis, A. Pnevmatikakis and Z.-H. Tan, "Three-Dimensional Adaptive Sensing of People in a Multi-Camera Setup," in *Proc. Of EUSIPCO 2010 – the 18th European Signal Processing Conference*, Aalborg, Denmark, August 2010.

[16] J. Wang, S. Zhai and J. Canny, "Camera phone based motion sensing: Interaction techniques, applications and performance study," in *Proc. of the 19th annual ACM Symposium on User Interface Software and Technology*, October 2006, Montreux, Switzerland.

[17] T. Luel and F. Mazzone, *Vision-Based Human Interaction Devices in a 3D Environment: Using Nintendo Wiimote, Webcam and DirectX*, Master Thesis, Aalborg University, 2009.

[18] M.H. Justesen, B. Hansen, A. Grønne, C.S. Petersen, M.D. Jensen, M.P. Andersen, *Design and evaluation of a handheld device creating an egocentric interactive space between user and device*, Project Report, Aalborg University, 2010.

[19] C. Harrison and S.E. Hudson, "Scratch Input: Creating Large, Inexpensive, Unpowered and Mobile finger Input Surfaces," in *Proc. of the 21st Annual ACM Symposium on User interface Software and Technolog*, 2008, New York, NY.

[20] M. Rohs and B. Gfeller, "Using Camera-Equipped Mobile Phones for Interacting with Real-World Objects," in A. Ferscha, et al. (Eds.): *Advances in Pervasive Computing*, Vienna, Austria, April 2004.

[21] http://www.violet.net/