Cellular Phones as Information Hubs

Minoru Etoh
Research Laboratories, NTT DoCoMo
3-5 Hikarino-oka, Yokosuka, Kanagawa 239-8536 Japan
etoh@ieee.org

ABSTRACT

This is a survey of how cell-phones are interacting in the real world with I/O devices for information retrieval. Many cell-phones are now equipped with I/O devices including GPS, microphones, CCD/CMOS cameras, and motion-sensors. Thus content delivery, interaction with contents, and e-commerce are going to be associated with the user's actual environment which will bring us context-aware(real-world aware) capability. This provides a clear distinction between cell-phones and mobile laptop PCs. Interaction is the key when considering mobile information retrieval. Based on 3G broadband and low-latency technologies, the "always-on" mobile infrastructure provides on-demand and real-time information retrieval capability that meshes well with today's communication culture, which we call "the third generation of the cell-phone culture". In the next two decades, billions of cell-phones will be connected to the ubiquitous server cloud. The result will be a different level of real-word aware information retrieval that cannot be predicted accurately.

Categories and Subject Descriptors

H.4.3 [Information Systems Applications]: Communications Applications; H.3.0 [Information Storage and Retrieval]: Web search

General Terms

Design, Experimentation, Human Factors, Standardization

Keywords

Mobile, Cellular Phones, Information Retrieval, Real-World Aware, Sensor Hub

1. INTRODUCTION

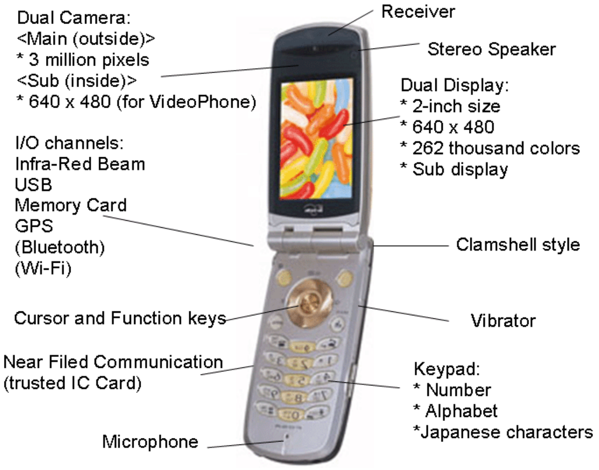


Figure 1: Typical Cell-phone in Japan.

The term "Ubiquitous Computing" has become a buzz word; it implies the paradigm of pervasive computing, ambient intelligence or real-world aware invisible computing, where many computers serve each person, and the nomadic devices carried by each person communicate with the environment as PC peripherals. This concept hasn't been implemented a commercial service, a similar concept is now being realized by cellular phones (cell-phones) equipped with several I/O devices. Fig. 1 illustrates a typical Japanese cell-phone; it has many I/O devices such as a microphone, cameras, GPS, and near field communication device.

Most discussions of the generations of mobile networks define generation as the attributes of the physical links, i.e. 3G and 3.5G stand for CDMA and HSPA/1xEVDO respectively[1]. From the viewpoint of their usage , however, we can make a different definition of generation. Here, to avoid any confusion in the terminology with " wireless technology generation" , let us use "stage" to describe different cell-phone cultures. As depicted in Fig. 2, the first stage of cell-phone culture existed up to and including early 2G . We used cell-phones for speech communication where the technical advances provided better spectrum efficiency. Late 2G (say 1997) saw the always-on packet switch connection network emerge simultaneously with the micro-browser in 1997; the first interconnection service called i-mode was launched in 1999. i-mode represents the beginning of the second stage of cell-phone culture. We used cell-phones as information tools for e-mail, web-browsing, and entertainment via multimedia contents. The evolution of 3G networks enriched

SIGIR 2008 Workshop on MobIR '08
July 24, Singapore

[1] For further details of mobile system technologies, see [5].

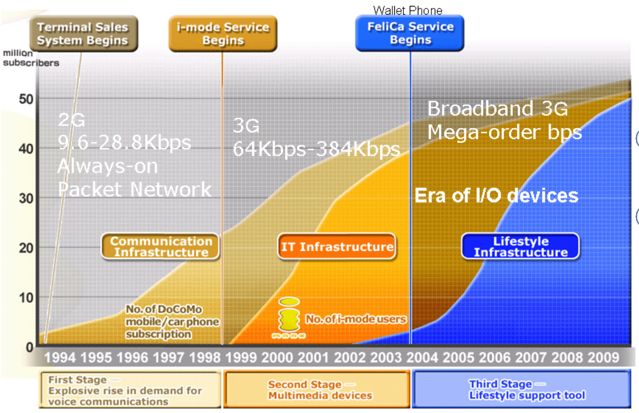


Figure 2: Three Service Stages.

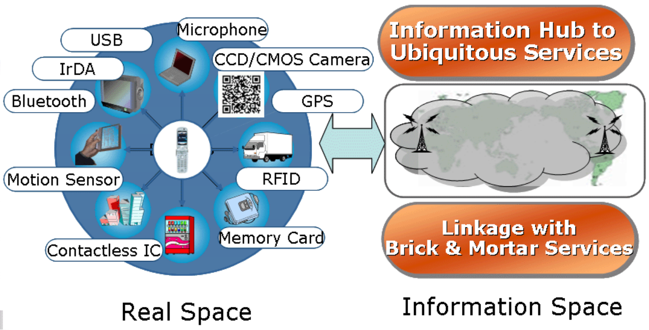


Figure 3: Cell-phone as Information Hub.

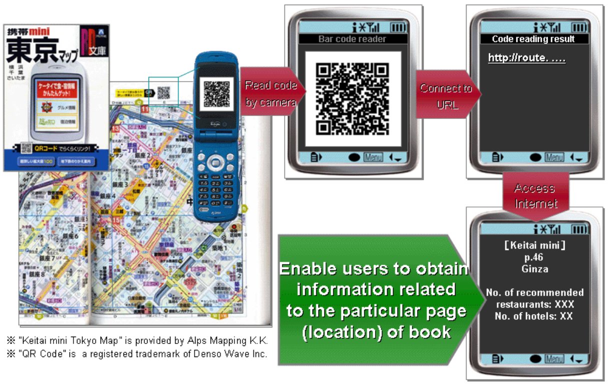
service quality. The third stage of cell-phone culture began around 2004 in Japan, when wallet-phones were launched with Felica IC card functionalities[2]. As already shown in Fig. 1, we have now entered the era of I/O devices and are testing the limits of technological convergence. Recent broadband and low-latency 3G technologies are strengthening the movement toward the third stage of cell-phone culture. This has created a remarkable divide between lap-top nomadic PCs and cell-phones. The former mainly provide information processing on the road, while the latter supports our daily life.

In the third stage, the devices are sensors connected to the Internet through the always-on cellular network. Each cell-phone works as an information hub that bridges the real world and the Internet as depicted in Fig. 3. The structure of this hub does not exactly equal ubiquitous computing though, the two concepts share the same view where sensors around each person interact with the real-world environment; for example, use of GPS enables locatoin-based search services, and use of a camera enables a two dimensional barcode reader which invokes the related information. Thus information retrieval (IR) via cell-phones is highly related to real objects. Please note that interacting with the real-world environment demands a low-latency network. Thanks to the recent progress of 3G-and-beyond mobile networks, latency is continually dropping and error is not crucial in actual operations[6]. In the following sections, we describe the cameras, microphones, GPS, and contactless ICs discussed as I/O devices. We will show how IR is related to the third stage of cell-phone culture.

2. CAMERA

2.1 Image Barcode

The camera is a powerful sensor that connects real objects to the Internet. One of most successful connecting applications is the “QR code.” The code was originally designed for logistics management in vehicle manufacturing; its original specification was standardized in 1999, and a corresponding ISO International Standard, ISO/IEC 18004, was approved in June of 2000 and updated in 2006 (ISO/IEC 18004:2006). Nowadays most (say, 90%) camera phones in Japan are equipped with QR-code reader software, which extracts QR Code images from the user’s environment uses the data, typically URLs, to realize IR via the phone’s browser (See Fig. 4). A two-dimensional QR code can contain up to 4,296 alphanumeric characters or 2,953-byte binary data [4]. Ohbuchi et al.[14] reported a detailed DSP implementation on a cell-phone platform. Owing to its robustness, QR codes are now used in many cell-phone applications.



※ "Keitai mini Tokyo Map" is provided by Alps Mapping K.K.
※ "QR Code" is a registered trademark of Denso Wave Inc.

Figure 4: QR Codes and Usage.

Two dimensional barcode access is not limited to QR codes. “ColorCode”[3] developed by a Korea-based company is attracting content providers for mobile commerce advertisement applications. The proprietary specification uses color information rather than 2D black and white dot patterns, ColorCode can provide more flexible representations that suit advertisements and natural pictures. If we colorfully paint objects or put color markers on those in a real-world environment with a meticulous calculation according to the code design rule, we can connect the real objects in a more natural way through the lens.

Regarding the acquisition of color-based access information through the lens, Microsoft and the International Standard Audiovisual Number International Agency (ISAN-IA), in 2007, announced that ISAN-IA has licensed Microsoft’s new High Capacity Color Barcode (HCCB) technology[13] which was developed to assist in the identification of commercial audiovisual works such as DVD contents. The HCCB format uses combinations of colors (i.e., 4 or 8 colors) to create each symbol. These and other color-based technologies will evolve to become an integral part of printed media which

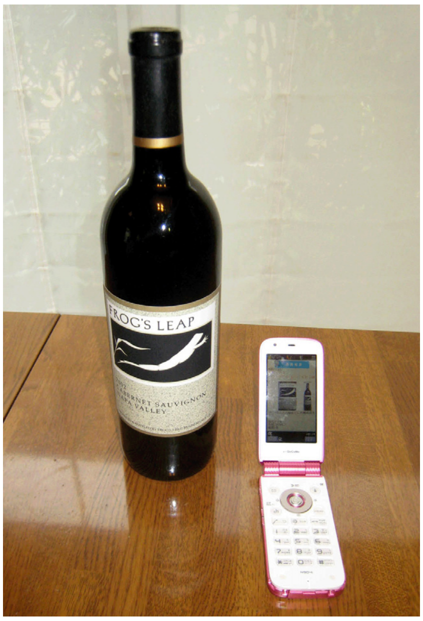


Figure 5: Wine Description by Scale Invariant Features.

will evolve into data-rich sources. Please note that the data-richness enables PKI infrastructures so as to realize secure connections between cell-phones and services[2].

2.2 Image Fingerprinting

Apart from the barcode approaches, connecting real objects to the Internet via natural scene capture is being developed. TRECVID, which is the abbreviation of TREC(Text REtrieval Conference) Video Retrieval Evaluation, is a well-known series of workshops specialized for information retrieval research areas for the content-based retrieval of video (see http: //www-nlpir.nist.gov/projects/trecvid/). A video retrieval method typical for that community takes the "bag of visual words" approach, where, analogous to document retrieval, keypoints extracted as salient image patches are collected from images. For keypoint representation, Scale Invariant Feature Transform (SIFT)[11] is typically used. That technology was adopted by a US-based company, Evolution Robotics, and has been commercialized as the visual search engine called "ER Search" for camera phone installation. Fig. 5 shows an example of an ER search, in which the wine label was captured, its features were transmitted to a backend server, and the wine's description is retrieved and displayed. Thus, image fingerprinting makes it practical to locate a registered instance in an image database.

Note that cell-phones well support feature extraction and result presentation. We introduce a similar framework for speech recognition in the following section.

3. MICROPHONE

The microphone is a vital I/O device for speech communication. All cell-phones are equipped with this I/O device and a speaker (i.e, receiver). The audio equivalents to video include a sound barcode for tagging, audio fingerprinting for music retrieval, and distributed speech recognition.

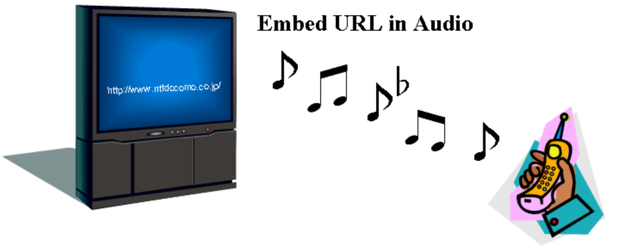


Figure 6: Sound Barcode (Audio Tag).

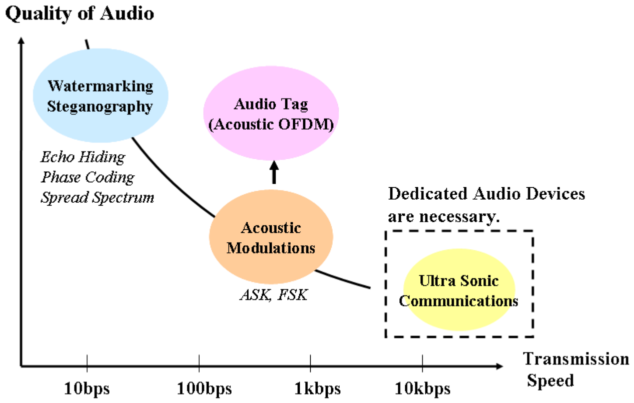


Figure 7: Existing Methods for Audio Data Transmission.

3.1 Sound Barcode

As shown by the image barcodes, we may tag audio contents with information as sound tags. Although the data rates offered by sound are relatively low and the sound can be annoying, it enables no-cost approaches to the linkage of traditional media with cell-phones. For example, a commercial message on TV or radio can transfer the URL of the company to user's cell-phones through the accoustic medium as depicted in Fig. 6. Existing acoustic communication methods that use data hiding techniques, offer very low data rates, less than 40 bps. These methods take many seconds to transmit even a short message such as a simple URL or e-mail address. Higher data rates, from hundreds of bps to more than 1kbps, can be achieved by sending the modulated signal directly, but the resulting sounds are noisy and often annoying. Ultrasound acoustic communication can achieve even higher data rates and is imperceptible to human ears, but ordinary low cost loudspeakers and microphones do not support ultrasound, which prevents the use of ordinary audio devices. Fig. 7 summarizes this discussion.

Matsuoka et al. proposed the audio tagging technique called "Acoustic OFDM" [12]. Adjusting the power of the OFDM carriers so that they match the spectrum envelope of the high-band component of the audio signal ensures that the modulated data signal does not degrade the original audio quality. It achieves data transmission rates of about 240-720 bps simultaneously with the playback of conventional audio streams when the audio streams have a certain power in the high frequency band such like rock, pop and jazz. Thus users of mobile cell-phones can access websites via URLs without manual input. To the best of our knowledge, this is the only audio-tagging technique that

[2] Authentication is one of major technical challenges in bridging the real world and the Internet.

achieves barcode-level information transfer rates (see also http://www.nttdocomo.co.jp/english/corporate/technology /rd/tech/main/acoustic_ofdm/index.html.).

3.2 Audio Fingerprinting

It is said in [18] that if the aim is not necessarily to identify a work, but a recording, audio fingerprinting techniques perform quite well. The instance identification has been commercialized, while the class identification is still a major research topic. All phone-based systems for identifying popular music use some form of audio fingerprinting. The query is a fragment of a recording captured by a cell-phone outdoors; the data center holds a complete music data base, contents with meta data. The search engine computes a similarity measure to identify the closest recording in the database and output its metadata. Research on instance-based music search engines became popular around 2000, and several have been commercialized. There are two major approaches: region(area)-based matching and feature point-based matching in the time-frequency domain. Histogram matching by Kashino[9] and binary spectrum and hashing by Haitsma[7] are categorized as the former approach, while Shazam by Wang [19] is categorized as the latter one. A hybrid approach which combines region-based and point-based tecniques may also exist (e.g., [10]). Those schemes have been proven robust in practical applications such as broadcasting monitoring, music fingerprinting (in a US-based company, Gracenote), and music retrieval by cell-phones.

3.3 Distributed Speech Recognition

In 2007, NTT DoCoMo released speech recognition capable cell-phones to the market[8]. Those cell-phones are high-end but have attracted a wide following. Its speech recognition functionality is based on ETSI (European Telecommunication standardization Institute) STQ-Aurora DSR Working Group's standard, DSR. DSR stands for distributed speech recognition, in which the recognition front-end is located in the terminal and is connected over a data network to a remote back-end recognition server. Fig. 8 shows the DSR architecture with standardized options which are FE(Front End), AFE(Advanced FE), XFE (extended FE) , and XAFE (Extended Advanced FE). The extended options realize speech reproduction at the server-side in addition to feature extraction, for which Mel-Frequency Cepstral Coefficients (MFCC) are used. Advanced options adopt 16Khs sampling. DoCoMo has adopted XAFE for their speech recognition service. Since the DSR uses 16Khz sampling for voice, where prior phones used 8Khz sampling for speech communication, this service has opened a new era in mobile speech processing. For further technical details please refer to the ETSI-DSR documents [15, 16]. Since ETSI doesn't specify the protocol between cell-phones and servers, DoCoMo developed its own DSR protocol, which runs over HTTP, after considering harmonization with web applications such as voice search. The DSR architecture uses the data channel, and thus it facilitates the creation of exciting new applications that combine voice and data. Please note that the separation of feature extraction and recognition has two major advantages: (1) reducing the burden at the client-side and (2) server-side scalability to support more time-space heavy technologies. The DSR framework is being used for diverse applications such as dictation for E-mail writing, route navigation in public transportation, and speech interpretation. As regards interpretation, current DSR applications with DoCoMo cell-phones (as of 2008) support Japanese, English, and Chinese.

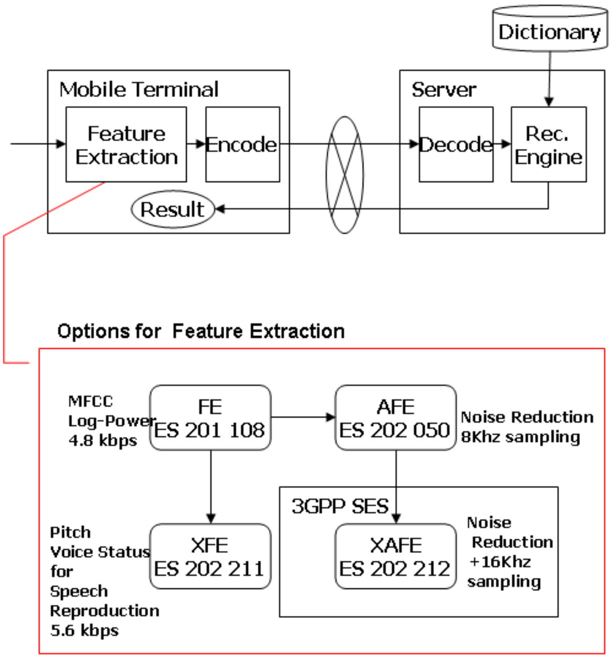


Figure 8: DSR System Architecture.

Please recall that the visual search engine "ER-Search" is presumed to have adopted a similar client-server architecture and so separates feature extraction from recognition. Back-end servers generally can bear both computational and storage burdens. One concern of such a wireless recognition system is a slow and narrow communication channel. With 3G-and-beyond mobile networks, however, that concern is eliminated. As the cell-phone evolves from a communication tool to an information hub, we will see the emergence of a distributed image/pattern recognition framework in a way of opening the specification of payloads and protocols such as XAFE and HTTP in the DSR framework. Suppose many cell-phones which are working as the front-end, through open APIs, connected to a server cloud. We may have next generation of cloud computing at the third stage of cell-phone culture. Here, cloud computing means a multi-user application cloud (also known as software-as-a-service or "Saas") based on Web APIs. Open web APIs like DSR foster interoperable mult-tenant and multi-user IR applications.

4. GPS AND NEAR FIELD COMMUNICATION

The new cell-phone culture makes the real world "event-driven." When you go to a station or a restaurant, when you buy some goods, all the events in which you interacted using your cell-phone can be connected to the internet. Let us see how we are connected in the event-driven world.

4.1 GPS

The penetration ratio of GPS cell-phones is about 40%-

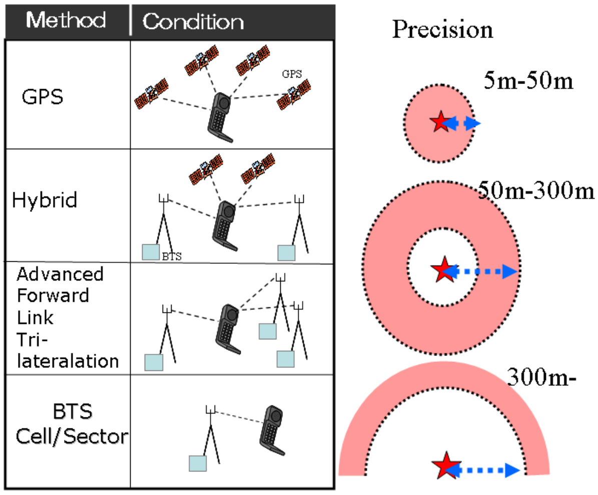


Figure 9: Cell-phone GPS and its precision

50% in Japan as of 2008, and the government is targeting 90% penetration with 3G cell-phones by 2011 for an emergency service as E911 in US. Given this high penetration ratio, GPS-enabled services such as GPS-assisted directory service and man-navigation will become extremely popular and become a commodity in our daily life. The broadband 3G mobile network supports the positioning process in two ways; (1) initializing satellite positions and (2) calculating the position even without adequate GPS links. Depending on the number of available GPS links, the precision varies as depicted in Fig. 9. We call such a scheme "Assisted GPS." For further details of assisted GPS with emphasis on communication protocols, please refer to [1].

When using the GPS function in the context of web-browsing and if you need to request the user's GPS position at HTML document, all that is needed is the rather simple tag of [3]

<A HREF="http:// www.docomo.co.jp/gps.cgi" lcs>

, where the lcs attribute is NTT DoCoMo's proprietary extension to invoke the GPS process at the cell-phone. This triggers the display of a pop-up window asking the user which location, current location, location from history, or location from phonebook, should be sent. If the current location option is selected, the cell-phone will then acquire the current location information and a screen for confirming the transmission of location information opens. If the user agrees to send the information, it is converted and sent as a location information URL. The example of

http://www.docomo.co.jp/gps.cgi?lat=%2B35.00.35.600&
lon=%2B135.41.35.600&geo=wgs84&x-acc=3

indicates the user's location as latitude: +35.00.35.600, east longitude: +135.41.35.600, additional information includes geodetic system:wgs84, and accuracy level: 3 which means better than 50m precision. Thus, the GPS data can be associated with web applications, and you can find restaurants, banks, or gas stations in your neighborhood via your GPS-equipped cell-phone. A comprehensive discussion of GPS-enabled applications is found in [17].

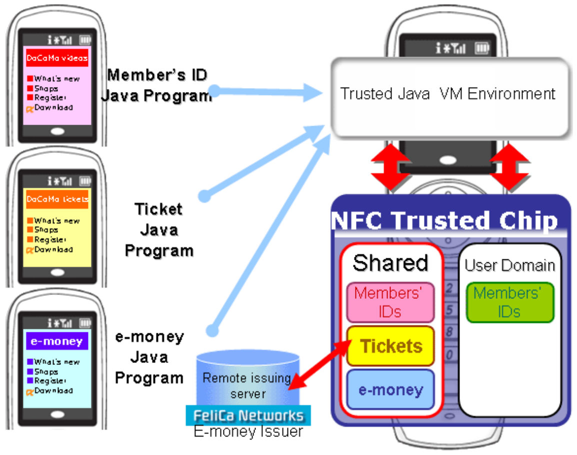


Figure 10: NFC and Java for Secure Transactions.

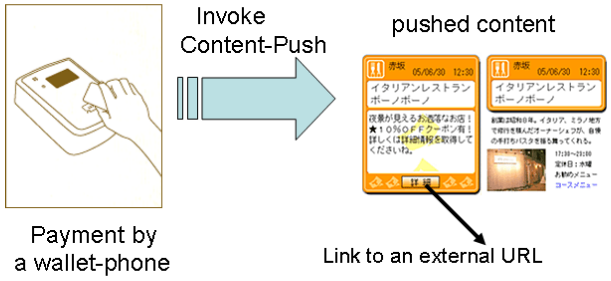


Figure 11: Event-driven Advertisement.

4.2 Near Field Communication

Near Field Communication(NFC) is the standard ISO/IEC 18092 developed by SONY and Phillips. It uses the frequency of 13.56Mhz. FeliCa[2] is SONY's commercial ISO/IEC 18092 compliant IC-chip card system, it has been widely used in Japan for public transportation such as train passes since 2001. It has been working as an IC card with or without cell-phones, and was adopted as DoCoMo's wallet-phone platform in 2004; it is now used throughout Japan.

The system was originally used for e-money service in Hong Kong, Singapore, and Japan, but current usages are quite diverse and include member identification (i.e., authentication), e-ticket, secure token transaction and so on. The penetration ratio of wallet-phones in Japan is about 40-50% as of 2008. The combination of the secure IC-chip card system with the secure Java execution environment provided by cell-phones promises significant synergistic effects: (1) viewing the current status of the trusted device (i.e., card), (2) charging/paying e-money from/to a remote e-money issuer/recipient over the mobile network, and (3) disabling remotely the function when lost.

Felica (or NFC in general) can trigger advertisements when a wallet-phone interacts with an e-money reader/writer[4]. Fig. 11 depicts how it works when paying e-money at a shop.

[3] See also http://www.nttdocomo.co.jp/english/service/imode/make/content/gps/

[4] See also http://www.nttdocomo.co.jp/english/service/osaifu/toruca.html

Simply wave your wallet-phone over the reader/writer at the store. Gift coupons or shop information are automatically stored on your cell-phone. By pressing the (details) button for the pushed content(e.g, coupon or gift card), you will be able to access details and the latest information via the Internet. Thus what you did in the real world invokes what you will see in the Internet.

5. DISCUSSION AND CONCLUSION

In the third stage cell-phone culture, cell-phones are vital life supporting tools, where the I/O devices are sensors connected to the Internet through the always-on 3G-and-beyond network. Among the sensors, as introduced already, GPS is highly associated with mobile search. A mobile search service sometimes specializes its search category, for example, to restaurants, toilets, parking places, transportation methods, local jobs, and so on. As of 2008, a Japan-based map database company, ZENRIN DataCom, offers a java-based navigation application, optimized for the small screen of a mobile phone with GPS, the DSR input, and a server-side map database. It is a best practice, to our knowledge, in combining useful sensors for local IR. Natural speech recognition by speaking “the nearest hospital” or “the nearest train station” to the cell-phone guides you to the destination.

Regarding the search inputs, Yi et. al. surveys mobile query patterns[20]. The author agrees on that survey in general and would like to comment as follows. In our observation, query patterns to the Internet with cell-phones are mixed up by two groups of general population: “solely or mainly from mobile” and “mainly from PC, mobile supplemental.” The mobile query patterns of the two groups are very different. The former group uses cell-phones as PC substitutes. On the other hand, the latter group owns PCs, and shows very purposive intentions with cell-phones, in other words, mobile specific IR activities to time-space localized contents as introduced above. In that sense, when discussing mobile IR, we should have a clear distinction between the usage patterns of the two groups.

What we have seen is that cell-phones connect real environments to the Internet. The key advance of cell-phones is bridging the real world and the virtual world. Cameras, microphones, and GPS and near field communication devices are producing real sensor data, where the cell-phones are assigned feature extraction while a backbend server carries the burden of time-consuming recognition processes; 3G-and-beyond mobile networks enable such a client-server architecture with low levels of latency. The combination of the low latency wireless broadband network, the sensors, and a server cloud as distributed service platforms with open APIs is the point to realize what is going on in the new cell-phone culture. It seems that the combination is fostering “real-world aware cloud computing”, where billions of cell-phones are connected to a server cloud. Thus as a result, we may retrieve a different level of real-world aware information, which has never been considered possible heretofore.

6. REFERENCES

[1] M. Aso, S. Shimada, T. Yamamoto, N. Sawai, and A. Kariya. Location information functions for FOMA terminals - location positioning function -. NTT DoCoMo Technical Journal, 7(4):13–19, 2008.

[2] J. Boyd. Here comes the wallet phone [wireless credit card]. Spectrum, IEEE, 42(11):12–14, Nov. 2005.

[3] C. Cheong, T.-D. Han, J.-Y. Kim, T.-J. Kim, K. Lee, S.-Y. Lee, A. Itoh, Y. Asada, and C. Craney. Pictorial image code: A color vision-based automatic identification interface for mobile. In Proc. 8th IEEE Workshop on Mobile Computing Systems and Applications, pages 23–28, Mar. 2007.

[4] Denso Wave Inc. QR code features. http://www.denso-wave.com/qrcode/qrfeature-e.html, 2003.

[5] M. Etoh, editor. Next Generation Mobile Systems: 3G and Beyond. Wiley & Sons., 2005.

[6] M. Etoh and T. Yoshimura. Advances in wireless video delivery. Proc. IEEE, 93(1):111–122, 2005.

[7] J. Haitsma and T. Kalker. A Highly Robust Audio Fingerprinting System With an Efficient Search Strategy. Journal of New Music Research, 32(2):211–221, 2003.

[8] T. Hamada, A. Tobe, A. Ichinose, and N. Torimoto. FOMA 905i application functions. NTT DoCoMo Technical Journal, 9(4):11–16, 2008.

[9] K. Kashino and H. Murase. Music recognition using note transition context. In Proc. ICASSP’98, volume 6, 1998.

[10] T. Kurozumi, K. Kashino, and H. Murase. A Robust Audio Searching Method for Cellular-Phone-Based Music Information Retrieval. In Proc. 16thICPR, volume 3, pages 991–994, 2002.

[11] D. G. Lowe. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 60(2):91–110, 2004.

[12] H. Matsuoka, Y. Nakashima, T. Yoshimura, and T. Kawahara. Acoustic ofdm: Embedding high bit-rate data in audio. In Proc. MMM2008, pages 497–507, Jan. 2008.

[13] Microsoft. High capacity color barcode. http://research.microsoft.com/research/hccb/, 2008.

[14] E. Ohbuchi, H. Hanaizumi, and L. Hock. Barcode readers using the camera device in mobile phones. In Proc. Int. Conf. Cyberworlds, pages 260–265, Nov. 2004.

[15] D. Pearce. Enabling new speech driven services for mobile devices: An overview of the ETSI standards activities for distributed speech recognition front-ends. In Proc. American Voice I/O Society, May 2000.

[16] T. Ramabadran, A. Sorin, M. McLaughlin, D. Chazan, D. Pearce, and R. Hoory. The ETSI extended distributed speech recognition (DSR) standards: server-side speech reconstruction. In Proc. ICASSP, volume 1, pages 1–53–6, May 2004.

[17] S. Shimada, M. Tanizaki, and K. Maruyarna. Ubiquitous spatial-information services using cell phones. Micro, IEEE, 22(6):25–34, Nov/Dec 2002.

[18] R. Typke, F. Wiering, and R. Veltkamp. A Survey of Music Information Retrieval Systems. Proc. Int. Conf. Music Information Retrieval, pages 153–160, 2005.

[19] A. Wang. An industrial-strength audio search algorithm. In Proc. ISMIR, 2003.

[20] J. Yi, M. Farzin, and J. Pedersen. Deciphering mobile search patterns: A study of yahoo! mobile search queries. In Proc. WWW2008, pages 257–266, 2008.