Multimodal Interaction in Gaming

Maria Chiara Caschera, Arianna D'Ulizia, Fernando Ferri, and Patrizia Grifoni

Institute of Research on Population and Social Policies (IRPPS) –
National Research Council (CNR)
00185, Rome (Italy)
{mc.caschera,arianna.dulizia,fernando.ferri,
patrizia.grifoni}@irpps.cnr.it

Abstract. Gaming environments are applications that have the great potential to increase people engagement in a participatory and collaborative way. Players interact with games under various situations, where the content, the form, and the modalities will be manipulated to fit the player's behaviours. This paper provides a multimodal environment for gaming by using a grammar-based approach for supporting the interaction process in the application scenario of scope card game, instantiating grammar by the elements and the rules of the game. Moreover, the paper focuses on the correct interpretation of the player's input during the game by the use of a HMM-based approach.

Keywords: Multimodal interaction, Gaming, Multimodal Language, Multimodal Grammar, Multimodal Ambiguity.

1 Introduction

Nowadays, innovations in user interface consist in making computer behaviour closer to human–human communication paradigm. In everyday life, during natural human–human communication, human beings use the five senses of touch, hearing, sight, smell, and taste in order to interact with external world. To create a natural and flexible human-computer communication paradigm, several efforts have been made to evolve traditional interfaces to multimodal interfaces. Therefore, multimodal interfaces have gained increasing importance as they allow a communication by simultaneous or alternative use of several channels of input/output at a time.

The most innovative solutions for user interfaces are usually proposed for gaming environments [1], as these applications generally have the great potential to provide new forms of engagement. The player interacts with games under various conditions where the content, the form, and the modalities will be manipulated to fit the player's behaviours. These considerations imply that multimodal interfaces play a fundamental role for the achievement of a high degree of interactivity during the gaming process.

In line with these considerations, this paper provides a multimodal platform for gaming for the achievement of a high degree of interactivity during the gaming process. In particular, the interaction process, which has been implemented in the application scenario of "scope card game", is based on a linguistic approach that uses the multimodal attribute grammar defined by the elements and the rules of the game, and the interpretation process is trained by examples of multimodal inputs of the players during the game.

Y.T. Demey and H. Panetto (Eds.): OTM 2013 Workshops, LNCS 8186, pp. 694–703, 2013.

The remainder of the paper is structured as follows. The state of the art on multimodal interaction systems for gaming is presented in Section 2. Section 3 presents the multimodal interaction gaming environment focusing both on the definition of the multimodal grammar for the game (section 3.1) and on interpretation of the multimodal player's input during the interaction process (section 3.2) in the application scenario of scope card game. Finally, conclusions are presented in Section 4.

2 Multimodal Interaction and Gaming

The application of the multimodal paradigm is increasing in computer interfaces in order to make computer behaviour closer to human communication. Indeed, communication among people is often multimodal as it is obtained combining different modalities, such as speech, gesture, facial expression, sketch, and so on. Similarly, multimodal interfaces allow several modalities of communication to be harmoniously integrated, making the system communication characteristics more and more similar to the human communication approach. The main features of multimodal interaction, such as the different classes of cooperation between different modes, the time relationships among the involved modalities and the relationships between chunks of information connected with these modalities, are described in [2].

Multimodal interaction paradigm are becoming more and more popular also in gaming platforms (e.g. Nintendo Wii) and in mobile games combining speech and graphics [3], using location (i.e., a passive/perceptual modality) as in the pervasive mobile game ARQuake [4], and including sensors such as accelerometers, proximity sensors and tactile screen with two handed interactions as in the recent iPhone [5]. A flexible multimodal interaction framework has been integrated in STARS [6] in order to provide a computer-augmented board games including simple mainstream board games such as Monopoly, and complex tabletop strategy or role-playing games. A multimodal multiplayer gaming environment has been developed in [7] where players interact by speech and gesture in a co-located home gaming complementarily using the speech for specifying abstract or discrete actions and gesture for describing location. In [8] a multimodal speech understanding game, which features incremental understanding, is provided. The work focuses on the relevant role of the incremental understanding with graphical feedback, through the use of highlighting, shading, and flashing. This work underlines the relevance of the use of grammar-based approach to provide incremental recognition results from the speech recognizer, as the user speaks.

In line to this consideration, this work provides a multimodal interaction gaming environment that uses an approach based on context free grammar in order to create an easy and intuitive interaction with the game. This approach utilizes a "by example" paradigm for defining the multimodal grammar through the use of game rules and examples of multimodal inputs in the game environments. The interpretation process is supported by a HMM-based approach that is able to disambiguate ambiguous multimodal inputs.

3 The Multimodal Interaction Gaming Environment

In order to provide a natural interaction between players in a game environment, the multimodal environment for gaming is proposed here. Figure 1 shows the client-server architecture of the gaming environment.

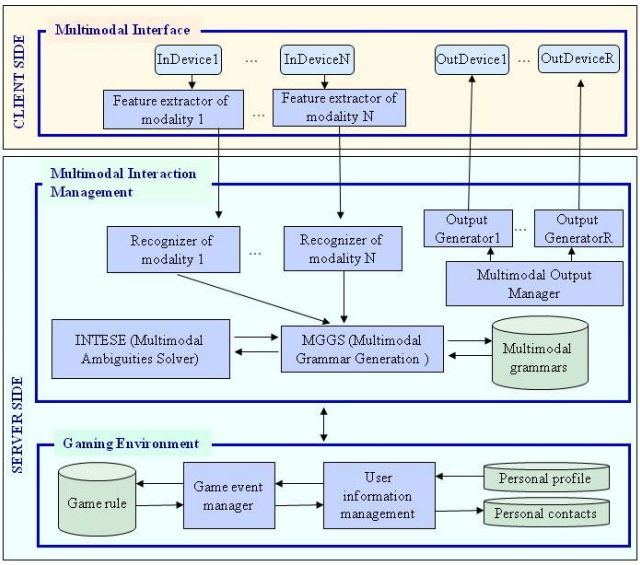


Fig. 1. Architecture of the multimodal gaming environment

The client side includes specific I/O devices (e.g. display, cameras, microphone, and loudspeakers) as well as the components for extracting features from the received signals. The server side consists of the "multimodal interaction management" and the "gaming environment". The multimodal interaction management has the role of recognizing unimodal inputs coming from the features extractors of each modality, correctly interpreting these inputs by the support of module for solving ambiguous input, integrating these different interpretations into a joint semantic interpretation, and understanding, which is the better way to react to the interpreted multimodal request by activating the most appropriate output devices. The game-based environment consists of: the game event manager, which provides the objects and rules definition for the game stored in the Game rule knowledge base, and the user information management, which is devoted to store and manage personal data of players, such as personal profile and personal contacts.

In the design of multimodal interaction environment, the process of combining information from different modalities, in order to have a comprehensive representation of the user's message, is a crucial point. In [9], a more extensive discussion about multimodal input fusion strategies has been provided. The combined information need to be interpreted by the system in order to provide an effective interaction. Therefore, the interpretation process is a further focal step. The interpretation of user input depends on different features, such as available interaction modalities, user's behaviours, conversation focus, and context of interaction. Moreover, an unambiguous interpretation of the user's input can be achieved by simultaneously considering semantic, temporal and contextual constraints. In [10], an overview of methods for interpreting multimodal input is provided. According to these considerations, this paper describes a multimodal interaction environment for gaming focusing on: the definition of the multimodal grammar for addressing the combination of information from different modalities (the MGGS-Multimodal

Grammar Generation in Figure 1) [11]; and addressing an unambiguous interpretation of the user's input by a HMMs-based approach (INTESE-Multimodal Ambiguities Solver in Figure 1) [12]. In detail, it describes the platform functionalities of the multimodal gaming environment used in order to play an Italian card game: scopa.

The following sections will describe how the platform acquires the multimodal grammar for playing "scopa card game", and how the player's input is correctly interpreted during the interaction process with the multimodal gaming environment.

3.1 The Grammar for the Multimodal Interaction Gaming Environment

In the Multimodal interaction gaming environment, a grammar for interacting with the gaming environment has been instantiated. The grammar has been generated by the multimodal attribute grammar generator system described in [11]. The grammar for the "scopa card game" has been inferred providing concrete examples of multimodal inputs of players, and the grammar inference algorithm automatically generated the grammar rules to parse the inputs. The MGGS interface allows the acquisition of the examples of the sentences and the concepts used during the interaction with the game. The MGGS takes the elements of the players' inputs and their semantic properties (i.e., actual value, syntactic role, modality, and kinds of cooperation between modalities), and integrates them opportunely, in order to generate a linear sequence of elements. In detail, the player interacts with the system in order to set for each element (see Figure 2): the actual value of the element; the modality used to express the input element; the syntactic role that the element has inside the unimodal sentence [13]; the kind of cooperation with the other input elements (e.g. redundancy, complementarity…). The linearization process combines the elements and groups them opportunely in order to generate their linear sequence.

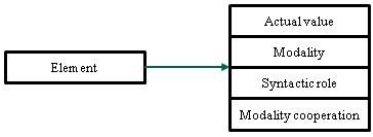


Fig. 2. Attributes of input elements

The linearization process considers: the modality cooperation for determining whether input elements convey information that has some relation with the information conveyed by the other elements; and syntactic roles for determining whether input elements can be considered close together in syntax (syntactic proximity criterion). The multimodal linearized input is sent to the MAG inference component for grammar inference [14]. By the grammatical inference algorithm, the MAG inference generates the set of production rules and the associated semantic functions that are able to parse the sentence. This process has defined the multimodal grammar, which was stored into the multimodal grammars repository (see Figure1).

In order to clarify this process, let us suppose that, in the definition process of the grammar, the player provides the multimodal input composed by the speech input "Select this card" and by the sketch modality for selecting card icon on a touch-screen display "". The sequence of input elements has associated sets of attributes as Figure 3 shows.

According to the modality cooperation (complementarity in this example), the speech elements "this" and "card" and the sketch element "" have to be close in the linearized sentence. The syntactic proximity criterion, which occurs because input elements with the same syntactic role are close together, reduces the number of acceptable linearized sentences.

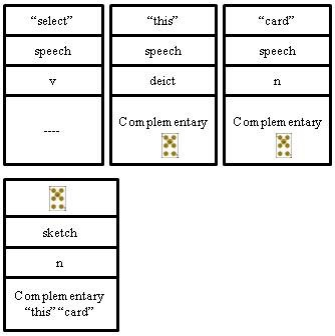


Fig. 3. Input element representation for the example

Therefore, the speech element "card" and the sketch element associated with "" are close together in the linearized sentence. The linearization process stops giving the following linear sequence of input elements: "Select" "this" "card" "".

In summary, the system enables the user to input the multimodal sentence by using devices corresponding to the inserted modalities (e.g., a microphone for speech modality, a touch screen for sketch modality). Moreover, the system converts the acquired inputs into concepts through the appropriate input recognizers (e.g., speech, sketch recognizers). Once the unimodal inputs are recognized, the language developer defines constraints both on syntactic roles and types of cooperation between modalities. After the specification of the multimodal sentence, the MGGS performs the linearization process, which translates the recognized unimodal inputs into a linear sequence of elements. This sequence is given as input to the revised CYK algorithm [14]. The running of this algorithm generates the set of production rules and the associated semantic functions.

3.2 The Interpretation in the Multimodal Interaction Gaming Environment

During the interaction process, each linearized multimodal sentence defined by the player's input can have: no interpretation, one or more than one; in the last case the multimodal sentence is ambiguous. This section describes how the player's multimodal input is unambiguously interpreted in the gaming environment on the basis of the INTESE method (the Multimodal Ambiguities Solver in Figure 1) proposed in [12]. Starting from the hypothesis that each multimodal sentence is associated to a syntax-graph, defined in [12], elements are combined to express the different interpretations by sentences in Natural Language (NL). When a multimodal sentence is ambiguous there are two or more different sentences in NL that are associated with it, i.e., the candidate interpretations of the multimodal sentence. Resolving ambiguities involves complex information, and methods. This paper uses the InteSe model developed in [12].

This model addresses the complexity of the multimodal input involving layered hidden Markov models (LHMMs) [15] structured by:

- hierarchical hidden Markov models (HHMMs) [16] that give a multilevel syntactic representation of a multimodal sentence and to resolve syntactic ambiguities using a training process [17];
- Hidden Markov models HMM [18] to assign only one meaning to each multimodal sentence, identifying the most probable sense for each element. This model is also influenced by HMMs that model context, where with context we are only considering the gaming domain.

A detailed description on how the model works is provided in [12] by showing the different steps of the multimodal dialog from a multimodal sentence in input, to the ambiguity resolution characterizing multimodal interaction.

In summary, the InteSe consists of three connected levels: the *context level*, the *syntactic sentence level* and the *semantic element level* (see Figure 4). The HHMMs syntactic sentence level enables the identification of the correct path on the syntax-graph for reaching each terminal element. Considering the semantic element level, lexical ambiguities are managed by the InteSe model assigning them the most probable sense. For dealing with issues of univocally assigning the sense tag, the InteSe model uses HMMs. The estimation of the probability is done by using information about the context, and, therefore, it is connected with the context level. The *context* and *semantic element levels* are used for the resolution of the *multimodal semantic ambiguities,* and the *syntactic sentence level* answers the need of resolving classes of *multimodal syntactic ambiguities*.

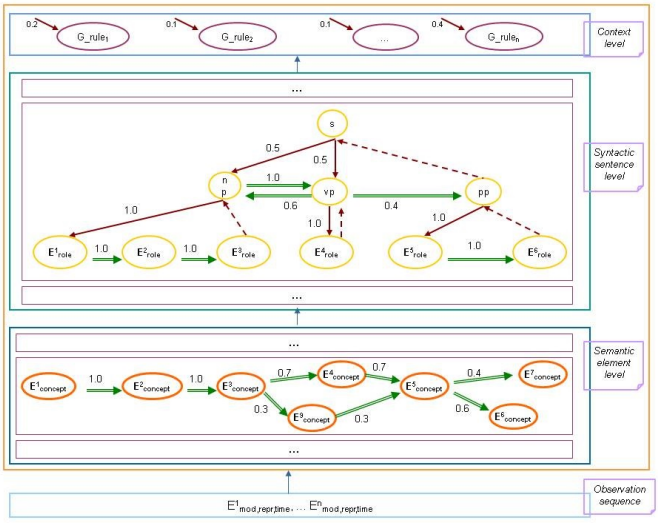


Fig. 4. Schema of the InteSe model for a multimodal sentence

In the gaming environment, it is crucial the *context game level* (see Figure 4 showing the schema of the InteSe model for a multimodal sentence), which consists of an HMM that associates the multimodal sentence to the context of the gaming environment according to the sequence of concepts that the sentence contains.

The context level supports the process of semantic element ambiguity resolution. In fact, the different meanings can be associated with the elements by considering their sequences as well as the gaming contexts in which they occur. The *context HMM* (context level in Figure 4) allows the association of each terminal element of the multimodal language with a semantic tag representing the meaning of the element in that context. The HMM states of the context level represent the rules of the games defined during the construction of the grammar for the game. Those states are linked to the game rules that are referred during the interaction process with the game.

In order to clarify the description of the used method, and example of ambiguous input is provided. In the example, the table card is shown in the Figure 5.

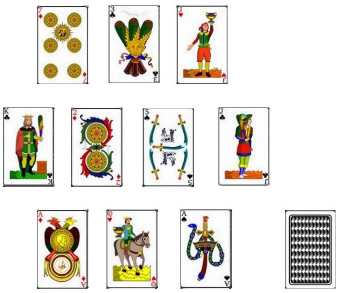


Fig. 5. Example of table cards

Let us suppose the player on the top of the Figure 4 is interacting with the game by sketch and speech.

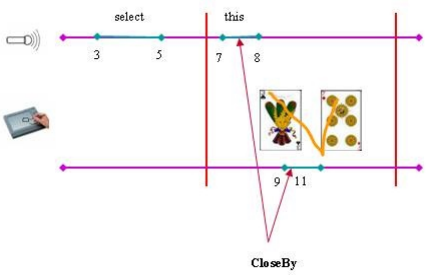


Fig. 6. Player's multimodal input that defines a target ambiguity

Using the speech modality the player says:

"select this card"

and, at the same time, the player selects by sketch two cards: the three of clubs and the seven of coins, as in Figure 6.

After, the player says:

"capture the two of coins and the five of swords"

The last sentence gives the discourse context.

The target ambiguity appears in the first player's multimodal sentence (Figure 6) because the player checks two different elements ("three of clubs" and "seven of coins") using the sketch modality. This player's sentence defines the syntax-graph in Figure 7.

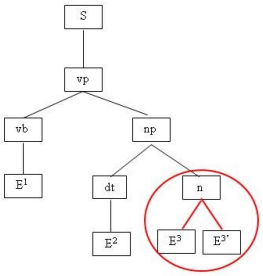
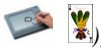


Fig. 7. Syntax-graph of the Multimodal Sentence that defines a target ambiguity

Elements defined by the speech are:

- E^1 *is!* $(E^1_{mod}=speech) \otimes ! (E^1_{repr}=$ "select" $)) \otimes ! (E^1_{time}=(3,5)) \otimes ! (E^1_{concept}=(verb)) \otimes !(E^1_{role}=(vb))$
- E^2 *is!* $(E^3_{mod}=speech) \otimes ! (E^2_{repr}=$ "this" $)) \otimes ! (E^2_{time}=(7,8)) \otimes !(E^2_{concept}=(deictic)) \otimes !(E^2_{role}=(dt))$

And using the sketch modality the player checks the following elements:

- E^3 *is !* $(E^3_{mod}=sketch) \otimes ! (E^3_{repr}=$  $)) \otimes ! (E^3_{time}=(9,11)) \otimes ! (E^3_{concept}=($ *three of clubs*$)) \otimes !(E^3_{role}=(nn))$
- $E^{3'}$ *is !* $(E^{5'}_{mod}=sketch) \otimes ! (E^{3'}_{repr}=$  $)) \otimes ! (E^{3'}_{time}=(9,11)) \otimes ! (E^{3'}_{concept}=($ *seven of coins*$)) \otimes !(E^{3'}_{role}=(nn))$

The alignment of the element E2 with the elements E3 and E3' detects a target ambiguity due to the fact that using sketch modality two different elements, "three of clubs" (E3) and "seven of coins" (E3'), are identified. Two candidate interpretations are:

1. "select this three of clubs";
2. "select this seven of coins";

For dealing with this target ambiguity, the HMM of the semantic element level is used. Figure 8 shows the hidden states, the observation sequences and the values of the probabilities of the model for dealing the provided example. The probabilities are computed starting from the information contained by examples of multimodal sentences defined by the players, the rules of the scopa game and the discourse context during the interaction (e.g. the speech sentence "capture the two of coins and the five of swords" in the example). The parameters of the model were estimated by a generalization of the Baum–Welch algorithm [19].

The trained model is able to correctly interpret multimodal sentences by identifying the most probable sequence of the LHMM states by the use of Viterbi algorithm [20].

In detail, the model returns the sequence of syntactic roles and concepts contained in the multimodal sentence having the highest probability value, according to the trained model.

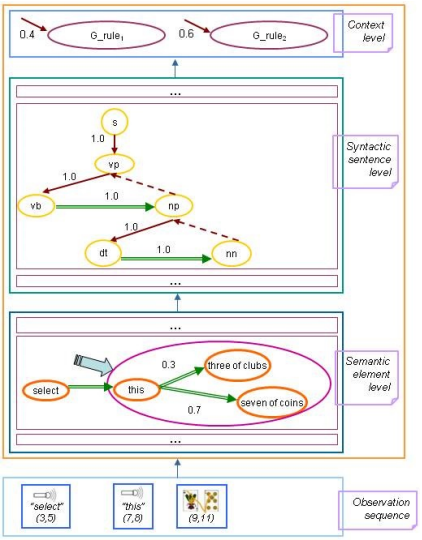


Fig. 8. Trained model for solving the target ambiguity

In this example, according to the model in Figure 8, the system interprets the multimodal player's input as follows:

"Select" "this" "card" "[image]"

4 Conclusion

This paper presented a multimodal interaction gaming environment to provide pervasive game applications on an interactive table and additional devices, such as camera and microphone. The work particularly focused on the grammar-based approach for supporting the interaction process. The definition of the multimodal grammar for the specific game, in the example scopa card game, and the method for correctly interpreting the multimodal user's input during gaming, were provided.

For the scopa card game, the multimodal grammar was inferred by concrete examples of multimodal inputs of players, and the grammar inference algorithm has automatically generated the grammar rules that were used to parse the inputs. The correct interpretation of the player's input was achieved by the use of the HMM-based approach that is able to incrementally learn the peculiar interaction features of each player, and in the specific gaming environment is crucial the training of the context layer of the model by the rules of the game.

In future works, a large-scale test of the environment will be developed for a set of meaningful games, in order to evaluate the level of user satisfaction in the use of multimodal interaction type.

References

1. Crawford, C.: Lessons from Computer Game Design. In: Laurel, B. (ed.) The Art of Human-Computer Interface Design, pp. 103–111. Addison-Wesley Pub. Co., Reading (1990)

2. Caschera, M.C., Ferri, F., Grifoni, P.: Multimodal interaction systems: information and time features. International Journal of Web and Grid Services (IJWGS) 3(1), 82–99 (2007)
3. Zyda, M., et al.: Educating the Next Generation of Mobile Game Developers. IEEE Computer Graphics and Applications 27(2), 96, 92–95 (2007)
4. Piekarski, W., Thomas, B.: ARQuake: The Outdoor Augmented Reality Gaming System. Communications of the ACM 45(1), 36–38 (2002)
5. iPhone, http://www.apple.com/iphone
6. Magerkurth, C., Stenzel, R., Streitz, N., Neuhold, E.: 2003. A multimodal interaction framework for pervasive game applications. Artificial Intelligence in Mobile System (Aims), Fraunhofer Ipsi (2003)
7. Tse, E., Greenberg, S., Shen, C., Forlines, C.: Multimodal multiplayer tabletop gaming. ACM Comput. Entertaint. 5(2), Article 12, 12 pages (2007), http://doi.acm.org/10.1145/1279540.1279552, doi:10.1145/1279540.1279552
8. Gruenstein, A.: Shape game: A multimodal game featuring incremental understanding. Term project, Massachusetts Institute of Technology (May 2007)
9. D'Ulizia, A.: Exploring Multimodal Input Fusion Strategies. In: The Handbook of Research on Multimodal Human Computer Interaction and Pervasive Services: Evolutionary Techniques for Improving Accessibility, pp. 34–57. IGI Publishing (2009)
10. Caschera, M.C.: Interpretation methods and ambiguity management in multimodal systems. In: Grifoni, P. (ed.) Handbook of Research on Multimodal Human Computer Interaction and Pervasive Services: Evolutionary Techniques for Improving Accessibility, pp. 87–102. IGI Global, USA (2009)
11. D'Ulizia, A., Ferri, F., Grifoni, P.: Generating Multimodal Grammars for Multimodal Dialogue Processing. IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans 40(6), 1130–1145 (2010)
12. Caschera, M.C., Ferri, F., Grifoni, P.: InteSe: An Integrated Model for Resolving Ambiguities in Multimodal Sentences. IEEE Transactions on Systems, Man, and Cybernetics: Systems 43(4), 911–931 (2013)
13. Mitchell, P.M., Santorini, B., Marcinkiewicz, M.A.: Building a large annotated corpus of English: The penn treebank. Comput. Linguistics 19(2), 313–330 (1994)
14. D'Ulizia, A., Ferri, F., Grifoni, P.: A Learning Algorithm for Multimodal Grammar Inference. EEE Transactions on Systems, Man, and Cybernetics, Part B 41(6), 1495–1510 (2011)
15. Oliver, N., Garg, A., Horvitz, E.: Layered representations for learning and inferring office activity from multiple sensory channels. Computer Vision and Image Understanding 96(2), 163–180 (2004)
16. Fine, S., Singer, Y., Tishby, N.: The hierarchical hidden Markov model: Analysis and applications. Machine Learning 32(1), 41–62 (1998)
17. Skounakis, M., Craven, M., Ray, S.: Hierarchical hidden Markov models for information extraction. In: Proc. of the 18th Int. Joint Conference on Artificial Intelligence, Acapulco, Mexico, pp. 427–433. Morgan Kaufmann, San Francisco (2003)
18. Rabiner, L.R.: A tutorial on hidden Markov models and selected applications in speech recognition. Proc. of the IEEE 77, 257–285 (1989)
19. Fine, S., Singer, Y., Tishby, N.: The hierarchical hidden Markov model: Analysis and applications. Machine Learning 32(1), 41–62 (1998)
20. He, J., Hu, S., Tan, J.: Layered hidden Markov models for real-time daily activity monitoring using body sensor networks. In: Int. Workshop Wearable and Implantable Body Sensor Network, Hong Kong, China (2008)