INFORMATION PROCESSING AND SEMANTIC CORRELATION
IN MULTIMODAL, MULTILINGUAL, HUMAN COMPUTER INTERFACES

A Dissertation in
Computer Science and Engineering
by
Namsoo Jung

© 2012 Namsoo Jung

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

December 2012
The dissertation of Namsoon Jung has been reviewed and approved* by the following:

Raj Acharya  
Professor of Computer Science and Engineering  
Head of the Department of Computer Science and Engineering  
Chair of Committee, Dissertation Co-Advisor

Padma Raghavan  
Professor of Computer Science and Engineering

Guoray Cai  
Associate Professor of Information Sciences and Technology  
Affiliate Faculty, Department of Computer Science and Engineering

Alan M. MacEachren  
Professor of Geography

Rajeev Sharma  
Associate Professor of Computer Science and Engineering  
Special Member, Dissertation Co-Advisor

*Signatures are on file in the Graduate School.
Abstract

One long-term goal in human computer interaction (HCI) study has been to migrate the “natural” means that humans employ to communicate with each other into the HCI system. Any HCI system that aspires to have the same naturalness as that of an interaction between humans should be multimodal, thus requiring a need for developing a multimodal HCI system.

Multilingual collaboration through a multimodal human computer interaction (HCI) interface needs to handle differences among the multiple human languages used by the collaborators. The difficulties of handling the multilingual information are partially caused by the linguistic differences in grammar, vocabulary, idiom, expression, word-formation, and the existence of unique homonym, polysemy, and homophone in each human language. Due to the difficulties, a semantic extraction and correlation process among the multilingual information becomes a further difficult task in a multimodal HCI system.

In this dissertation, a deterministic information processing framework for multilingual collaboration in the context of geo-spatial collaborative crisis management is proposed. The deterministic information processing framework comprises an information unit augmentation step, called object verification, and operating mechanism for the augmented information unit, called polymorphic operator, as a part of the proposed approach to process the multilingual information. In the subject of the multilingualism, the deterministic information modeling approach is an attempt to reduce the complexity in handling the multilingual information by replacing the language modeling process with the information modeling approach.

In this dissertation, the term “rule“ is defined as an agreement or guideline based on the intention for achieving a goal among the users of a particular embodiment of the approach in a specifically predefined domain for defining a set of relationships among the verified information units that are used in the specifically predefined domain.

It is an objective of the proposed approach to see whether the deterministic information processing approach can efficiently handle the multilingual collaboration and semantic correlation with feasible computational complexity and processing time through algorithm analysis and discussion of exemplary experimentation. The proposed approach is concerned with the reduced corpus of the multilingual information in the specified domain. The proposed
approach is also restricted to the multilingual collaboration and understanding in the medium of a multimodal HCI system.
# Table of Contents

List of Figures ix

List of Tables xi

Acknowledgments xii

1 Introduction .................................................................................................................. 1
   1.1 Overview .................................................................................................................. 1
   1.2 Problem Description ............................................................................................... 4
      1.2.1 Problem Statement .......................................................................................... 4
      1.2.2 Objective ......................................................................................................... 7
   1.3 Scope of the Research and Restrictions ................................................................. 7
   1.4 Overview of the Proposed Approach .................................................................... 9
   1.5 Contributions ......................................................................................................... 10
   1.6 Dissertation Outline ............................................................................................. 11

2 Background .................................................................................................................. 13
   2.1 Relevant Work ........................................................................................................ 13
      2.1.1 Multimodal Human Computer Interaction and Information Processing .... 13
      2.1.2 Computer Supported Collaborative Work (CSCW) ...................................... 16
      2.1.3 Object-Orientation in Research ....................................................................... 17
      2.1.4 XML and Web Services ................................................................................ 19
      2.1.5 Multilingualism in a Multimodal HCI System .............................................. 19
   2.2 Scientific Challenges ............................................................................................. 21
   2.3 Objectives and Significance of The Research ..................................................... 22

3 Multilingual Multi-user Multimodal Human Computer Interaction ....................... 24
   3.1 Overview ................................................................................................................ 24
   3.2 Multimodal Human Computer Interaction .......................................................... 25
   3.3 CSCW and Geospatial Collaboration ..................................................................... 27
   3.4 Multilingualism and Computational Linguistics .................................................. 27
   3.5 Proposed Approach ............................................................................................... 30
      3.5.1 MultiLUM ....................................................................................................... 30
      3.5.2 Key Features and Methodology ...................................................................... 31
      3.5.3 Scientific Hypothesis .................................................................................... 32
   3.6 Discussion .............................................................................................................. 33
6.7 Semantic Correlation Module (SCM) ................................................................. 124
6.8 Discussion ........................................................................................................... 126

7 Object Verification and Components in Interface Design ........................................ 128
7.1 Overview ............................................................................................................ 128
7.2 Database Management ........................................................................................ 128
  7.2.1 Introduction .................................................................................................. 128
  7.2.2 Relevant Work ............................................................................................. 130
  7.2.3 VeRification EnAbled Database (VREAD) .................................................... 133
7.3 Output Presentation ............................................................................................. 140
  7.3.1 Introduction .................................................................................................. 140
  7.3.2 Relevant Work ............................................................................................. 141
  7.3.3 Verified Object and Multimedia Output (VOMO) ........................................ 143
  7.3.4 Output of the Multilingual Information ....................................................... 145
7.4 Web-based Collaboration and Mobile Extension .................................................. 146
  7.4.1 Introduction .................................................................................................. 146
  7.4.2 Relevant Work ............................................................................................. 147
  7.4.3 Web Services and Collaboration .................................................................. 148
  7.4.4 Extension to Mobile Devices ...................................................................... 148
7.5 Discussion ........................................................................................................... 149

8 Experimentation ..................................................................................................... 151
8.1 Overview ............................................................................................................ 151
8.2 Evaluation Methodology ...................................................................................... 151
8.3 Exemplary Experimentation ................................................................................ 153
8.4 Implementation of the LISN .............................................................................. 161
  8.4.1 Technical Aspects and Configurations of the LISN Application .................. 164
  8.4.2 Contribution of the Implementation ............................................................. 165
8.5 Analysis and Evaluation ..................................................................................... 166
  8.5.1 Simulation .................................................................................................... 166
  8.5.2 Algorithm Analysis ...................................................................................... 172

9 Conclusion ............................................................................................................ 183
9.1 Summary ............................................................................................................ 183
  9.1.1 Suggested Approach .................................................................................... 183
  9.1.2 Evaluation of the Approach ........................................................................ 185
9.2 Future Work ....................................................................................................... 187
  9.2.1 User Collaboration Handling Based on the Information Processing Scheme .................................................................................. 187
  9.2.2 Usability and Human Factor Study ............................................................... 187
  9.2.3 Measurement and Evaluation for Other Aspects of the Approach ............... 188
9.3 Related Contribution ......................................................................................... 188
Bibliography

Appendix
List of Figures

3.1 Multilingual multi-user multimodal HCI system (MultiLUM) .................................. 30
4.1 Overview of the OVEN .................................................................................................. 35
4.2 Object verification process .......................................................................................... 45
4.3 An example of explicit verified objects ......................................................................... 46
4.4 Single object verification for an object ......................................................................... 47
4.5 Multiple object verifications for an object ..................................................................... 49
4.6 Relationship between verified objects and multiple objects ....................................... 51
4.7 Example of a rule based on the Boolean logic in a segment of a XML file .................. 55
4.8 Pseudo example of the polymorphic operator using the Boolean logic ....................... 55
4.9 Application of the rule to verified objects by the polymorphic operator at the semantic group level .................................................................................................................. 59
5.1 An exemplary state diagram for the interactions with application contents in a multimodal HCI system .................................................................................................................. 67
5.2 An exemplary multimodal HCI system ......................................................................... 72
5.3 An application of object verification enabled network (OVEN) in the context based reasoning module for a multi-user multimodal HCI system ................................................. 81
5.4 An exemplary processing sequence and conversational segment of a particular interaction ............................................................................................................................................... 85
5.5 Exemplary rules for processing a command using spatial/gesture contextual information and conversation contextual information ................................................................. 87
5.6 An exemplary processing sequence and conversational segment in which the vision handler is notified by the CBRM in the feature recognition level for resolution adaptation event .................................................................................................................. 89
5.7 Exemplary rules for sending messages to handlers based on spatial, gesture-based, symbolic, and conversation contextual information to process a command ............... 90
5.8 Block-by-block processing and Semi-curvature treatment in the viscosity approach (application of the sticky agent) ........................................................................................................... 97
5.9 Exemplary rules for helping a filtering algorithm to filter out a reverse trajectory with stronger confidence based on contextual information ....................................................... 99
5.10 Guided movement by the CBRM .................................................................................. 100
5.11 Architecture of the context based reasoning module and a set of handlers .................. 103
6.1 An exemplary configuration of the LISN application ................................................. 110
6.2 Components of the language independent speech network (LISN) ............................... 118
6.3 An integration of the language independent speech network (LISN) into a multimodal HCI geo-spatial collaborative crisis management application ........................................ 119
6.4 Steps of multilingual speech understanding process in the LISN ............................... 124
6.5 Mapping of semantics for a multilingual system and object representation language as a medium for the semantic mapping in the LISN ......................................................... 126
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.1</td>
<td>Architecture of a system that utilizes the VeRification EnAbled Database (VREAD), a VREAD management system (VMS), and an object verification controlling module (OVCM)</td>
<td>134</td>
</tr>
<tr>
<td>7.2</td>
<td>Network of the object verification enabled databases (VREADs) and the VREAD management system (VMS)</td>
<td>139</td>
</tr>
<tr>
<td>7.3</td>
<td>Model for the verified object and multimedia output (VOMO)</td>
<td>144</td>
</tr>
<tr>
<td>7.4</td>
<td>Screen shot of an exemplary output for the verified objects in the VOMO</td>
<td>146</td>
</tr>
<tr>
<td>8.1</td>
<td>Languages and information conveyance tools used in the language independent speech network (LISN) for an exemplary scenario</td>
<td>154</td>
</tr>
<tr>
<td>8.2</td>
<td>Exemplary grammar for “get” command in English</td>
<td>159</td>
</tr>
<tr>
<td>8.3</td>
<td>Exemplary grammar for image type in English especially for a web application</td>
<td>159</td>
</tr>
<tr>
<td>8.4</td>
<td>Exemplary Grammar for “get” command in Japanese</td>
<td>160</td>
</tr>
<tr>
<td>8.5</td>
<td>Various configurations of the LISN applications</td>
<td>163</td>
</tr>
<tr>
<td>8.6</td>
<td>Diagram of processes in a configuration of the LISN application</td>
<td>165</td>
</tr>
<tr>
<td>8.7</td>
<td>An exemplary state diagram for the information correlation of a homonym ‘bank’ in the LISN</td>
<td>169</td>
</tr>
<tr>
<td>8.8</td>
<td>Exemplary rules for state transition conditions in a state diagram for the information correlation of a homonym ‘bank’</td>
<td>170</td>
</tr>
<tr>
<td>8.9</td>
<td>Exemplary semantic data structure using hash tables</td>
<td>176</td>
</tr>
<tr>
<td>8.10</td>
<td>Union of disjoint sets in a semantic data structure implemented in a hash table</td>
<td>181</td>
</tr>
</tbody>
</table>
List of Tables

5.1 Exemplary classes of keywords for speech inputs in the context based reasoning module ................................................................. 87

7.1 Relation by the object verification criteria for the verified objects utilizing the relational database model ................................................................. 133
7.2 Exemplary “many to 1” and “1 to many” relationship in VREAD ............................ 137

8.1 An exemplary interaction segment from the scenario ........................................ 155
8.2 Another exemplary interaction segment from the scenario ................................ 156
Acknowledgements

This work was partially supported by the VideoMining Corporation, State College, PA, and the NSF Geospatial Collaborative Crisis Management (GCCM) Project (National Science Foundation under Grant No. EIA-0306845.)
Chapter 1

Introduction

1.1 Overview

Sharma, et al. [118], Krahnstoever, et al. [69], and Kettebekov, et al. [63] noted that one long-term goal in human computer interaction (HCI) has been to migrate the “natural” means that humans employ to communicate with each other into the human computer interaction system. Any HCI system that aspires to have the same naturalness as that of an interaction between humans should be multimodal, thus requiring a need for developing a multimodal HCI system. The “XISM” system developed by Sharma, et al. [118] or DAVE_G by Rauschert [107] is a good exemplary multimodal HCI system.

As it is well-known, the natural human to human interaction is achieved through activities for issuing and receiving information utilizing five basic senses, such as audio information by voice conversation, visual information in gestures by hand and body movements, tactile information by touch, and so on. In the case of computer aided devices and human to computer interaction, the activities for issuing and receiving information can be achieved by various sensors. In recent years, keyboard and mouse have been used as popular sensors to capture the tactile movement. However, over the years, the demand for more natural sensors that are closer to the actual human to human interaction has been increased. For example, the natural sensors like hearing and seeing can be better replaced by using a microphone for voice conversation and a camera for visual image sensing rather than a graphical icon or menu selection on a graphical user interface (GUI) using a mouse or keyboard.

According to Sharma, et al. [118] and Liu, J., et al. [80], the integration of multimodal input for HCI can be seen from the perspective of multisensor data fusion. Different sensors can be related to different communication modalities although a multimodal HCI system can take advantage of multiple computer-sensing modalities in order to mimic the natural interaction.

Sharma, et al. [118] also pointed out that studies have shown that people prefer to interact with computers using the multimodality. It is known that the multimodal HCI eases the need for
specialized training, and fusion of multiple types of sensors may increase the accuracy by reducing the uncertainty in decision making.

Especially, a multilingual multi-user multimodal HCI system can be used as an efficient medium for a specialized computer supported communication. A possible contribution of such system includes a capability of allowing the multiple collaborators, with different native spoken language, to collaborate with each other speaking in their own native language or most comfortable language given the choice for alternative languages.

Information flow bottleneck is one of the well-known problems in the multi-modal HCI community, as described in Sharma, et al. [118]. In a relevant work, MacEachren, et al. [89, 90] noted that crisis management is an important application domain that needs to handle a large data sources. Although the information flow bottleneck is caused primarily by the large amount of data to handle or by the inefficiency of the interface, other human or cultural factors can cause the bottleneck also. Such factors include a communication delay due to a need for the translation among multiple human languages, a need to handle multilingual data, or users’ deteriorating comfort level that can occur by the usage of the second language.

The examples of various multilingual collaboration areas where the information flow bottleneck can be found comprise:

- International humanitarian relief efforts for natural disasters, such as the 2004 Indian Ocean earthquake and tsunami, December 26, 2004, or 2005 hurricane Katrina, August, 2005,
- Collaborative tasks in an international organization, such as the Red Cross or the U.N,
- Geo-spatial collaboration in emergency operation centers distributed internationally,
- Handling of a huge amount of multilingual intelligence data against a foreign terrorist attack, and
- Scientific collaboration in an international group of people.

One might expect that they can share a specific common language for the collaboration. However, assuming each of the collaborators in the multilingual collaborative group has the same level of language capability or education level for the specific common language can be fundamentally flawed.
If the collaborators come from diverse backgrounds, each collaborator’s proximity to the specific language will be different. It is likely that their comfort level for the specific language can influence the performance result in actually performing a task. Usability studies often show that failures in speech recognition prompt users to choose other input modalities, such as gesture or mouse, while interacting with a multimodal HCI system. Speech recognition for the words that are not pronounced by the native speakers is certainly more challenging.

Ellis, et al. [39] decomposed a collaboration scenario into a 2x2 matrix based on the time and space attributes. Similarly, we can consider different cases of collaboration, such as:

- a synchronous face to face collaboration among the collocated collaborators in a physical space,
- a synchronous communication device mediated collaboration in a distributed environment,
- an asynchronous computer mediated collaboration in the same location, or
- an asynchronous computer mediated collaboration in different locations.

Depending on the case of the collaboration scenario, the dependence of the collaborators to the computer-based system may vary. For example, in the case of a synchronous face to face collaboration among the collocated collaborators in a physical space, the collaborators may prefer to communicate with each other directly rather than relying on any device. However, if the nature of the task requires a computer mediated information processing, the collaborators’ dependence to the computer-based system will certainly increase.

Therefore, in any combination of the synchronous/asynchronous computer mediated collaboration in the same/different location, if the information required to process is in multilingual form, the need for an intelligent multilingual HCI system will certainly increase. Especially, an intelligent multilingual multimodal HCI system will help the collaborators by fully utilizing the strength of the multimodality in the multimodal HCI system.

In the various multilingual collaboration areas, an efficient strategy for computer supported communication among the collaborators will facilitate their collaboration and maximize the result of the collaboration, in handling the complex multilingual data and information, while the
scientists, workers, emergency managers, field agents, or representatives from different language backgrounds try to collaborate.

**1.2 Problem Description**

Although it is evident that there is a need for the multilingual multi-user multimodal HCI system to exist as discussed above, there are also multiple challenges that need to be overcome in order to actually develop such a system. The essential research topics for the challenges are discussed from three different perspectives in this section:

- from the multimodal HCI system perspective,
- from the computer supported collaborative work (CSCW) perspective, and
- from the multilingual speech processing perspective.

In general, the research topics in a multimodal HCI system are related to novel input multimodalities, interaction modeling with regard to the novel input methods, improved efficiency, improved robustness, and usability. The studies in these areas help the user interaction from various points. For example, a novel input method enables a broader accessibility and usability by allowing an alternative tool for the users.

However, it is a non-trivial task to come up with an implementation of a multilingual multimodal HCI system that is practically usable. As a starting point, availability for speech recognition engines for desired human languages is limited at the moment. Apart from the multilingualism, even the monolingual speech recognition technology itself has many known research topics that need to be overcome. While improvements are expected to be made in the relevant research community, it appears that many challenges in the technology may not be resolved in the foreseeable near future.

Given the situations, this dissertation carefully discusses applicability of a new algorithmic approach to the modeling of a multilingual multimodal HCI system. We are interested in learning how the proposed heuristic information processing approach helps in handling the complex multilingual information in a multimodal HCI system. Therefore, our research focuses
on an approach for handling complex multilingual information in a multimodal HCI system rather than improving any individual feature in a speech recognition system, e.g. Chen, et al. [28], such as the speech recognition rate for any particular human language.

The input information from the speech recognition modality in a typical multimodal HCI system should be seamlessly coordinated with other input information from other types of input modalities, such as gesture recognition. Therefore, one of the topics that this dissertation discusses is how the multimodal HCI system can handle the multilingual speech input seamlessly with other types of input in the system, utilizing the proposed information processing approach.

Research topics in the CSCW area relate to how a computer-based system can help in a collaborative work environment. Although it is a possible scenario that a single user uses a multilingual multimodal HCI system, one of the areas of interest in our research is concerned with the collaborative interaction by multiple users that use multiple human languages.

Therefore, another of our study topics is to discuss how the proposed information processing approach can help in the computer supported multilingual collaboration. We discuss the functional roles of the proposed information processing approach in the computer supported multilingual collaboration, using a language independent speech network as an example of the computer-based system in a collaborative multilingual work environment.

With regard to the multilingual speech processing perspective, the areas of research topics in the multilingual speech processing can include the following topics, according to Schultz, et al. [115]:

- Language characteristics in multiple human languages and consequences for multilingual speech processing
- Development of multilingual data repositories and resources
- Multilingual acoustic modeling in automatic speech recognition
- Multilingual language modeling, including portability of existing language modeling techniques, morphological complexity, word segmentation, and cross-linguistic language modeling
- Multilingual text-to-speech generation and synthesis
- Automatic language identification
- Problems in non-native speech in the context of multilingualism
• Different approaches and strategies in speech-to-speech translation
• Issues in multilingual automated dialog system

In this dissertation, the multilingualism is partially related to the language characteristics in multiple human languages, development of multilingual data repositories and resources, multilingual language modeling, automatic language identification, and issues in a multilingual automated dialog system, among the topics above.

The difficulties of handling the multilingual information are partially caused by the linguistic differences among the human languages, such as the differences in grammar, vocabulary, idiom, expression, and word-formation. Unique homonym, polysemy, and homophone in each human language make the handling of the multilingual information harder. Due to these difficulties caused by the linguistic differences, a semantic extraction and correlation process among the multilingual information becomes a further difficult task in a multilingual multimodal HCI system. These create a research need for an efficient approach for the multilingual information processing and its semantic correlation with regard to a multilingual multimodal HCI system.

1.2.1 Problem Statement

The problem statement of this research is summarized as follows:
“A multimodal HCI interface that enables multilingual computer supported collaborative work (CSCW) needs a new information processing approach to solve issues and challenges from the multimodal HCI system perspective, including a seamless fusion of multiple modality information, from the CSCW perspective, including a unified collaboration among multilingual users, and from the multilingual speech processing perspective, including a precise semantic correlation in the multilingualism.”
1.2.2 Objective

In summary of the above discussion, if a design of a multilingual multimodal HCI system is contemplated in order to facilitate a computer supported multilingual collaboration, it is important that the design approach has a capability of handling semantic correlation among different linguistic features in multiple human languages while enabling a seamless integration of the multilingual speech input modality with other modalities in the multimodal HCI system.

From the HCI interface design perspective, a shift of focus from individual human language specific features to a common information processing framework that the multiple human languages share can reduce the necessary amount of data and complexity that need to be processed in order to facilitate the multilingual collaboration.

Therefore, the main objective is summarized as follows:

A deterministic information processing framework is proposed as a solution for information handling in a multilingual multimodal HCI system. In the proposed approach, it is expected that the framework provides a novel and efficient mechanism for both 1) a semantic correlation among multiple human languages and 2) a seamless integration of the multilingual speech input modality with other modalities in the multimodal HCI system, wherein the interface design for the multilingual computer supported collaborative work is approached from the perspective of an information processing framework rather than from the perspective of linguistic features in a group of languages.

1.3 Scope of the Research and Restrictions

Automatic speech recognition technology has been improved over many decades although the technology is not mature enough to support all the desired human languages in the multilingual application yet. The previous research in the speech recognition technology generally relates to improving the performance in monolingual speech recognition technology for a specific individual human language or developing a multilingual speech recognition engine, e.g. Wang, X., et al. [144].
The research in this dissertation is not related to the study of improving the speech recognition technology for any individual human language or developing a multilingual speech recognition engine. The research specifically focuses on studying and understanding the issues in multilingual collaboration with the medium of a multilingual multimodal HCI system in a specified domain, from a HCI system design perspective.

Therefore, the research assumes that the recognition performance of any monolingual speech recognition does not affect the multilingual aspects of the study. In many of the discussions, the speech recognition for any individual language is assumed to be already performed with a normal success rate regularly found in a real-world execution at the feature level.

In addition, the research is not concerned with any language specific features as they are discussed in linguistics. For example, although the experimentation is performed based on English and Japanese as the exemplary language, the research is not particularly interested in the language specific features in these languages. Although each language in the discussion has its own language specific features and an approach is suggested to handle semantic correlation over the unique features, the research is mostly oriented toward the issues of language independent operations from information modeling perspective, such as semantic correlation for commonly understood tasks, among the multiple languages. Therefore, the research is not particularly interested in how to solve the language specific features for certain languages from the perspective of linguistics.

The study direction in this dissertation is not intended for the specific topics in the natural language processing research area, a multilingual speech recognition engine, nor a multilingual translation system, e.g. Becker, et al. [15], either.

In the experimentation and in the discussion for the feasibility of the proposed approach, the sample multilingual data will be restricted to the reduced corpus of the multilingual information in the geo-spatial collaborative crisis management domain. The sample data will be discussed within a well defined geospatial context with the restricted vocabulary.
1.4 Overview of the Proposed Approach

The research proposes a set of novel approaches that enable the implementation of a multilingual multimodal HCI system as a part of the study. The importance of the multilingual multimodal HCI system, the requirements, and the difficulties in implementing and utilizing such a system in a specific domain are discussed. Due to the complexity of processing multilingual information among possibly many multiple human users in an unconstrained and geographically distributed environment, the research aims for an efficient and robust set of algorithmic approaches.

A set of algorithmic approaches, called language independent speech network (LISN), are introduced as a multilingual multi-user multimodal HCI system. The LISN utilizes a novel information processing approach in order to embody a real-world applicable multilingual multimodal HCI system.

A deterministic information processing approach, called object verification enabled network (OVEN) is introduced as one of the key approaches in the set. The OVEN provides a framework that handles the information processing to correlate semantics among the multiple languages through an information augmentation process, called object verification. In this approach, the term “rule“ is defined as an agreement or guideline based on the intention of achieving a goal among the users of a particular embodiment of the approach in the specific domain, which establishes a set of relationships among the verified information units that are used in the specifically predefined domain. The OVEN is based on deterministic rules, but different from the rule-based modeling in relation to the traditional language modeling, such as the BNF notation. For example, the OVEN does not use the production rules and notations in the context-free grammars in Scott [116].

In addition, the proposed approaches in this dissertation also briefly introduce novel approaches for database management, output production, and mobile extension.

Traditionally, a multimodal HCI system used to have a dialog management system as a sub-routine that handles the language understanding and communication between the user and the system independently, while the multimodal HCI system handles other parts of the system execution, such as feature level speech recognition and gesture recognition among the input modalities. The “Dave_G” system by Rauschert [107] is an exemplary multimodal HCI system.
that uses a dialog management system. The proposed approach can handle both the language understanding and the feature level system operation in a single framework. This makes the execution of the multilingual multimodal HCI system seamless and simpler than the traditional way from the point that the traditional multimodal HCI systems execute the language understanding and the feature level system operation separately in two different components.

More importantly, in the shifted focus mentioned above, the job of the interface design becomes much easier by focusing on the common framework rather than each individual language’s specific features. This novel strength can be extended to any situation where the multimodal HCI interface should handle other types of data than human languages also.

1.5 Contributions

The research effort in this dissertation can contribute in the circumstances of studying and implementing computer-mediated tools for the various multilingual collaboration areas mentioned above. The research will help us to further broaden our understanding and knowledge about the issues in multilingual computer supported collaborative work. What we can learn through this research also includes:

- Issues, challenges, and difficulties in enabling multilingual computer supported collaborative work (CSCW) by studying relevant areas based on the literatures.
- How a heuristic information processing approach, proposed in this dissertation, can help in the tasks of complex multilingual information processing for the multilingual CSCW based on an experimentation and an evaluation of the approach.

It is also hoped that the dynamic nature of the proposed approach can help in the information flow bottleneck situation in the areas discussed above, so that any unnecessary delay by the differences in human languages or by the limitations in the individual’s linguistic ability can be alleviated in certain cases. For example, it is possible that the humanitarian relief efforts for natural disasters, the collaborative tasks in an international organization, or any geospatial
collaboration in emergency operation centers, could be facilitated by the utilization of what we learned and discussed in this dissertation.

In addition, it is hoped that the research also contributes in other areas, where the benefits of the multilingual collaboration are greatly needed. For example, the features in the research could be used to develop a whole new kind of computer mediated interface that can handle a huge amount of multilingual intelligence data, or help in an international scientific collaboration.

As an information conveyance tool, the proposed approach reduces the complexity of processing time for mapping the key words from a group of keywords to a group of semantic units from $O(n^2)$ to $O(n)$, where $n_i$ is total number of keywords in a language $l_i$. A semantic data structure is discussed using multiple hash tables in two layers, and a successful search time to find a keyword and its object verification in the exemplary data structure takes $\Theta(1 + \alpha)$, where $\alpha$ is the load factor. In the exemplary semantic data structure, depending on how the polymorphic operator (PO) is designed, the execution of the PO may require two disjoint data structures to be merged together, and an amortized time of the operations is $\Theta(m)$, where $m$ is the total number of operations for making new sets, union, and finding a set. Thus, the operations in the proposed approach are found to spend linear execution time.

1.6 Dissertation Outline

Chapter 2 discusses the background of the overall research by analyzing relevant work and the scientific challenges from the relevant work, which eventually lead to the objectives and significance of the research.

Chapter 3 discusses the key points in the relevant sub-research areas that build up the multilingual, multi-user, multimodal human computer interaction and proposes an approach for the multilingual multi-user multimodal human computer interaction system, called MultiLUM.

Chapter 4 discusses one of the key features in the proposed MultiLUM, called object verification enabled network (OVEN), as a novel information processing approach.

Chapter 5 discusses some issues in the information processing areas in the relationship with the object verification enabled network (OVEN) to provide better understanding of the novel
approach. Some exemplary areas where the verification enabled network (OVEN) can be best applied, such as context based reasoning processes in a human computer interface, are discussed.

Chapter 6 discusses a proposed approach that can be used as a basis for developing a multilingual multimodal human computer interaction system. The basis can be both algorithmic and architectural. The proposed approach is called language independent speech network (LISN). Chapter 6 also discusses the detailed features of the real-time speech recognition processing feature in the LISN and scientific analysis behind the features, including language switching and semantic correlation.

Chapter 7 discusses the detailed features of the speech database processing and management in the LISN and an analysis behind the features, including the VeRification EnAbled Database (VREAD) and topics in the information retrieval. Chapter 7 also discusses visualization of the multilingual information in multimodal HCI in relation to the proposed approach. An exemplary approach is given as verified object and multimedia output (VOMO). Chapter 7 further discusses web-based collaboration and mobile extension in relation to the proposed approach.

Chapter 8 discusses an experimentation that attempts to prove a feasibility of the concepts in the proposed approach utilizing an implementation for the LISN. An evaluation methodology is designed, and an analysis and evaluation based on the experimentation are discussed.

Chapter 9 concludes the dissertation with a suggestion for future work.
Chapter 2

Background

2.1 Relevant Work

The human computer interaction (HCI) study is an interdisciplinary study area, Preece, et al. [106], Carroll, et al. [23], Dix, et al. [37], Laurel, et al. [73]. Preece, et al. [106] also noted that there are about a dozen areas that are related to the HCI study. In general, studying a multilingual multimodal HCI system involves many different disciplines. Therefore, this chapter discusses relevant work from various perspectives with regard to the topics in the dissertation, but mainly from the multimodal HCI system design perspective. Especially, due to the importance of the information verification approach in the dissertation, a portion of the discussion will be assigned to the relevant work in connection to the information verification approach.

2.1.1 Multimodal Human Computer Interaction and Information Processing

Since a multilingual multimodal HCI system is one type of HCI systems, issues and theories in designing a HCI interface are also closely related to the multilingual multimodal HCI system. The following paragraphs introduce some of the well-known theories in the field and discuss in relation to the approaches in this dissertation.

The debate between the “direct manipulation” and the “software agents” by Ben Shneiderman and Pattie Maes in 1997 [122, 123] raised up many interesting and important points that are worth while for an interface designer to think about in the HCI interface design. As the debaters pointed out, both approaches may work as complementary metaphors. In the complementary roles, the software agents take care of certain processes inside a complex task as an aid, while a user utilizes a direct manipulation interface with its own intelligence capability in the visual domain and a great control for the highly complex tasks.
The interface for the information processing approach proposed in this dissertation is related to both of the two different interface design approaches. The scope of the proposed information processing approach is related not only to the front end interface design, but also to the functionalities, with which an interface can process tasks efficiently and seamlessly. Therefore, the interface that is developed in connection with the proposed information processing approach will share many similar questions and problems to solve from both approaches, as discussed in the debate.

For example, Shneiderman [122, 123] showed the “FilmFinder” as one of the exemplary interfaces, to which the direct manipulation approach can be suited well. In the example, Shneiderman [122, 123] pointed out how the approach can give the user a sense of context for what to look at in addition to the sense of being in control, i.e. how the user can efficiently select desired information details and be responsible for the decision among the 1500 films.

The proposed information processing approach in this dissertation is also concerned with a large amount of data in many cases. Therefore, the strengths of direct manipulation, such as being in control for the information processing based on the expert user’s human intelligence on the information visualization, can definitely become one of the significant design considerations for the interface.

On the other hand, the multilingual information could be out of reach from the capacity of the expert user’s intelligence. The kinds of the multilingual information need not necessarily be uniformly defined. For example, multiple objects labeled with different human languages can show up on the same information visualization screen and require a user to have a multilingual understanding to process the multiple objects according to a goal at the specific time of interaction. In this case, if the user does not have proper knowledge for the multilingual objects, the accomplishment of the task can be hindered with much hassle. Therefore, the approach will need a help beyond the expert user’s human intelligence in this situation.

As an example of the potential help, the dissertation discusses a novel approach that enables the expert user to handle the multilingual information processing, at least partially, without requiring having full knowledge for the multilingual data that the user is handling. By focusing on the information content rather than the linguistic features, the lack of knowledge for the multilingual information from the language point of view can be overcome. The novel approach
is to utilize a processing structure for an augmented data based on the predefined rules. The details will be discussed in chapters 4 and 6.

The novel approach is independent from the idea of the software agents. Although the approach could be partially integrated into a software agent, where the software agent is implemented specifically with the intention of being used for the same purpose, it is not appropriate to consider the approach as a software agent. The approach is essentially a part of the algorithmic approach, but it is not a software agent. However, what they share is the idea of getting extra help beyond the expert’s intelligence and capability.

In another study with regard to the direct manipulation, Khella [64] pointed out some exemplary limitations in the direct manipulation as follows:

- Difficulty in executing a series of actions
- Difficulty in performing a pipelining technique
- Performance rate could be reduced as the time of interaction prolonged for a certain case.

The information processing approach in this dissertation can help a direct manipulation based interface to overcome the limitations. Since the approach is related to a whole multimodal HCI system rather than just the front end interface, a careful usage of the information processing approach can take care of the internal processes in a more efficient manner so that the limitations can be compensated by the other strengths with which the approach can contribute in the multimodal HCI system, particularly with the strengths of the ‘multimodality’. For example, although the pipelining technique may not be easily applied to the direct manipulation-based front end interface, the information processing approach can be carefully designed to take care of the pipelined processes internally, facilitated by the fusion of multimodal inputs.

The object-action interface (OAI) model discussed in Shneiderman [122, 123] is focused on task objects and actions and their isomorphic interface objects and actions, where the modeling starts with understanding the task. Thus, it was argued that the OAI model helps a user to focus on the task itself more than the tool the person is using, reducing the load of device dependent syntactic details in the memory.

Shneiderman [122, 123] noted that a graphical user interface (GUI) reduces the syntactic details in issuing commands to achieve the goals in the tasks compared to the command
languages. Therefore, he also noted that the users who know the task domain objects and actions can easily learn the interface due to the reduced syntactic details required to learn.

The OVEN is an information processing approach not a pure interface design model like the OAI. Therefore, the two approaches have significant differences that make a side by side comparison of the approaches inappropriate. On the other hand, it is worthwhile to compare a part of the concepts in the OVEN with relevant concepts in the OAI because it helps us to understand the OVEN approach better. For example, it facilitates an understanding of the term “object” in the OVEN by comparing the two approaches.

According to Shneiderman [122, 123], a task is composed of a set of real-world objects and actions in the OAI model. The objects are regarded as media with which the users work to accomplish their intentions. The actions are the operations that the users apply to the objects. The objects and actions are decomposed into sub-level objects and actions in a hierarchical structure. They also have metaphoric and isomorphic representations of the interface objects and actions.

In the OVEN, the primary target of interest is the verified object rather than the raw object. Therefore, in the OVEN, how an object can be verified with the appropriate additional, i.e. intended, semantics at the verification step has the key importance as far as the object is concerned.

The difference in the term “object” is a non-trivial matter. In the sense that both approaches emphasize the importance of an object, their interest shares a common goal to a certain degree. The differences will be further clarified in chapter 4.

2.1.2 Computer Supported Collaborative Work (CSCW)

The research area of the computer supported collaborative work (CSCW) studies how people collaborate with each other for a task and how the computer-based technology can help them to accomplish a successful collaboration.

As Ellis, et al. [39] also divided the cooperative work paradigms into 2X2 matrix in a taxonomy, we can observe and divide the collaborative work domain into four sub-divisions based on the coupling among the temporal and spatial attributes, such as 1) synchronous face-to-face interaction, 2) synchronous distributed interaction, 3) asynchronous same place interaction,
and 4) asynchronous distributed interaction. As discussed, the OVEN is a network, and it is related to all the four sub-divisions of the collaboration paradigms.

Ellis, et al. [39] defined “groupware” as computer-based systems that support groups of people for a task and that provide an interface to a shared environment. Groupware is closely related to the CSCW.

It is important to note that the proposed approaches, such as the OVEN or the language independent speech network (LISN), in this dissertation are not entirely groupware although some features in the groupware can be found in them. The OVEN is an algorithmic approach for information processing. The LISN may have similarities partially with the groupware according to the definition of groupware in Ellis, et al. [39]. One of the modes of the LISN is concerned with the collaboration among multilingual multi-users. However, the LISN does not necessarily operate only for the collaboration mode. For example, the LISN can also provide its capabilities to a single user who wants to process multilingual information as virtually all collaborative systems support individual work.

Still, parts of the methodologies and functionalities in OVEN and LISN may be implemented as groupware. For example, a set of the polymorphic operators in OVEN can be implemented in such a way that the group of people in the collaboration can share the same set of polymorphic operators for certain tasks while they can still maintain their own set of polymorphic operators. In another example, an interface can be developed as groupware for the collaboration mode of the LISN.

2.1.3 Object-Orientation in Research

The concept of “object” has more significance in some research areas than others. This section briefly notes the concept of “object” in traditional research areas in relation to the concept of the verified object in this dissertation. Detailed discussion of the relevant work in the research areas will be shown in chapter 4.2. The traditional research areas include studies of:

- object-oriented user interfaces (OOUI),
- object-oriented programming (OOP), and
Mandel [92] discussed that the object-oriented user interfaces (OOUI) is a means for taking the graphical interface environment beyond the simple representation of applications by truly focusing on the objects. According to Tibbetts [130], an OOUI invites the user to explicitly recognize and manipulate the objects on the screen. The ideas in OOUI are related to the approach for the output presentation of verified objects in this dissertation. The details will be discussed in chapter 7.3. In short, similar to the ideas in OOUI, the output method for verified objects in the discussion preserves the semantic importance of the verified object that was assigned at the object verification process. In order to express the verified objects according to their intended semantics and fulfill the purpose of the object verification, a novel and intelligent approach for the presentation of the verified objects is suggested in chapter 7.3.

In object-oriented programming, the key programming paradigm is shifted from implementing functions or procedures to classes and objects that are instantiated from the classes in achieving the goals of the programming. Stroustrup [125] noted that the class is a user-defined type. The OOP contains practices that can be highly useful in the programming environment. However, a novel approach for using the concept of object is introduced as object verification in this dissertation. The object verification has significant differences from the object orientation in the OOP. The details will be discussed in chapter 4.

The object-oriented database management system (OODBMS) is one way of designing a database management system. According to Ullman, et al. [135, 136], the object definition language (ODL) is a proposed standard language for specifying the structure of databases in object-oriented terms. The OODBMS and the ODL are particularly related to the method of how to store the multilingual information objects in this dissertation. The detailed discussions about the relationship between the OODBMS and the data management topics in this dissertation will appear in chapter 7.2.
2.1.4 XML and Web Services

XML is one of the markup languages that are used to develop web pages in the World Wide Web. XML has gained significant popularity in recent decades due to its versatility. Web Service Definition Language (WSDL) is a programming language for describing network services in a XML format. The Web Services are a sort of pre-developed utility programming codes available on the World Wide Web as the network services. A study in Dymetman, et al. [38] shows usage of the XML for multilingual document authoring.

The WSDL is an example of the guideline for object verification. Although, the object verification does not require any rigid form of verification procedure, following a guideline such as the WSDL certainly helps in organizing the verified objects and reducing their processing time in a manageable state. Further discussion of XML and the proposed approach will appear later, including chapter 7.4.

2.1.5 Multilingualism in a Multimodal HCI System

Topics for the multilingual multimodal system are found in Lavie, et al. [74, 75]; Burger, et al. [21]; Costantini, et al. [30, 31]; Zue, et al. [154]; and Almeida, et al. [3, 4]. In addition, Vicsi, et al. [137, 138], Perugini, et al. [105], Lyu, et al. [86, 87], Turunen [134], and Tanaka [128] also discussed their work with regard to a multilingual multimodal HCI system.

Challenges in the semantic correlation are noted in the previous work. Linguistic differences and their resulting problems have been understood explicitly and implicitly by researchers focused on multilingualism, including the area for the multilingual multimodal HCI systems. One of the objectives in this dissertation is to complement the previous work by addressing a deterministic semantic correlation approach among the multilingual information pieces within a multimodal HCI system.

The research in this dissertation includes a discussion for designing an experimental multilingual multimodal HCI system that supports the proposed ideas in the dissertation as a proof of concept. The experimental system is called “language independent speech network”
(LISN). The following sections briefly introduce relevant past work with regard to the LISN. More detailed discussions of the literature will appear in chapter 6.3.

First step of multilingual understanding is to identify the language the user is speaking before the uttered speech signal is processed by the system if the language is not manually identified. Navrátil [98] described a phonotactic-acoustic system as a successful approach to the automatic language identification (LID), in the topic of language identification. According to Navrátil [98], a compromise between full lexical decoding and a simple phonotactic principle must be found in a LID system in order to make it feasible for real applications.

With regard to identifying the speech recognition engines (SREs), the language switching module (LSM) of the LISN takes an approach from a different perspective. Currently, the implementation of the LSM uses simpler, non-automatic language switching methods. This is partially due to the fact that the primary interest of the LISN for the multilingualism lies in the multilingual network rather than in an individual multilingual speech recognition engine. Another reason is that the LISN is developed for a multimodal HCI system, which already provides a simpler and more reliable input control modality for language identification and switching although it may not be automatic.

Waibel, et al. [141] mentioned two different methods for multilingual speech recognition. One is to integrate multiple monolingual recognizers with a language identification module, and the other is to use a multilingual speech recognition engine.

With regard to multilingual speech recognition, the LISN can work with either of the methods, or even both simultaneously. Regardless of whether the speech recognizer is a multilingual recognizer or a collection of monolingual recognizers, the LISN is designed to be extendible with any type of recognizers. Extendibility is one of the key ideas in the LISN because the LISN aims to serve collaborators from various resources.

Issues for multilingual features are discussed in many application domains, such as multilingual spoken language processing, Lamel, et al. [72], multilingual web sites, Huang, et al. [55], Liu, Z., et al. [81], Maeda, et al. [91], Mehta, et al. [94], Morgan, et al. [96], multilingual agent, Massaro, et al. [93], Read, et al. [108], information retrieval, Lin, C. H., et al. [79], Perlman [104], information retrieval from web-documents, information access to multilingual database, Hsiao, et al. [54], indexing multilingual information, Yip, et al. [152], video text detection, Kim, J. H., et al. [66], text mining, Lee, C. H. [77], text encoding, Mudawwar, et al.
[95], and text-to-speech synthesis, Black, et al. [18], Evans, et al. [41]. However, this dissertation mainly focuses on the multilingualism in a multimodal HCI system.

2.2 Scientific Challenges

Main scientific challenges found in the previous work with regard to the multilingual collaboration research area can be summarized as follows.

- Non-trivial tasks in 1) integrating the object orientation concept into a multimodal HCI system and 2) modeling a novel interface design based on the object orientation.
- Lack of knowledge and previous work about the issues in designing an interface that processes multilingual information in a computer supported collaborative environment.
- Lack of a novel information processing approach and a language independent speech processing framework that enable a dynamic processing of highly complicated multilingual information.
- Complexity in mapping from one speech segment in a language to another speech segment in a different language, where the choice of the languages by the collaborators are random from the system point of view.
- Handling the differences in grammar, vocabulary, idiom, expression, word-formation among human languages.
- Handling the existence of unique homonym, polysemy, and homophone in each human language.
- Semantic correlation and searching algorithm for matching keyword in the target language, including the application of the contextual information to the semantic correlation.
- Multilingual visualization, data representation, and feedback to the collaborators in multilingual audio, visual, textual format.
- Multilingual database management and retrieval.
- Language identification of the speech input among multiple human languages.
2.3 Objectives and Significance of The Research

The discussions in this dissertation extend studies in the relevant previous work partially as summarized above and complement their work by suggesting a deterministic information processing framework as a part of the attempts to solve the challenges.

The objectives of the dissertation and its expected significance on the state of the art can be summarized as follows.

- It is an objective of this dissertation to propose a novel approach for integrating the ideas in object concept into a multimodal HCI system and for modeling an interface design based on the novel information processing approach, called object verification.
- It is another objective of the dissertation to introduce a novel information processing approach and a language independent speech processing framework that enables a dynamic processing of highly complicated multilingual information.
- It is expected that a meta-language, called object representation language (ORL), in the novel approach, reduces the complexity in mapping among the multiple languages. Based on the fact that the usage of a meta-language will eventually reduce the mapping complexity from $O(n^2)$ to $O(n)$, the dissertation will discuss the expected complexity improvement in the evaluation study.
- It is another objective of the dissertation to include a discussion for handling the differences among the human languages and the issues of unique homonym, polysemy, and homophone in each human language using an information modeling approach while having a discussion about how the rules can be constructed in the study of the deterministic rule-based information processing approach.
- It is another objective of the dissertation to provide an efficient strategy for processing the semantic correlation and the application of the contextual information. From the previous work by the Costantini, et al. [30, 31], we can derive a conclusion that the multimodality in addition to the deterministic information processing approach will help in semantic correlation process. Therefore, it will be worthwhile to discuss how the
combination of the multimodality and a deterministic information processing approach further enhances the semantic correlation among multiple human languages.

- It is a further objective of the dissertation to study language identification of the speech input among multiple human languages through a novel approach called language switching module (LSM).
- It is another objective of the dissertation to discuss an efficient and novel strategy for processing multilingual output in various data formats.
- It is also hoped that the discussions with regard to the multilingual database management, information retrieval, web-based collaboration, and mobile extension, based on the novel deterministic information processing approach enables us to understand the potentials of how the information processing approach can be utilized in the areas, respectively.
- It is a strict objective of the dissertation to limit the domain of the research within the context of the geo-spatial collaborative crisis management. Therefore, the corpus of the multilingual data for the dissertation will be also limited.

Overall, it is hoped that the objectives of the dissertation serve to improve the understanding of multilingual collaboration in a multimodal HCI system and to provide a proof for the feasibility of developing a multilingual multimodal HCI system within the restricted application domain.
Chapter 3
Multilingual Multi-user Multimodal Human Computer Interaction

3.1 Overview

This chapter introduces a multilingual multi-user multimodal HCI system, called MultiLUM, as an approach that embodies the ideas in this dissertation. The research problems in the multilingual multi-user multimodal human computer interaction (HCI) are divided into three different areas in this dissertation, as described in the chapter 1:

1. study of multilingual multi-user human computer interaction in a multimodal HCI system
2. the influence of the proposed approach in a computer supported collaborative work (CSCW)
3. topics in multilingual speech processing

In relation to the areas, this chapter discusses the following topics:

- known issues in designing a HCI interface, particularly with regard to a multilingual multimodal HCI system in a computer supported collaborative work (CSCW) environment - The issues include empirically found facts in a real-world application of a HCI interface.
- some of the well-known theories in the HCI interface research field
- an overview of the proposed approach as a solution to the discussed issues - The components of the proposed approach are also discussed.
3.2 Multimodal Human Computer Interaction

According to Sharma, et al. [119], the multimodal interfaces allow users to interact via a combination of modalities such as speech, gesture, pen, touch screen, displays, keypads, pointing devices, and tactile sensors. In other words, a multimodal HCI system is a computer-enabled system that allows users to have a capability of selectively using one or more available interaction methods among multiple interaction modalities in a human and computer interaction situation. Relevant work is also found in Oviatt, et al. [101], Reeves, et al. [109], Amores, et al. [5], Jiang, et al. [59], Karpov, et al. [61], Teh, et al. [129], Yu, et al. [153]. In this dissertation, the interaction modality is primarily concerned with the input modality rather than the output modality. The other output modalities or output sensory than a display screen, such as a tactile output, are not discussed in this dissertation.

In modern era, humans heavily rely on various computing devices that process many different forms of information. In the past decades, the keyboard and mouse have been the primary interfaces between the computer and the user. This is partially due to the way the popular operating systems were designed for personal computers. In this keyboard and mouse interface approach, the coordinates associated with the mouse movement on the screen were used to represent what the user wants to do with the graphical objects depicted on the screen. For example, the interaction with a mouse was used for selecting an object, i.e. point and click, or for moving an object from one point to another on the screen, i.e. click, hold, drag, locate, and release. For this approach, the organization of the complicate information was simplified through the abstraction of the information by graphical objects, such as the file system. The interaction with a keyboard was used to insert texts to a memory space representing a file, to send commands in a process queue, to transmit a combination of codified information, e.g. alphabets, numbers, and symbols, to another work station in a network, and so on.

However, as the interaction paradigm evolves, the reliance on the keyboard and mouse began showing limitations in certain applications and tasks. Sharma, et al. [119] discusses the well-known problem of information bottleneck in using the conventional interfaces. Therefore, it seems that having an alternative interaction modality becomes more important as the interaction paradigm evolves, especially in order to handle more complicated tasks. In addition to this, as the types of computing devices are further changing from a personal computer to smaller and more
portable computing devices, such as mobile phone, smartphone, tablet PC, automobile communicator, multimedia player, motion-based video gaming console, and computing devices with a 3D display, the importance of alternative input modality seems to be increasing.

MacEachren, et al. [88] and Sharma, et al. [119] discussed the relationship among the crisis management, geospatial information, and collaborative group of people. According to MacEachren, et al. [88] and Sharma, et al. [119], most crisis management relies on geospatial information, because the geospatial information is essential for an assessment before an event, responses during the event, and recovery efforts after the event in the crisis management. Crisis management also frequently relies on a group of people who handles the geospatial information in collaboration. However, MacEachren, et al. [88] and Sharma, et al. [119] pointed out the fact that geospatial information technologies until now have not been designed to support such a collaborative group work. Especially, MacEachren, et al. [88] and Sharma, et al. [119] pointed out the lack of scientific understanding of how the group of people works with the geospatial information in the collaborative work environment.

Furthermore, there has been only a few work with regard to the multilingual collaboration in such a crisis management situation, dealing with the geospatial information, Tomaszewski [132], although the importance of geospatial hotspot was noted by the researchers, Patil [103]. Therefore, it is one of the objectives of this dissertation to study the scientific analyses and understanding in multilingual collaboration in the context of dealing with geospatial information. The proposed approaches in this dissertation can be discussed in other multilingual information processing applications, but the discussion is related to the multilingual collaboration dealing with multilingual geospatial information in this dissertation. In addition, the crisis management situation is an exemplary incidence to which the study of multilingual collaboration can be well applied.

As Sharma, et al. [119] discussed, the knowledge of natural integration patterns will play a key role in designing and using a multimodal interface system. Since not all the users feel comfortable in using all the available modalities in a multimodal interfaces, they will select the type of interfaces that they feel more comfortable with. Thus, information processing among the multiple input modalities needs more scientific study.
3.3 CSCW and Geospatial Collaboration

Rauschert, et al. [97] noted that geospatial information is critical for collaborative decision making in an emergency management situation. Further extending this view, it is our belief that the geospatial information is also important for tasks in a non-emergency management situation. As the study in Kettebekov, et al. [63] shows, geospatial information is also closely related to the deictic gestures for the spatial information as an input modality in a multimodal HCI system.

When the geospatial information consists of multilingual information or if the collaborators come from diverse linguistic backgrounds, the geospatial collaboration could become more complicated. When the targeted geospatial information comes from internationally distributed physical spaces with multilingual form, achieving the goal of the collaboration could depend on how well the collaborators understand the geospatial information that contains multiple human languages.

As it will be discussed later, if an interface enables the collaborators to speak their own languages in dealing with the multilingual geospatial information, such as local names that are possibly written in the local language for the objects on a shared map displayed on a screen, it will help them to concentrate more on their collaborative tasks rather than paying too much attention to peripheral linguistic issues.

3.4 Multilingualism and Computational Linguistics

Grishman [48] noted that computational linguistics is the study of computer systems for understanding and generating natural languages. As discussed in chapter 1, Schultz, et al. [115] showed different aspects of the research problems in the multilingual speech processing in a comprehensive manner. Delgado, et al. [35], and Gibbon, et al. [46] discussed the multimodal multilingual spoken dialog systems, and Ipsic, et al. [56] discussed multilingual spoken dialog systems.

In the rule-based language modeling, the rules are usually written in Backus Naur Form (BNF) notation. The multilingual grammars are typically written separately for each language, and the model poses stricter restrictions on the user’s input sentences. BNF is for language modeling whereas the LISN uses a heuristic information processing approach to map semantic correlation among the multiple languages based on user-defined rules. It is important to clearly point out that the rules in the LISN are for defining the relationship among the linguistic objects in a higher level of abstraction rather than defining the multilingual grammars or modeling a language.

In statistical language modeling, the system can deal with spontaneous input, but it requires a large dialogue corpus for the specific application domain and each of the languages, which is a difficult task.

Delgado, et al. [35] noted that a multilingual system can have a speech recognition module for each language and a language independent common semantic representation. Delgado, et al. [35] also discussed three approaches to the multilingual dialogue system architecture as follows:

- **interlingua approach**,  
- **semantic-frame conversion approach**, and  
- **dialogue-control centered approach**.

The interlingua is basically used as the language independent common semantic representation. The semantic-frame conversion approach is a derivation of the interlingua approach. Depending on the level of the interlingua, the multilingual dialog system can be 1) task-independent or 2) task dependent.

The task-independent system focuses on syntactic relations and enables a grammar to be reusable for other tasks. However, it does not guarantee that a semantically identical object is represented in the same interlingua representation. On the other hand, the task-dependent system focuses on semantic relations in the task and its rules are written using semantic categories instead of syntactic categories. Thus, it is difficult for the grammar to be reusable for other tasks since they will have different semantic structures.

While the information modeling structure of the object representation language (ORL) in the LISN focuses on the semantic relation among the multilingual information elements, its
processing is controlled by a set of polymorphic operators (POs). The set of POs provides the flexibility of the semantic correlation to a large extent. The set of POs can also be re-used for other semantic correlations easily, depending on the level and characteristics of the rules that define the relationships among the multilingual information elements, although it is true that the set of POs may have a weakness of causing ambiguities among the similar semantic representations by multiple multilingual elements that have the similar meaning among each other.

Therefore, the object representation language (ORL) in the LISN can be understood as an interlingua-like layer that provides a hybrid approach to a multilingual dialog system. The ORL is essentially a layer of a set of object representations that tells the semantic relations among the multilingual verified objects.

Another reason that the ORL is more than the interlingua is that the ORL represents the most significant meaning or most significantly intended meaning for the goal of the multilingual information element. This feature assures that the ORL naturally emphasizes a single most important meaning among other possible multiple meanings in each of the multilingual information units.

As pointed out by Delgado, et al. [35], only a few machine translation systems have been implemented using the interlingua approach because it is not a trivial task to establish a language independent common semantic representation, handling the differences among the multiple human languages.

From the points of the LISN and the ORL, the object verification process is performed by the expert users, so that unique language specific features that are too exotic to the language-independent common semantic representation can be filtered out at the object verification steps in a specific domain. Although this process may seem to limit the scope of semantic representation for the filtered out elements in some languages, the process helps the multilingual system to organize the command structure and promote unified understanding among the collaborators. Thus, the ORL can be a solution for an efficient language independent common semantic representation.

This is one of the reasons why it was emphasized that the approaches through the LISN and the ORL are not for language modeling but for information modeling. Unlike the rule-based modeling in the traditional approach, such as the BNF notation that is related to the language
modeling, an approach of using information modeling is used to correlate semantics among the multiple languages through an information augmentation process, called object verification, in the LISN and the ORL. In addition, the LISN and the ORL are not approaches for a machine translation system either.

3.5 Proposed Approach

3.5.1 MultiLUM

The following discussions cover various aspects in the design and development of a multimodal HCI system although not all of them will be covered with the same level of depth. The research is mainly focused on the deterministic information processing framework and its role for the multilingual collaboration in a multimodal HCI system. Therefore, the study investigates the answers to the objectives of the research, described in chapter 2.3.

![Multilingual Multi-User Multimodal HCI System (MultiLUM)](image)

Fig. 3.1 Multilingual multi-user multimodal HCI system (MultiLUM)
The Fig. 3.1 shows the components in the proposed approach. We call the approach that embodies the ideas in the research the multilingual multi-user multimodal HCI system (MultiLUM). The MultiLUM is not just a multilingual speech recognition engine or a generic HCI system. The MultiLUM further designates the ideas, algorithms, and practices for designing a multilingual multi-user multimodal HCI system.

### 3.5.2 Key Features and Methodology

The key features with regard to the components of the MultiLUM are:

- **Language Independent Speech Network (LISN):** An approach for enabling multilingual speech understanding and semantic correlation for the speech segments in various configurations of the system, such as for distributed speech understanding and collaboration.

- **Object Verification Enabled Network (OVEN):** An algorithmic approach and framework that is used for handling multilingual information and for enabling other modules to collaborate with each other seamlessly within the multilingual multi-user multimodal HCI system. The OVEN defines an augmented information unit based on data and event as a ‘verified object’ and provides a processing structure, called a ‘polymorphic operator’ that applies relational operations among the verified objects based on predefined rules.

- **Verified Object and Multimedia Output (VOMO):** An approach for the output of verified object and multimedia defined by the OVEN. Geo-spatial information could comprise various types of multilingual information, such as multilingual legend and names. The VOMO discusses a possibility of facilitating the multilingual geo-spatial information visualization.

- **Context based Reasoning Module (CBRM):** An approach for utilizing the contextual information to enhance the information handling.

- **VeRification EnAbled Database (VREAD):** An approach for storing, organizing, and retrieving the verified objects in a database.
- Object Verified Gesture Recognition (OVGR): An approach for understanding the gesture recognition in relation to the verified objects.
- Mobile Extension: An approach for integrating and utilizing mobile devices to the multilingual multi-user multimodal HCI system.
- User Collaboration Handling Module: An approach for helping in the collaboration among the multi-users based on an information processing model, such as the OVEN.
- Measurement and Evaluation: Methods to test the proposed approach.

This dissertation utilizes the OVEN as an exemplary information processing approach that will enable the proposed features in the MultiLUM. The foundation of the ideas in the OVEN is based on years of collective experiences in developing multimodal HCI systems by many researchers and engineers. Based on the foundation, we created the OVEN as an information processing approach that operates predefined tasks based on deterministic rules that are defined by expert users in collaboration. We have been also applying the OVEN approach to develop the key features discussed above. One philosophy in the OVEN is to come up with a scientifically organized deterministic approach rather than a collection of ad hoc knowledge. It is believed that the original ideas and core philosophical contributions with regard to the adaptation of the information processing model uniquely represent an approach to the development of the novel multilingual multi-user multimodal HCI system in a specific context. The approach is based on the practical needs in a real-world situation. Thus, one of the contributions in the approaches is to come up with a practical suggestion that is applicable to real-world multimodal HCI systems.

### 3.5.3 Scientific Hypothesis

According to the ‘Merriam-Webster online dictionary’, the word ‘hypothesis’ is defined as follows:
1. a : an assumption or concession made for the sake of argument
   b : an interpretation of a practical situation or condition taken as the ground for action
2. a tentative assumption made in order to draw out and test its logical or empirical consequences.
Accordingly, a hypothesis in this dissertation can be constructed as follows:

“If a multilingual multi-user multimodal HCI system utilizes the proposed deterministic information processing framework, it can handle the complexity of multilingual information processing in a more efficient way than the conventional method, through the seamless integration of the multimodal dialog management to the main multimodal HCI system and through the shift of the focus from the language specific features to the information processing features, while providing a practical solution to the problems in previous work.”

3.6 Discussion

In this chapter, a novel approach, called MultiLUM, was suggested for a multilingual multimodal HCI system. The approach was briefly discussed from the perspectives of three research areas: the multimodal HCI system area, the computer supported collaborative work (CSCW) area, and the multilingual speech processing area. A brief overview of the proposed approach and its key components are also introduced. The MultiLUM and its components are discussed throughout the following chapters as a part of the attempt to prove the hypotheses described above and in chapter 1.2.
Chapter 4

Information Verification

4.1 Overview

This chapter describes a novel information processing approach that can improve the performance of a complex information processing system. A multimodal human computer interaction (HCI) system is a good example of the complex information processing system that can benefit from an intelligent information handling approach. This is even further true for the multilingual multimodal HCI system. The novel information processing approach can enhance the processes of handling multilingual data in a multilingual multimodal HCI system.

Our approach for the novel information processing includes an idea of applying information verification techniques derived from a framework called object verification enabled network (OVEN), Jung, et al. [60] (Usage of the proprietary technology is granted for academic purpose in this dissertation by the VideoMining Corporation.). The Fig. 4.1 shows an overview of the OVEN. The approaches in the OVEN can provide a heuristic and feasible computational framework to a multimodal HCI system that needs to handle complex information units created from the complex data and events at various levels in the system. The OVEN comprises an approach to augment information units, called ‘objects’, in a complex information processing system through the ‘object verification’ and to process the verified objects through the application of processing structures, called ‘polymorphic operator’, based on the rules that define the relationship among the verified objects. In short, the OVEN comprises the following steps of:

- converting an information unit into a verified object,
- expressing the verified object by object representation, and
- processing the verified object through a processing structure, called a polymorphic operator, based on the predefined rules in a particular application domain. The predefined rules can be stored in a rule-base, and the rules are defined by the users of the system.
A particular embodiment of the OVEN can comprise a collection of networked modules, so that the OVEN can extend the information processing paradigm to the domain of web-based applications, such as the web-based geo-visual analytics application by MacEachren, et al. [88], in addition to the standalone applications. The OVEN can utilize the strengths of a multimodal HCI system in the distributed web environment, along with the traditional multimodal HCI applications. For example, the approaches in the OVEN in a web application can facilitate multiple user interactions where the interactions are based on multiple human languages and modality in the highly distributed computing environment. In this embodiment, the OVEN can naturally rely on remote resources, such as a remote database. For example, the OVEN can utilize a remote verification process for the objects in the network environment. This is one of the reasons why flexibility is important in the OVEN approach, although it may increase the
complexity in processing the information units. Balancing between the flexibility and the correctness of information handling is one of the key tasks in the OVEN, and its validity will be discussed in a numerical analysis in chapter 8.

4.2 Relevant Work

This section includes further detailed discussion of the relevant work that was briefly introduced in chapter 2.1.3 with regard to object-oriented user interfaces (OOUI), the object-oriented programming (OOP), and the object-oriented database management system (OODBMS), which are the representative research areas where the concept of the object is the key focus.

4.2.1 Object-Oriented User Interfaces (OOUI)

Mandel [92] discussed that object-oriented user interfaces (OOUI) are means for taking the graphical interface environment beyond the simple representation of applications by truly focusing on the objects. Mandel [92] further described the differences between the OOUI and the application oriented user interfaces (AOUI) in detail. For example, in OOUI, the product consists of a collection of cooperating objects and views of objects, whereas in AOUIs, an application consists of an icon, primary windows, and secondary windows. The focus is on inputs and outputs for objects and tasks in OOUI, whereas it is on the main task in the AOUIs as determined by the application.

Tibbetts [130] explained the characteristics of the OOUI by pointing out differences between the OOUI and the graphical user interface (GUI). According to Tibbetts [130], an OOUI invites the user to explicitly recognize and manipulate the objects on the screen. An OOUI appears to be a simulation, not a representation, of reality as in the GUI. For example, an icon itself is an object rather than a representation or link to another object, and clicking on the icon is a command to the icon to execute the assigned action. From the OOUI point of view, the GUI is just a low-level user interface UI abstraction. Most importantly, Tibbetts [130] pointed out that
the OOUI utilizes a GUI for its set of components but adds to it a significant semantic content on the meaning of objects and user interaction.

The GROW by Barth [13] is an example of an object-oriented system for building a GUI. In GROW, the graphical objects are organized and shared in a taxonomic hierarchy. In the taxonomy, the graphical objects are related to each other through composition and graphical dependency. A separation between the interface and the application is also one of the key points in the GROW. Flamingo by Anderson [6] is another example of the object-oriented user interface especially for a distributed environment. Voisard [140] showed an interesting work for integrating a graphical geographic database user interface (GDUI) into a database management system.

The ideas in OOUI are related to the approach for the output of verified objects in this dissertation. The details with regard to the output presentation will be discussed in chapter 7.3.

4.2.2 Object Oriented Programming (OOP)

In object-oriented programming, the key programming paradigm is shifted from implementing functions or procedures to classes and objects that are instantiated from the classes in achieving the goals of the programming. Stroustrup [125] noted that the class is a user-defined type. A class defines member data and functions or methods, and an encapsulation mechanism groups the data and the sub-routines together and hides irrelevant details from the users of an abstraction, according to Scott [116]. Classes can have a hierarchical structure, so that a derived class can inherit partial or whole features from the base class. The derived class can add its own members. A concept of virtual functions enables a polymorphic behavior of the objects, where the binding of the member function is decided dynamically at the running time of the program.

The OOP contains practices that can be highly useful in the programming environment. However, within this research, a novel approach for using the concept of object is introduced as object verification. The object verification in this research has significant differences from the object orientation in the OOP, as discussed in chapter 4.3.
4.2.3 Object-Oriented Database Management System (OODBMS)

The object-oriented database management system (OODBMS) is one way of designing a database management system. The object definition language (ODL) is a proposed standard language for specifying the structure of databases in object-oriented terms, in which the concepts of the object orientation is similar to those in the object oriented programming languages, according to Ullman, et al. [135, 136].

The OODBMS and the ODL are particularly related to the method of how to store the multilingual information objects in this research. The detailed discussions about the relationship between the OODBMS and the data management topics in this dissertation will appear in chapter 7.2. The ‘entity-relationship’ (E/R) model is a graphical approach to database modeling, which is very similar to the object-oriented approach. The differences between the E/R model and our database model based on the OVEN will also be discussed in chapter 7.2.

4.2.4 XML and Web Services

The web service definition language (WSDL) is an example of the guideline for the object verification. Although, the object verification does not require any rigid form of verification procedure, following a guideline similar to the WSDL certainly helps in organizing the verified objects and reducing their processing time to a practically manageable state. Especially, for a collaborative task, the participating collaborators should share the same object verification information for the verified objects. Therefore, the guideline can help them to provide common knowledge with regard to the object verification information.

If a guideline is not available, the processing of verified objects by a polymorphic operator could be costly because a part of the verification information in the verified objects may not match with the targeted verification information by the polymorphic operator.

From another point of view, not having a guideline could become a cause for generating unexpected results for the polymorphic operations. In other words, when a polymorphic operator is designed to search for certain targeted or meaningful information among the unknown verified objects, the process can lead to a result that may be beyond the boundary of the guideline. For
example, a database may consist of representative but incomplete verification information that may not match all the available verification information for a certain group of verified objects, yet still covering them in a significant enough manner. When a polymorphic operator searches and utilizes the database for an operation, the result of the polymorphic operator processing can still return a valuable result utilizing the partial verification information. The polymorphic operator can be designed in such a way that it uses only partial verification information in the verified objects when the guideline is not easily available. This can help the users to find out how closely the verified objects can be matched for the given search task.

4.2.5 Object-Action Interface Model

With regard to the object-action interface (OAI) model discussed in chapter 2, the OVEN cannot be directly compared side by side with the OAI because the OVEN is an information processing approach whereas the OAI is an interface design model. However, it is worth while to compare some concepts in the OVEN with seemingly relevant concepts in the OAI, in order to facilitate further understanding of the OVEN approach. For example, it is meaningful to understand how the term “object” is differently defined and used between the two approaches.

According to Shneiderman [122, 123], the task is composed of a set of real-world objects and actions in the OAI model. The objects are regarded as media with which the users work to accomplish their intentions. The actions are the operations that the users apply to the objects. The objects and actions are decomposed into sub-level objects and actions in a hierarchical structure. They also have metaphoric and isomorphic representations of the interface objects and actions.

In the OVEN, the primary target of interest is the verified object rather than the raw object. Therefore, in OVEN, how an object can be verified with the appropriate additional semantics at the verification step has the key importance as far as the object is concerned.

The verified objects are represented by the object representation. The object representation does not have to be only represented as the visualization of the verified object. The object representation of the verified object in the OVEN can be expressed in various methods and types of data.
The actions among the verified objects are performed by the applications of the polymorphic operators according to the predefined rules. In this scheme, the verified objects and their executions, i.e. actions, are completely separated. From one point of view, the intentions of how to utilize the objects are delegated to the external processing structures, and the key role of the verified objects is only to provide their targeted semantics into the object representations in the OVEN.

4.3 Object as an Information Unit based on Data and Event

In the approach of the OVEN, the information units in an information processing system are called ‘objects’. Among many different types of information units, the OVEN focuses on the information units based on data and event and further defines the definition of an object as follows:

- First, the object can be an ‘information unit based on data’, which can appear in the input modalities, output representation, or internal processes within the system handled by the processor.
- Second, the object can also be an ‘information unit based on event’. The creation of information pieces is often event driven. The OVEN treats each single event, or a combination of such events, as an individual object.

4.3.1 Information Unit based on Data

When a multimodal HCI system processes speech recognition, it applies a speech recognition algorithm through a speech recognizer to the speech signal. The recognized speech segments are understood based on grammar files. Likewise, when the multimodal HCI system processes a sequence of captured video images, the two-dimensional array of image data stored in a memory space is passed to the processor that applies computer vision algorithms to the data. These input
modalities essentially create information pieces based on the data with which the system understands the user’s command and possibly intention in the semantic level.

From the output modality point of view, the response or feedback that is given to the user can form an object that is augmentable with additional semantics. Visualization of the pre-stored output data, such as graphics, or textual messages, is one of the ways to send the information piece to the user. Audio instruction, such as a ‘.wav’ file in a memory space, is another example of output information delivery. In this case, it can be said that the information units are created based on the output data.

Some information pieces are created internally by the internal processes within the system, without involving a direct input or output process between the system and the user. These information data pieces are also created as a result of the interaction between the system and the user. For example, when an input is gathered and processed, the processor can create a message formed in a data structure that contains a data member for the processor’s internal processing. The message in the data structure is an exemplary information data piece created by an internal process. Resulting data that is directly queried from a database by a system in an automatic data query procedure is also another example of the information unit based on a data piece created by an internal process.

In the OVEN, all of these information units that appear in the input modalities, output representation, or internal processing steps are regarded as the object, i.e. ‘information unit based on data’.

4.3.2 Information Unit based on Event

The object also includes an information unit based on event in the OVEN. The creation of information pieces is often event driven. The OVEN treats each single event, or a combination of such events, as an individual object. The definition of event in the OVEN includes any computing processes or outcome of the processes, which comprise the following:

- Result of a process, such as the return value of a function,
- Relevant attributes to a data,
• Concept derived by executing a procedure in the application, like contextual information,
• Functionality of the process, and
• Relationship among the processes. For example, in a sequence of processes, the relationship can be represented in the history of executing the processes or how a process influences the execution of another.

The information unit based on an event will eventually be represented in a data form that a computer can access and process. Therefore, the final object produced from an information unit based on event may appear to be the same as the information unit based on a data piece created by an internal process in a computer. However, the information unit based on event emphasizes contextual information of how the information unit is generated, such as the relationship of the information unit with other variables in the event. Therefore, the operation that is applied to this type of object, i.e. information unit based on event, is differently formulated than that of the other type of object, i.e. the information unit based on data piece created by an internal process.

The event can be triggered by

• a user interaction or a direct command,
• a system in response to a user interaction, as in a feedback,
• another process, as in a callback function,
• an interaction by another participant in a multiple user collaboration situation,
• a resulting process derived from a group of parent events, or
• a processing instruction for a data piece from a network.

4.3.3 Further Details of the Term ‘Object’

The definition of the object is abstract, but the embodiment of the abstraction can exist in various computer accessible media forms, such as in textual, numerical, visual, or audio form. The embodiment can exist in a local system or in a distributed database remotely. One good example of the embodiment of the object representation is a XML-based data message file that contains relevant information for an event. Another example can be found in a speech
recognition process. When a speech segment is not recognized, the system can apply contextual information in order to guess the speech segment based on the history of interaction. In this case, the fact that the speech segment is not recognized, the fact that the system can apply contextual information to the speech segment, and the fact that the system maintains a history of interaction, can be defined as events. In a system, these events can be embodied into a message of parameters for Boolean value or a real numbered value providing a measurement for the degree of confidence for each of the events. These types of objects are classified as the objects created by the information unit based on the events.

The term ‘object’ was chosen rather than ‘information’ or ‘event’ for the concept described above in order to emphasize that the ‘object’ will be eventually transformed to the ‘verified object’ that is defined as an augmented information unit in the OVEN. The augmentation of the object is accomplished by a verification process’. This differentiates the object in the OVEN from a simple information piece or event in an information processing system in general. Furthermore, the term ‘object’ is clearly different from that of the object-oriented programming. The object in the OVEN is not relevant to the key features like data encapsulation, inheritance, or instantiation in the OOP.

The Fig. 4.1 shows that an object, such as ‘object A’, from a ‘database A in location A’ and another object, such as ‘object B’, from a ‘database B in location B’ are verified as the "verified object A” and the ‘verified object B’, respectively. If the verification is processed offline on these data, the process can be called a ‘static object verification’ process. An object database processing module (ODPM) can be used as an exemplary software module that helps the static object verification process. The Fig. 4.1 also shows a “dynamic object verification” process, which can be facilitated by a software module called “real-time object processing module (ROPM)”. In the Fig. 4.1, the “dynamic object verification” process can verify an exemplary information unit based on event, such as an “object C” from an “event in user interaction C” and an “object D” from an “event in system processing D”, into verified objects, such as “verified object C” and “verified object D”, respectively, in real-time.
4.4 Object Verification

The information unit is converted into a “verified object”, through a verification process in the OVEN. The verification process is a step to assign the object an additional meaning based on the characteristics of the object, in which the additional meaning is eventually intended to be the essential meaning or goal-oriented meaning of the object for a particularly intended application domain.

The need for the object verification is based on the observation that the information unit based on data or information unit based on event, i.e., an object, may not represent the object’s most desired characteristics by itself, especially when the application domain, where the information unit is processed, varies. For example, the information sometimes reveals its most appropriate meaning to the goal of the target process only when the contextual attribute to that information is known. Thus, it is meaningful for the object to be augmented with the key idea that helps the object to behave and be utilized as it is truly intended.

Accordingly, we assume that the information in the object needs more intelligent definition in order to make an application that processes the information to behave in a more contextually relevant manner. In order to make the information to be more intelligent, the information object can contain processing information, not member functions, inside itself. The processing information tells how the information in the object is to be treated. For example, the information can contain processing information for its contextual attribute. The step for augmenting the object with this kind of additional meaning and making it more intelligent is called the object verification (OV) process in the OVEN.

The Fig. 4.2 shows that the object verification process for multiple verified objects, such as the exemplary verified objects, “verified object A”, “verified object B”, “verified object C”, and “verified object D”, can be performed from various types of information units that reside locally or remotely, such as “data in input and output modality”, “data in user interaction and system processing”, “data in database at location A”, and “event”, which could possibly be created locally or remotely, with the exemplary object representations, such as “object representation A”, “object representation B”, “object representation C”, and “object representation D”, respectively.
The Fig. 4.3 shows an example of explicit verified objects. Verified objects can be represented explicitly in a data format as in the Fig. 4.3, or implicitly through constructing the polymorphic operators in such a way that it operates processes in relation to the verified objects according to the verification goals in a guideline.
4.4.1 Single Object Verification for an Object

Fig. 4.4 shows an exemplary single object verification for an object and relationship among the “object”, the “verified object”, the “object representation”, and the “object verification” process.

In the object verification process, the main characteristics of the object are represented in a single set of the verification information in the verified object. This representation is called “object representation” (OR). The object representation shows the key concepts of the object as the result of the verification regardless of the fact that the object is a simple object or a composite object, i.e. a collection of simple objects. The object representation does not include the object itself, but it is the representation for the object according to the verification. The embodiment of
the object representation can be just a name of the object in the simplest form or a collection of important characteristics of the object in a literal data type. Entire key concepts of the object can be represented in one composite object representation, or only a partial concept of the object can also be used as the object representation.

As mentioned earlier, the concept of ‘object’ and ‘object representation’ in the OVEN is different from the ‘object’ in the ‘object oriented programming’ (OOP). There is no concept of instantiation for the object in the OVEN. The ‘object representation’ is different from the concept of instantiating an object from a class in the object oriented programming practice, where one or a plurality of member data and functions are included as the definition of the class. The object representation can follow a class-like data structure as a guideline or conventional protocol to form the object representation. The concept of using a guideline or protocol further distinguishes the object representation from the concept of instantiation in the OOP.

The XML including the XML Schema is one of the exemplary tools that can be used to implement the object representation for the verified objects. The XML can also be used for other...
concepts of the OVEN, such as rule construction for a processing structure. One exemplary usage of the XML for the object representation is a XML-based data message file that forms relevant information for the verified object.

However, the key approach and interest in the object representation in the OVEN is more related to the representation of the semantics or meaning of the object into a verified object by giving the object an additional meaning based on the characteristics of the object rather than the tools. Thus, the concepts of formatting in a data document or transformation in the XML are less relevant to the object representation.

Unlike the XML Schema, the object representation in the OVEN does not have a predefined set of built-in data types. A verified object has a single set of the verification information, and each verified object is regarded as a complete single unit by itself. Although a verified object can contain processing information or type information of the verified object as a part of the verification information, they are not necessarily member function or built-in data type.

Although the object representation does not have a predefined set of built-in data types, the OVEN system should know the data type of the verification information in the object representation in the application of the processing structure, called the polymorphic operator (PO), to the verified objects. For this practical purpose, the processing information or type information of the verified object will have to match the data type that the processing structure, i.e. the polymorphic operator (PO), can handle.

### 4.4.2 Multiple Object Verifications for an Object

Fig. 4.5 shows exemplary multiple object verifications for an object. The object verification process may need to define an object in multiple concepts, i.e. multiple verified objects, depending on the goals of the target processes in the system. Examples are "verified object 1", "verified object 2", and "verified object N" in Fig. 4.5. The object can be represented in multiple object representations for multiple verified objects derived from the object, such as "object representation 1", "object representation 2", and "object representation N" in the Fig. 4.5.
Fig. 4.5 Multiple object verifications for an object
Each of the multiple verified objects has its own single set of the verification information through the multiple object verification processes, such as "object verification 1", "object verification 2", and "object verification N" in Fig. 4.5. When multiple object representations exist for an object, each of the verified objects is treated differently as a separate verified object, depending on the object representation. Accordingly, their embodiments exist separately in order to avoid ambiguity.

Each verified object and its object representation after the object verification is regarded as unique. In relation to the OOP, there is no concept of sibling objects instantiated from a parent class or assignment among the objects. Therefore, if multiple verified objects have exactly the same object representations for a single object, then they are regarded as the same verified objects for the single object, and only one embodiment is needed. For example, in the Fig. 4.5, each of the "verified object 1", "verified object 2", and "verified object N" is unique.

There can be multiple objects whose object representations may be identical. In this case, although the objects and their verified objects may not be identical, the verification information in the object representation, i.e. semantics, can be identical in their particular verified objects. Since a verified object consists of two parts, i.e. a pointer to the object and an object representation, the verified objects are not identical just because the object representations are identical. In this case, the same object representation means that the verified objects are semantically identical although they are not syntactically and their actual objects are not the same.

By being semantically identical, the information units can be essentially organized according to their object verification representation, i.e. semantically. If there exist multiple exactly same representations for multiple objects, then the objects are regarded as belonging to the same verification group. However, each of the verified objects in the same verification group is regarded as the unique verified object for each unique object.

Fig. 4.6 shows the relationship between the verified objects and the multiple objects. As shown in Fig. 4.6, each verified object among the multiple verified objects for a single object is unique and independent from each other for the single object.
Fig. 4.6 Relationship between verified objects and multiple objects
However, there is a possibility that multiple objects can have one or multiple identical object representation. For example, the “verified object i for object O1” and the “verified object 1 for object Om” have an identical “object representation i”. In this case, the two different verified objects are semantically identical based on the object verification information in the object representation, and the processing structure will apply the same rules for both verified objects in turn when the verified objects are sent to the processing structure.

4.4.3 Characteristics of Object Verification

The object verification process is often dependent on the context in which the objects reside. The information about the context is integrated to the object representation, in case the contextual information plays an important role. If the system is intended for dealing with multiple contexts, it is important to distinguish one context from another context. In this case, each contextual information can contribute to the generation of each object representation in multiple object representations, as discussed.

Depending on their scientific backgrounds, people in different research communities have different definitions for the meaning of ‘context’. In other words, the term ‘context’ has been used to indicate many different aspects and attributes for the different research environments. The definition of the context in our system comprises any information and computational environment that could characterize the object and influence the way the object is understood and applied to a particular information processing system in a specific domain. There is no restriction on which aspects of the research environment are used for the object verification process in our approach.

The object representation can use any human languages for the literal embodiment. For example, the object representation can contain “name” and “type” elements, in English or different human languages, as the representation for a verified object. Furthermore, the object representation can use numerical, symbolical, graphical, or any other form of data representation, as long as the relationship among the object representations is defined by the rule-base, so that the polymorphic operator understands the relationship among the object representations. Due to
this characteristic, multiple human languages can be simplified by the appropriate object representation.

In the case of the object as the information unit based on data, the object verification is performed primarily for the objects that are closely tied to the goal of the application among possibly numerous objects based on data. In this case, the object verification for the objects based on data can be done manually by the expert user prior to using the system, unless the database contains verified objects that had been verified previously. The expert user can use any available text editing tool to produce the verified objects, such as the explicitly written verified objects using XML tags in the Fig. 4.3.

In the case of manual object verification, not all of the objects in the task domain may need to be verified. Practically, it is not feasible to verify entire objects in a database of an information processing system, especially when the size of the database is huge. Even if entire objects in a database are verified, not all of them may be used. Therefore, the object verification is selectively processed according to the application goal. This exercise can be referred to as the “verification on demand”. When there is no verification representation in an object, i.e. the object may not be a verified object, the OVEN treats it as a normal object by default.

In the case of the object based on event, the object verification can be performed automatically by a verification module. For example, when an event is created by the internal processes, such as speech recognition or gesture recognition, the verification module can verify the event as a verified object according to a predefined verification procedure. For example, the speech segments or gesture units can be augmented with time stamps, demographic classes, or any available values for predefined parameters.

The module for the automatic object verification works based on a set of the predefined verification rules. It is important to note that the verification rules are different from the rules for polymorphic operators. The differences can be summarized as follows:

- The rules for the object verification define how the objects are verified, e.g. which attribute of the information unit will be augmented.
- The rules for the polymorphic operators define the relationship among the verified objects.
The verification rules are used to create primarily explicit verified objects, whereas the rules for polymorphic operators can be used to create implicit verified objects.

The rules can be programmed in the verification module if the number of rules in the set of verification rules is fixed. If an implementation needs flexibility in the automatic verification of the objects, the rules can be externally defined in a document file, such as a XML based document, and the verification module reads in the rules at the initialization step.

In any case, it is important for the author of the polymorphic operators to know about the object verification rules. It is because the author of the polymorphic operators needs to know what kinds of verifications are available within the verified objects. The operation rules for the polymorphic operators are applied to the available verification information in the verified objects. Therefore, as mentioned earlier, a guideline for the verification rules needs be propagated among the relevant collaborators for a specific task.

As was the case for the object based on the data, it will be difficult to verify entire objects based on events, especially when the processes for creating and handling events are complex. The number and type of objects based on event, to which the verification module can apply the object verification, are controlled by the rules according to the application goal. The result of the automatic object verification can be stored in a database.

It is worthwhile to note that the object verification is different from the concept of disambiguation, in the sense that the object verification aims to augment the target object with intended meaning whereas the disambiguation is to remove the ambiguity among the involved entities.

### 4.5 Polymorphic Operator

Once the object representation is assigned to each object, and as soon as the system recognizes the verified objects, they are sent to the processing module in the system. This can be triggered by the request from the user or the system. Then, the OVEN provides a processing structure that applies an operation among the verified objects and produces an outcome from the application of the operation. The processing structure is called the polymorphic operator (PO). Fig. 4.7 shows an exemplary rule for the polymorphic operator based on the Boolean logic in a
XML file, and Fig. 4.8 shows the “pseudo example of the polymorphic operator” based on the Boolean logic.

```xml
<rule type="and">
    <object type="literal">input_val_and_0</object>
    <object type="literal">input_val_and_1</object>
    <result type="literal">then_result_val_and_0</result>
    <result type="literal">else_result_val_and_0</result>
</rule>
-<rule type="or">
    <object type="numeral">input_val_or_0</object>
    <object type="numeral">input_val_or_1</object>
    <result type="numeral">then_result_val_or_0</result>
    <result type="numeral">else_result_val_or_0</result>
</rule>
-<rule type="not">
    <object type="literal">input_val_not_0</object>
    <result type="literal">result_val_not_0</result>
</rule>
-<rule type="nested_if">
    <object type="conditional">input_val_nested_if_0</object>
    <result type="numeral">then_result_val_nested_if_0</result>
    <result type="numeral">else_result_val_nested_if_0</result>
</rule>
```

**Fig. 4.7** Example of a rule based on the Boolean logic in a segment of a XML file

```xml
<rule> IF  'entity1' [Boolean logic] 'entity2' [Boolean logic] … [Boolean logic]
    'entity n'
    THEN  then-result
    { ELSE  else-result } // optional
</rule>
```

**Fig. 4.8** Pseudo example of the polymorphic operator using the Boolean logic
The polymorphic operator can handle both quantifiable and non-quantifiable operations among the verified objects. The relationship of operands is governed by predefined rules that are constructed by the expert users in a particular application domain where the system is used. The rules are stored in a rule-base that is also constructed by the expert user. The rules can exist in a distributed network, so that a remote rule-base can be referenced by the system remotely.

The rules can also define the type and format of the output result caused by the operation. How the type and format of the output result are created from the input is also defined by the relationship among the verified objects. The polymorphic operator (PO) derives an output based on the relationship that is defined by the rules in the rule-base.

The rule can define the relationship among the verified objects at either

- the information content level or
- the semantic group level, of the verified object.

When the rule is defined at the information content level, the PO mostly evaluates the value of the verified objects. For example, numerical evaluation of each entity may be performed for the numeral entity. The numeral entity can contribute toward constructing the numerical hypothesis in the “if” conditional statement as a confidence parameter. When the rule is defined at the semantic group level of the verified objects, the PO often compares the object representations of the verified objects using the keywords in semantic groups in the rule.

One exemplary data structure in operating the PO for each semantic group node is the disjoint set data structure. Based on the disjoint set theory, we can group or ungroup the verified objects using the polymorphic operator. A derived output does not necessarily have to combine two disjoint sets, but if the relationship is defined to merge the two or multiple disjoint sets of object representation, the disjoint set data structure and its operation is a good exemplary data structure to enable the operation. This will lead to a more complex reasoning after combining a set of objects. In this case, a divide and conquer algorithm-based approach for reasoning the newly created composite object can be used.
4.5.1 Architecture for Polymorphic Operator

For the concept of the PO and its operation, the architecture for the system is designed in such a way that various computational models for the PO can be integrated by the following design principles:

1) The architecture is physically independent from any specific information processing system. It is a standalone library, and it communicates with a domain specific application by an interface.

2) The architecture uses an external data file, such as a data file written in XML, which contains the rule information for the PO. The XML file allows users to define their own rules in the domain specific context. Thus, the computational model in the rules can be changed according to how the rules are constructed in the XML file.

Based on these design principles, the flexibility of the system allows adaptation for various computational models. Regardless of which computational model is used for constructing the PO, the system is designed to provide a framework that connects the driver application to the processing module of the POs.

Once a computational model is selected, the computation is performed through the polymorphic operators. The computational model itself can provide rules that define the relationships among the verified objects and among their corresponding semantic groups. The POs execute the rules onto the object representations that match the semantic groups.

The operands are organized in a hierarchical semantic data structure based on an abstraction such as a hash table, a linked list, or a tree. The expert user can construct the hierarchical data structure. Each operand is treated as a node for a semantic group, a group of semantically relevant concepts, in the hierarchical data structure. In case of a tree, the node may or may not have the children nodes. Based on the domain knowledge, an expert can assign a keyword to each semantic group that represents the key idea of the semantic group.

The semantic data structure, i.e. semantic table or semantic tree, is also an abstraction that is used as a basis for defining the relationship in the rules. In case of a semantic data structure based on a hash table, a mapping function, i.e. as a hash function, can be used to match the
keyword pattern in the first level of the data structure. The keyword matching can be performed in the next level of the data structure, i.e. a sub-structure. In a tree data structure, each node represents a semantic group. The keyword pattern matching is primarily processed at the semantic group level not the individual semantic level. However, the keyword pattern matching can be performed in the individual semantic level in the sub-structure. It is similar to a process of finding a “representative keyword” in the disjoint set where the “target keyword” belongs.

As discussed, one exemplary data structure for each semantic group node is a dynamic disjoint set, where the dynamic disjoint set is a member of a collection of dynamic disjoint sets in the disjoint set data structure. When the sub-structures are grouped or ungrouped according to the disjoint set data structure theory, set theoretic operators can be utilized in the polymorphic operator definition. For example, a set theoretic union operator is used for grouping multiple nodes or sub-structures in the semantic data structure. Likewise, a set difference operator is used for separating a sub-structure out of a larger structure by removing an intersection. In a divide and conquer algorithm-based approach for reasoning complex semantics, the polymorphic operators can be first applied to sub-structures of the semantic data structure.

In the Fig. 4.7 that shows an exemplary rule for the polymorphic operator based on the Boolean logic, the rule can be applied to the relationship-based reasoning process. In the exemplary rule, the basic structure of the reasoning of the semantics is formed based on a conditional statement. The “if” segment not only holds unary/binary Boolean logic, but also any number of operands in the Boolean logic. In simple cases, basic Boolean logic relationships are used for the reasoning, such as AND, OR, NOT, XOR, NAND, NOR, and NESTED_IF. We can construct highly complicated logic through the composite of the basic Boolean logic and nested “if” statements. In the Fig. 4.7, the NESTED_IF allows one-step deeper reasoning, and further recursive processes in deeper levels are avoided in the example. Evaluation of each entity at the semantic group level of the objects is often decided as existential quantification. This can be especially true for the literal entity.

The structure of the polymorphic operator is constructed in such a way that the PO can seamlessly follow the predefined rules, such as those in Fig. 4.7. For example, each rule in Fig. 4.7 has an attribute called “type”. The attribute will provide an operation information by matching the [Boolean logic] in the “IF” phrase of the exemplary PO in Fig. 4.8. The first “result” element within a rule will provide an instruction for the “THEN” processing portion, and
the second “result” element within the rule will provide an instruction for the “ELSE” processing portion.

Fig. 4.9 shows the architecture of how the polymorphic operator applies a rule to verified objects at the semantic group level of the object.

Fig. 4.9 Application of the rule to verified objects by the polymorphic operator at the semantic group level
One example that shows how the POs, rules, and semantic groups work can be found in a speech recognition process. In general, the number of vocabularies for the speech recognition used in the interaction is restricted, based on the expert knowledge in the domain. It will be meaningless to issue arbitrary commands in the interaction between users and a multimodal HCI system since the system will not understand the arbitrary commands. In addition to this, the environment, where the exemplary information processing system for the speech recognition is used, is often domain specific, in which only a restricted number of vocabularies are most relevant to the application. When the object representations of verified objects contain these vocabularies, they can be matched to the keywords in semantic groups.

In Fig. 4.9, an exemplary PO, such as the “pseudo example of the polymorphic operator” in Fig. 4.8 can apply exemplary rules in an XML file, such as “an exemplary rule for the polymorphic operator” in Fig. 4.7, to the exemplary verified objects, “verified object A” and the “verified object B”. In order for the POs to apply the rules in an XML file, the operators first match the object representation, such as the “object representation A” and “object representation B”, to the keywords, such as "keyword for semantic group 1" and "keyword for semantic group 2", in the semantic groups by a pattern matching process. Then, they perform relationship-based reasoning based on the “semantic group level relationship”. The “result process” is the result of applying the relationship in the rules, and it can be a simple result or lead to a further process as defined in the rule.

The pattern matching is a mapping process of the keywords to semantics (partially based on concepts in discrete mathematics). Some exemplary types of pattern matching can be as follows:

- Many to 1 mapping: different keywords are mapped into a single meaning.
- Hierarchical mapping: same keywords have different meaning, depending on the level where the keywords are used in the hierarchy.
- Chronological mapping: previous mapping influences on the current mapping.

In this computational model, the process going from “pattern matching” to “relationship-based reasoning” facilitates the fusion of multiple input modalities, whereas the other way facilitates the feature level recognition processes, which is explained in the chapter 5.
4.5.2 Challenges with Polymorphic Operator

The operands of PO can be virtually any type of objects as long as they exist in a digital data format. Although the PO does not restrict the operands in any specific type, it knows how to recognize the operands and process them according to the predefined rules for a particular task since the object verification process verifies the information units into recognizable verified objects for the task. However, it is possible that logical error and complexity can be caused when the OVEN allows the PO to have highly heterogeneous types of objects as operands.

An example of incorrect operation that is caused by semantic ambiguity is given as follows:

Let us assume a goal of an interaction using a multilingual multimodal HCI system is to find a possible emergency shelter in a flooded area, and the multimodal HCI system tries to examine whether a particular building could be used as the possible emergency shelter by applying a PO. For simplicity, let us also assume that the PO is defined to produce a positive return if the object is a ‘building’ and ‘safe’ from the flood damage. If 1) one object instantiated from a certain event, such as an event of sending a query into a database of the buildings in the area, related to the building has an object representation of ‘building’, and 2) the other object instantiated from another event related to the building has an object representation of ‘safe’, then this building could be considered as a candidate for the emergency shelter that can be used temporarily in the flooded area according to the PO.

However, let us further assume that there is another building, for which one object representation is ‘building’ and another object representation is ‘safe’, but the type and meaning of the ‘safe’ in the object representation is totally different from the type and meaning of the previous ‘safe’. For example, the meaning of the ‘safe’ for the second building may be defined to indicate the investment to the building is estimated to be at a safe level by a realtor. In this case, the application of PO will return an unintended result.

Therefore, it is important to construct the PO with precise definition so that the PO is associated with intended verified objects. Additionally, a check-up process may be needed for monitoring the progress of the PO application.
4.6 Discussion

It is important to note that the fundamental ideas in both object-orientation and object-verification focus on the object. However, the object verification specifically focuses on the semantic aspect of the object, where the semantic aspect is verified by the human users as the targeted meaning that the object is intended to represent for a task. This information augmentation gives a direction to each verified object how the information unit needs to be interpreted in a given context. In the adaptation of the OVEN, this dissertation discusses a framework that enables the object verification.

Although there are other ways the object verification in the OVEN may be utilized, one of the significances of this dissertation lies in the fact that the proposed approach suggests a novel and unified way to apply the entire set of object verification ideas, algorithmic approaches, and computational structures in an information processing system like the multilingual multi-user multimodal HCI system.

One of the roles of the OVEN is to provide a tool for defining semantics to each event encountered in an interaction with a multimodal HCI system. With the growing complexity of the web environments, the need for a tool that assigns appropriate semantics to each object is also increasing. Usage of such a tool will enable the expert user to construct a well-organized data structure that will eventually expedite the information processing.

The information sources or flows in an information processing system are not necessarily context dependent. In the case of processing contextual information, the OVEN does not have any assumption that the information comes from only human users. Thus, the concept of ‘intention’ or ‘attention’ is not treated differently from other contextual information in the OVEN.

Dealing with the data representation and handling of the verified objects in the distributed network becomes a challenge when the amount of data is huge or the nature of the data is complex. A multimodal HCI interface can greatly ease the process, by providing additional modalities for interpreting the objects. For example, a voice command, “Show me the result of applying this (e.g. a polymorphic operator) from here (e.g. a verified object class A1) to there (e.g. another verified object class A8),”, along with gestures pointing to the object class A1 representation and object class A8 representation, is more efficient than using a keyboard or
mouse, in getting the same result by directly accessing the related objects in a more descriptive way. The multimodal HCI system parses and understands the descriptive commands automatically and processes the sequence of information in a much faster way, which would have taken more cumbersome intervening steps from the user if it were not facilitated by the multimodal HCI system.

In general, there are several issues that need further research and consideration in the concepts of the OVEN and PO, which are listed as follows:

- The object verification can take a longer time than it is necessary. In order for the application of the OVEN to be successful in a system, an agreement and control of the resources, i.e. object verification time and memory space, needs to be established among the collaborators.
- The depth of rule application by the PO needs to be synchronized with the depth of the semantic data structure. Therefore, a method for communicating between a set of POs and a semantic data structure needs to be established.
- The level of complexity for the rules and the semantic data structure depends on the expert users’ knowledge and skill in the domain. Therefore, coordinated efforts are needed to check the quality of the rules and the semantic data structure, while keeping a track of the hierarchy of the rules.
- The OVEN allows the PO to have heterogeneous types of objects as operands although they may not be comparable among themselves. Although this is one of the most powerful strengths in the OVEN, the operation might appear illogical especially to the users who were not informed to the intention of using the heterogeneous types of objects as operands for the computation in prior. Therefore, coherence of logical or illogical operation in the PO needs to be commonly understood among the collaborators.
- When there is a conflict among the rules for defining the relationship, a strategy for resolving the conflict needs to be applied. The strategy will require collaborative work for an agreement among the participants in collaboration in prior.
- When there are multiple rules, in which one rule is a subset of another rule, a decision needs to be made for which level of rule among the rules should be applied.
Chapter 5
Information Processing in a Multilingual Multimodal HCI System

5.1 Overview

An information processing system is most prominently embodied in a computer-based system in the modern era. The amount and type of information that a typical information processing system deals with is overwhelming. Human errors that are often introduced in interacting with such information processing systems almost certainly warrant a need for an intelligent approach to reduce the errors as much as possible. Physical and environmental conditions in crisis situations are also other factors that contribute to the increasing need for an intelligent information processing approach. These are also the same obstacles that a multilingual multimodal HCI system faces as an information processing system.

In this chapter, an innovative method for information processing is introduced. The approach includes a usage of an information augmentation process, called object verification, and its framework, called object verification enabled network (OVEN). The object verification approach can be applied to various parts of an information processing system. Particularly, it is used for practically improving the process of achieving the goals that are originally assigned to the parts of the information processing system.

As an exemplary application of this approach, this chapter also introduces a concept of “object verified gesture recognition” (OVGR). The OVGR is an approach to utilize the object verification in the OVEN for the image-based gesture recognition. The gesture recognition is an important input modality for a multimodal human-computer interaction (HCI) system. Therefore, this chapter discusses the OVGR in relation to a multilingual multimodal HCI system.

This chapter also discusses a model for applying ‘contextual information’ to a multimodal HCI system, as a specific method in the steps for utilizing the OVGR to the improvement of the gesture recognition. The application of contextual information provides a capability to utilize a
high level understanding of the user-interaction circumstances for the gesture recognition. The exemplary approach will be mainly discussed from the theoretical point of view.

Therefore, the goal of this chapter is to understand how an approach for a structured information augmentation and operations with the augmented information can help in a traditional information processing system, in which the processes of gesture recognition is facilitated by the object verification.

Although the method covers multiple aspects of applying contextual information within a multimodal HCI system, the focus is given to the hand gesture recognition input modality. Hand gesture recognition is one of the most important input modalities in a multimodal HCI system. Its importance especially increases in an actual real-world multimodal HCI system. Research for more robust hand gesture recognition has been focused mainly on finding new or better numerical algorithms for recognizing gestures by the hand images captured through cameras Kettebekov, et al. [63], or using smart cameras, Ham, et al. [50].

In addition to the effort for better numerical algorithms, a few previous attempts for utilizing contextual information for the gesture recognition have been shown, Sage, et al. [113], Hofemann, et al. [53], Kuniyoshi, et al. [71], Iwai, et al. [57], although there are not many.

Our method for applying contextual information in a multimodal HCI system is based on the previously discussed independent heuristic information processing framework; the object verification enabled network (OVEN).


According to the “Merriam-Webster online dictionary”, the definition of context includes:

1 : the parts of a discourse that surround a word or passage and can throw light on its meaning
2 : the interrelated conditions in which something exists or occurs.
Abowd, et al. [1] suggested that the “five W’s” of context are a good minimal set to understand the context. More importantly, Abowd, et al. [1] pointed out that the importance of the context representation.

Due to the various aspects for the meaning of the context, the definition of the term “context” in this dissertation is defined as any quantifiable and non-quantifiable information that characterizes an information object and influences the way the object is understood and applied to a particular multimodal HCI system, in a specific domain.

There is no restriction on which aspects of the context are used for the object verification process in our system, although the exemplary model shown later in this chapter is mainly related to ‘geo-spatial’, ‘geometric’, ‘temporal’, ‘relational’, and ‘history of interaction based’ aspects of the context.

The definition of the context in this dissertation is further limited to the situational information in a state diagram of the user interaction with a HCI system, where the situational information comprises following characteristics:

- **Locality:** The contextual information is affected more by local information than a global environment so that the situational changes at a local state dominate the contextual information at that specific local state.
- **Dynamicity:** The context is dynamic in the sense that the contextual information changes according to the situation at a certain location at a specific time rather than having a static value governed by a global environment. Therefore, the contextual information can be randomly changed depending on the situation and the result of the user interaction, given the almost equal probability distribution for the next interaction choices.
- **Practicality:** The context is preferably represented by a set of processing information and events rather than a numerical parameter. Therefore, the practical usage of the contextual information in this dissertation is achieved by a software module that converts the processing information and events into a message data structure.

Fig. 5.1 shows an exemplary state diagram for the interactions with application contents in a multimodal HCI System. In the example, the contextual information means the situational information at the current state at the current time. For example, the context is defined based on
the interaction history, i.e. the transition history among the state nodes at the time of \( t_i \).

As discussed above, the contextual information changes dynamically depending on the state where the user interaction stage is located. Therefore, the contextual information at \( t_{i-1} \) can be different from that of at \( t_i \). The contextual information in an actual HCI system can be gathered and used to formulate a message file according to a data structure. Then, the message file is passed around among the processes that require the usage of the contextual information.

Fig. 5.1 An exemplary state diagram for the interactions with application contents in a multimodal HCI system

Typically, an application of the contextual information to hand gesture recognition process has been introduced either in low level feature recognition steps or in a high level semantics related processing steps in an interaction model, such as a discourse model, for a multimodal HCI system, Cai, et al. [22], Chai, et al. [24, 25]. Our approach is related to both levels, and it tries to handle the information flow between the two levels.
The connection of the two levels is typically possible through interconnected communication channels among several different possible sub-modules in a multimodal HCI system. For example, an application of context to the gesture recognition at the feature recognition, such as color-based and motion-based detection and tracking, can contribute to the semantics in the high level interpretation of the user gesture through the interconnected communication.

When multiple users use a multimodal HCI system, the gesture recognition in the collaborative interaction paradigm requires more complex handling of the information flow. One important fact to understand in this situation is that no matter how good the feature recognition algorithms are for a single user, the performance of the gesture recognition for the single user may not necessarily be translated to the performance level for the multiple users’ gesture recognition. Interference among the hand tracks in the images for the multiple gesture recognition is one of the reasons why the multiple user gesture recognition becomes more complicated. Increased complexity in detecting and reducing the image noise in the image processing is another reason.

This remains true regardless of the geospatial location of the collaborative users’ existence, whether they stay together in a confined space and interact with a HCI system locally or whether they collaborate through a communication tool, such as a video conferencing, and interact with a HCI system using a network remotely. Thus, the discussion in this chapter also briefly includes how the proposed approach can be deployed in situations for multi-user collaboration and image-based gesture recognition.

5.2 Relevant Work

In the following previous work that suggested methods of using the contextual information in gesture recognition, the first three papers show how context is used in 1) Hidden Markov Model (HMM), 2) condensation algorithm, and 3) neural network. These three algorithms are well-known approaches for the gesture recognition. Later papers show how context is used for a higher level of gesture recognition processing in a multimodal HCI system.

Sage, et al. [113] introduced methods for building cognitive vision systems to understand activities of expert operators. Their approach for the gesture recognition is to learn the generic
models and develop methods for contextual bias of the visual interpretation with the Hidden Markov Model (HMM). A computational model for the Hidden Markov Model (HMM) and its extensions to allow both top-down bias in the contextual processing and bottom-up augmentation by moment-to-moment observation of the hand trajectory is introduced. Particularly, they used a “context control variable” in the HMM training process to apply the contextual information to a system.

Hofemann, et al. [53] showed an approach for recognizing deictic gestures on mobile robot by analyzing the image data from the robot camera. Their approach is to extend a particle filtering based trajectory recognition algorithm for the hand gesture with symbolic information, i.e. contextual information. The contextual information is taken from the objects in the vicinity of the hand as a spatial or symbolic context and incorporated in the condensation algorithm. A “context factor” is introduced in their paper, which depends on whether an object, previously bounded with a motion, is present in the context area or not. Like the “context control variable” in Sage, et al. [113], Hofemann, et al. [53] used the “context factor” to apply the contextual information to their system.

Kuniyoshi, et al. [71] introduced an artificial neural network model for visual recognition of actions. Their model learns to recognize object-directed actions with causal chains of events. An example of the object-directed actions is as follows: “He threw the ball at the window and broke it”, which is also called as “true” action. In their approach, a network, called spatial relationship network, learns and recognizes “spatial relationship”, such as contact/non-contact and movement towards another object. Then, another network, called movement pattern network, learns and recognizes “movement patterns”, such as classes of trajectory patterns. Their output vectors are integrated by the third network, called “temporal context” network, which combines the two symbol vectors in time and detects specific temporal sequences corresponding to an action class.

While Sage, et al. [113], Hofemann, et al. [53], and Kuniyoshi, et al. [71] introduced the approaches to represent the contextual information through parametric variables or symbol vectors, Chai, et al. [24, 25] proposed a system, which handles higher level contextual information from the discourse level with their semantics-based modeling scheme and a context-based approach for understanding ambiguous inputs from the user. Chai, et al. [24, 25] used a variety of contexts, such as domain context, conversation (history) context, user model context, environment model context, linguistics context, gesture context, and visual context to enhance
multimodal fusion. Unlike most previous work, e.g. Wu, et al. [149], on multimodal interpretation, which focus on interpreting user inputs through modality integration, Chai, et al. [24, 25] applied a context-based approach that uses a variety of contexts to enhance multimodal fusion. The goal was to reduce ambiguities and incomplete understanding of inputs in a multimodal HCI system.

The processes of the “Multimodal Interpretation for Natural Dialog” (MIND) in Chai, et al. [24, 25] are divided into three major understanding modules, such as uni-modal understanding, multimodal understanding, and discourse understanding. Their focus is given to the multimodal understanding, and it is further divided into two major techniques, semantic model and integrated interpretation. The semantics-based modeling in MIND deals with modeling user inputs, such as precise meaning of user inputs and the user’s input styles, and modeling discourse-level semantics, such as the entire progress of a conversation to provide the conversation context, based on Grosz, et al. [49]’s conversation theory. Their context-based multimodal understanding is based on these semantic models.

Heidemann, et al. [52], Strobel, et al. [124], Kolesnik, et al. [68], De Angeli, et al. [32, 33], Iwai, et al. [57], Sherrah, et al. [121], Kramer, et al. [70], Munk, et al. [97], and Wolff, et al. [148], also discussed issues in incorporating the contextual information with regard to the gesture recognition. Iwai, et al. [57] showed another example of using the contextual information in a HMM model, in which the dependencies on gesture model at different times are regarded as the context information. Here, the concept of contextual information is too simple to be applied to a higher level multi-modal HCI system. Kolesnik, et al. [68] briefly mentioned a simple and general usage of the contextual knowledge for the hand pointing gesture as a parameterized constraint. Munk, et al. [97] used temporal context to find a way of improving the conditions for robust tracking in addition to a priori knowledge.

The applications of the context in gesture recognition by Sage, et al. [113], Hofemann, et al. [53], Kuniyoshi, et al. [71], Iwai, et al. [57], Kolesnik, et al. [68], Munk, et al. [97] were proposed in a very specific low-level feature recognition processes. The processes may not utilize other available contextual information in the target environment and system as it was shown in Chai, et al. [24, 25].

On the other hand, the applications of the context in Strobel, et al. [124], De Angeli, et al. [32, 33], and Sherrah, et al. [121], were proposed at a very high-level concept, in which their
descriptions about the application often lack necessary details for designing and implementing a concrete system in real-world environment. Their works were based on just an abstract level of simulative experiment, such as the Wizard of Oz experimentation. Kramer, et al. [70] shows a good exemplary categorization of using context in interface design in several processing steps, but it suffers a lack of practical implementation details also.

From the architecture point of view for the multimodal HCI system, the framework in Flippo, et al. [42] is partially related to the system architecture proposed in this dissertation. Another work for rapid system development can also be found in Araki, et al. [9], Barricelli, et al. [12], Elting, et al. [40]. Flippo, et al. [42] suggested a multimodal framework for rapid development of multimodal interfaces, by using an application-independent fusion technique, based on the idea that a large part of the code in a multimodal system can be reused. The key idea for the application-independent fusion technique is to have a separation of three tasks in the multimodal system design. The three separated tasks are 1) obtaining data from modalities, 2) fusing the data to come to an unambiguous meaning, and 3) calling application code to take an action based on that meaning. In Flippo, et al. [42], a semantic parse tree with time stamps was used for the fusion. The semantic parse tree maps natural language concepts to application concepts, and ambiguity resolving agents used contextual information from a context provider in the mapping process. The context provider was either from an external sensor, such as a gaze tracker, or from more abstract data sources, such as dialog history.

Flippo, et al. [42] is foreign to the concepts of augmenting information units in an information processing system through a structured information augmentation method. From a viewpoint, the proposed information augmentation framework can enhance the performance of an application-independent architecture for a multimodal HCI system, like the approach in Flippo, et al. [42]. For example, the component of fusing the data for an unambiguous meaning can utilize the verification technique to disambiguate the meaning of data in a deterministic way by verifying how the data will be used in different contexts. The ambiguity resolving agents that use the contextual information can also be formulated based on an application of the structured information augmentation and polymorphic processing structures. In this adaptation, the reusable part of the code can be programmed to verify the objects automatically since the objects used in the reusable part of code are likely to be known in prior and behave in predictable patterns.
5.3 Context based Reasoning in a Multimodal HCI System

A multimodal HCI system is generally known as an efficient interaction system because it enables a user to use a combination of not only conventional interfaces, such as a keyboard or a mouse, but also other types of interfaces, such as speech or gesture based input modalities, for interacting with the system. Fig. 5.2 shows an exemplary multimodal HCI system, in which a single user typically uses a speech input or a touch free hand gesture in order to interact with the system.

For the last decades, many researchers in the multimodal HCI human computer interaction (HCI) community have been trying to come up with applicable multimodal HCI technologies for the real world environment. There has been much advancement in the technologies, and a few prototypes of commercial real world multimodal HCI applications were shown.

Fig. 5.2 An exemplary multimodal HCI system (Courtesy of VideoMining Corporation)
The realization of such multimodal interfaces is quite different from that of the conventional interfaces. Unlike the keyboard or mouse that transforms a human hand’s touch, push, or various tactile movements into electrical signals through the physical devices, the input modalities by natural speech or gesture recognition require much more complicated processing to interpret the user’s intention. This is a reason why there are not many existing multimodal interfaces today.

For example, in a particular embodiment of a multimodal HCI system, the gesture recognition generally requires multiple intermediate steps before the system eventually extracts meaningful information from the user’s touch free hand gestures. First, one or a plurality of cameras captures the hand gesture visual signals and a frame grabber digitizes the images. The intermediate steps comprise face or hand detection, face or hand tracking, i.e., a trajectory of the hand movement, and steps for deriving meaningful gestures from the given tracking information. The steps further include steps for utilizing available contextual information from various points, such as geometric relation among target objects, temporal information, environmental conditions, and conversation/interaction history.

In order to recognize a gesture, a body part like a hand or a head that is used for the interaction needs to be detected first. A feature based approach, such as skin-color or intensity difference of the target object in captured images, is one of the most popular human body part detection approaches. A model based approach, using a template or prior knowledge of the shape for the target object, is also one of the most popular algorithms for the body part detection. A rigid model or a deformable template (especially for 3D hand model) is used in the model based approach. Other approaches for the body part detection include a facial feature based approach, a learning machine based approach, such as using a neural network or a support vector machine (SVM), and a genetic algorithm approach. Details of various face detection approaches that can be considered in relation to the head detection can be found in Yang, et al. [151].

The feature-based approach has also been used for tracking the human body parts after the detection process. The delay and jitteriness in a hand tracking can be alleviated by the help of an efficient filtering algorithm, such as the Kalman filter.

Using the information from the detection and tracking, the gesture recognition process finds meaningful information for the user interaction. This is usually performed by analyzing a sequence of coordinates tracked in input images as an isolated unit of gesture and mapping the group of coordinates to a target gesture in a set of predefined gestures based on predefined
decision criteria for the target gesture. Stochastic model based approach, such as a Hidden Markov Model (HMM), is one of the most broadly used hand gesture recognition approaches.

These approaches have been studied for many decades and have made some progresses in improving the feasibility for operating in a controlled environment. Unfortunately, none of the known hand tracking and gesture recognition approaches is reported to be robust enough to make a multimodal HCI system to be truly usable by general public users in an uncontrolled real world environment. Based on years of empirical studies, we have observed there is a discrepancy between what was reported from the laboratories and the actual results a system produces in the real-world environment, including unexpected abnormalities in the gesture recognition.

One of the reasons why the discrepancy exists is based on the fact that the real world environment itself poses non-trivial challenges to the multimodal HCI system when the system is deployed to the real world environment. The obvious challenges often invalidate fundamental assumptions that are established at the system design stage under a well-controlled laboratory type environment. Even a slight variation caused by the challenges in the middle of executing the internal steps inside the already highly complicated system eventually leads to an unexpected and unpredictable abnormal results. For example, this includes the well-known jitteriness or jumpiness problem in the hand movement trajectories caused by an arbitrary interference of surrounding objects.

Gesture recognition and hand detection/tracking technologies in a multimodal HCI system that aims to be usable in a real world environment need to overcome the following challenges:

- **Image Quality:** The system should be able to capture and process the input image sequences at a reasonable frame rate in order to handle the real-time interaction and information processing, e.g. 15~30 frames per second (fps), without losing the image details. This is related to the issues of speed for processing the imagery data vs. size of the imagery data per frame.

- **Automatic Resolution Adaptation:** Automatic resolution adaptation is sometimes needed due to the imprecise location of the user stance in front of the system or movement of the target object in the field of view. For example, a user might stand too close or too far with regard to the proximity to the designated optimal system interaction space.
• No Static Threshold: Ideally, no threshold or no fixed size template should be used for executing the instructions and image processing algorithms in the system architecture. This is to make the system flexible enough for the environmental and user specific variations. In practice, this is highly difficult to achieve because many algorithms depend on a predefined thresholds or templates. Therefore, it is aimed to reduce the usage of fixed size thresholds and templates. Adaptive algorithms can be used for this goal.

• Automatic Initialization: Automatic initialization does not require user’s manual or interactive involvement to initiate the system and its fundamental algorithms. In general, users do not like to be involved in the system initialization process. It requires extra time, learning period, and cumbersome steps for the users to participate in the initialization, without the automatic initialization function. Therefore, this is a necessary element in the system design in a real world environment.

• Tolerance to Environment: The system should be tolerant to environmental variations, such as intrusion of unexpected object, image noise, and ambient lighting variation in the environment. For example, if the color of a floor mat is similar to the skin color, within the field of view, the skin color based hand detection and tracking algorithm experiences a greater chance of failure. The system should detect the situational phenomenon if the performance of the detection and tracking degenerates, while applying the hand detection algorithm to the images. However, if an environmental variation exceeds the limit the tolerance level of the system, the system should provide an alternative solution that minimizes the surprise and frustration the user might face due to the system failure.

• Independence: The system should avoid too much dependence on specific user characteristics or demographics in the geographical region, such as different skin color, height, or shape variation. For example, average skin color templates for a population in a region may not be applicable to a group of people in the region or another population in another region. Therefore, any assumption that is too much dependent on user characteristics, i.e. average skin color in this example, for the hand shape and movement should be avoided. Similarly, any assumption that is too much dependent on other demographic attributes in the geographical region should be avoided.

• Easy On-demand Training: Cumbersome training requirement for using the system should be avoided or minimized. This goal is also closely related to the issues how to
design the application contents and the transition of the contents. For example, minimal training contents can be embedded in the main application contents seamlessly with minimal interruption. The training contents can also be made to be revisited from another stage of application contents when a user needs them by making a transition to them. The training step should also be skipped if the user elects to do so. Similarly, the secondary contents, such as help menu, should be easily accessible or skipped based on user demand.

One important point to note is that although the usage of static information, such as the geographical specifics, need to be avoided as discussed above, the static information may contribute as contextual information in processing the gesture recognition if an intelligent and structured way to convert the static information to contextual information.

If the goal of a system development and its usage are merely to show study results and concepts of experimentation in a laboratory, many of the requirements mentioned above can be avoided relatively easily, by controlling the situation and incidence of the interaction scenarios. However, if the usage of the multimodal HCI system is intended for a real world environment, the solutions to the issues listed above are non-trivial goals to achieve.

For example, a hand-tracking process can produce incorrect coordinate sequences of the trajectory, partially due to blocking objects, lighting variation, abrupt changes in the image by the auto-exposure or auto-focus function of the camera, or noise on the captured images. In this case, no matter how good the original hand gesture recognition approach is, it is subject to error since the original raw image data itself is damaged from the possible causes. Likewise, with many practical reasons mentioned in the list above, there is a risk that even a very sophisticated algorithm for the gesture recognition could fail. We have implemented many real world multimodal HCI systems, and some of them were used as bases that had been commercially deployed to real world interaction environments based on the work by Sharma, et al. [118, 119]. During the experiences with the systems, the scientific challenges discussed above have been empirically observed in multiple occasions.

As an attempt to reduce the system operation failure and abnormal behaviors, often ad-hoc solutions were introduced. For example, introducing additional parameters, thresholds, extra conditions, or error handling pieces of code, are sometimes used as quick fixes to the challenges discussed above in a real-world environment. As Flippo, et al. [42] pointed out, only quick fixes
are applicable to handle the challenges in this situations in order not to risk the primary performance and fundamental assumptions due to the complexity and rigidity of the multimodal HCI system.

However, the ad-hoc solutions are often non-scalable and incompatible from one system to another. Cumbersome training processes can be requested to the users for deploying the ad-hoc approach. Unfortunately, the training process makes the system highly user-dependent and sensitive to even slight changes among the users and in the environment. Thus, the ad-hoc solutions need to be avoided or minimized in a multimodal HCI system, as mentioned above.

The performance improvement by the temporary quick fixes is also limited in an actual interaction process in the real-world environment. It was observed that the discrepancy between the expected numerical results and the actual performance in the real-world continued to exist even after parameters and variables for the numerical algorithms were refined in the system development. For example, as it is well-known, the Hidden Markov Model (HMM) based gesture recognition requires cost inefficient training processes from each user if the system were to have higher recognition rate due to the conditions in the real world environment. Another example can be found in the speech recognition. The speech recognition technology provides another prominent input modality in a multimodal HCI system. However, it is also well-known to have a limitation in its performance depending on the speakers, not to mention environmental situations, such as noise. A speaker-dependent speech recognition tool is known to have a higher accuracy rate than that of a speaker-independent speech recognition tool. However, it also requires individual user’s voice training process prior to the actual usage of the system in order for the system to be usable with the reasonably acceptable accuracy rate.

The proposed information handling approach provides a robust, scalable, and well-organized framework that overcomes the issues in the ad-hoc solutions and efficiently converts the ad-hoc solutions into a manageable tool that may lead into a feasible solution to handle the requirements for designing a multimodal HCI system discussed above.

In another perspective with regard to the nature of the interaction, the user interaction is often tightly related to the semantics in a certain context at a specific time. Therefore, recognizing a user’s hand trajectory alone may not show the user’s true intention clearly. The single image feature-based gesture recognition approach alone could cause a misinterpretation of the user’s intention as the outcome. This could possibly lead to misrecognition of the gestures in turn. One
of the concepts in the proposed approach is that it can utilize high level information to make an impact on how the feature level algorithms process. For example, a set of contextual information can be embedded into object verifications of the hand tracking coordinates and a group of processing structures can be defined to utilize the contextual information to aid the interpretation of the hand coordinates.

Continuous research efforts are needed for developing better algorithms for maturing the conventional input modality technologies for the gesture and speech recognition as they have been attempted in past decades, Bonnyman, et al. [19]. In addition to these traditional research efforts, an approach for using the available contextual information in the processes for a HCI system has a potential to provide a complimentary aid to solve the above-explained obstacles.

Therefore, it is meaningful to note that the novel approach helps making the HCI system closer to a usable real-world product. The improvement can be made at the following levels of the system:

1) in the feature recognition level, by solving the problems in the feature recognition level, at least partially, and
2) in a higher level processing of the system for handling the semantics, produced within the interaction.

Our approach for applying the contextual information to a multimodal HCI system is embodied in an application independent module, called the “context based reasoning module” (CBRM). The CBRM is connected to the key processing units in a particular multimodal HCI system, such as gesture recognition and speech recognition modules, and communicates with the key processing units through a message passing method. The CBRM comprises its own sub-modules and connected to a set of handlers and the object verification enabled network (OVEN), as shown in Fig. 5.3. Altogether with 1) the CBRM, 2) the set of handlers, and 3) the OVEN, it is called “context based reasoning network” (CBRN).

The sub-modules in the CBRM are composed of reasoning modules for handling reasoning process and user handling modules for handling user interaction and collaboration among users. The set of handlers processes the contextual information from the related processing units in the system. The CBRM also uses a collection of networked modules in the OVEN in order to
provide a computational model for identifying events that occur while processing the tasks and user interaction in the system. The OVEN also provides a model for applying pre-defined operation to the events through the polymorphic operators based on the heuristic rules in the OVEN.

The sub-modules, the set of handlers, and the OVEN communicate among themselves to derive necessary information for assigned tasks. For example, a processing unit of the system reports an event to a reasoning module. Then, the reasoning module communicates with a handler to apply a polymorphic operator to the event in the processing unit of the system.

Due to the nature of the information processed in a multimodal HCI system, a computational model for this approach needs to handle both quantifiable and non-quantifiable contextual information. Quantifiable information is often integrated as parameterized contextual information, but non-quantifiable information needs a different approach. This is one of the key reasons why a structured information augmentation method like the OVEN, is needed in the CBRM for a multimodal HCI system design.

So far, we discussed the research problems from the perspective of a single user multimodal HCI system. When multiple users interact with a HCI system for a collaborative task, detection and tracking of multiple hand trajectories are needed for simultaneous multiple gesture recognition. In a collaborative interaction that requires multiple gesture recognition, the motivation for each gesture by each individual is based on a unique mental status. Intentions and plans for the collaborative interaction among the users become highly intricate in this case. Therefore, a simple feature level interpretation of the gestures may not be sufficient to truly understand the collaborative users’ intentions. For example, when multiple hand trajectories overlap over a set of coordinates in the camera view for a period of time, the low level feature recognition by itself may not provide appropriate interpretation of the gestures why the users’ hand trajectories are collocated at the coordinates during the particular moment. Often it becomes clearer when high level information is available, such as a history of interaction or conversation contents in a dialogue.

The task for detecting and registering hand positions for each user is a non-trivial step for the multi-user gesture recognition. It is a further challenging task to reliably track the registered hands and correctly match the hands with the users. Especially, interference among the hand trajectories further aggravates the difficulty of matching process. For example, overlapped hand
positions by multiple users or cross over of the trajectories often cause the hand trajectories are matched with incorrect users.

As for the speech input modality, it is well-known that the task of distinguishing different voice inputs among multiple users is a very difficult task when the voice signals overlap. Discerning the initiatives among users is also another challenging problem to solve. Communication among multiple users makes the interaction even more complicated in a complex scenario.

Sharma, et al. [118] believes that a probabilistic evaluation of all possible speech gesture combinations promises a better estimation of user’s intent than either modality alone. This belief can be extended to the case of multi-user multimodal HCI systems. When evaluating all possible speech-gesture combinations among multiple users, how the modalities can seamlessly cooperate becomes an interesting point to observe.

The interference in sensing signals among multiple users can occur both at the lowest feature level of the system processing and at the highest semantic level of the user interaction. Regardless how the architecture for sensing users’ input modalities is constructed, the interference phenomenon exists among the signals while using a multi-user multimodal HCI system. Depending on the application scenarios for the interaction, the interference could be alleviated through a step of applying restrictions to the interaction. For example, the step can include a process of executing the gesture recognition in sequence rather than executing it simultaneously. Setting up priorities, particularly for preemptive gestures, in the interaction is another way to alleviate the interference phenomenon. However, this will restrict the way users interact with the system because it forces the system to be less responsive to the gesture in the limited application scenario. As a result, this method also makes the interaction unnatural to the users’ perception. As an exemplary solution to reduce the signal level interference, a registering algorithm can be used as indicated earlier. In this solution, each signature of the signal is registered to each user and matched to the correct user at the execution. We can either dynamically initiate and terminate the registration of the signal, or statically assign the registration.

Another thing to notice in a multi-user multimodal HCI system interaction is that the relationship of the gesture and speech information flow is not one-to-one mapping always. For example, a speech information flow from a ‘user A’ does not necessarily have to match the
speech information flow from another ‘user B’. While the ‘user A’ is dealing with the speech information flow, the corresponding information can be mapped to the gesture information flow of the ‘user B’. Therefore, how to find the meaningful intent of the multi-users within the intermixed relationship of the information flows needs to be carefully considered.

Fig. 5.3 shows an exemplary multi-user multimodal HCI system. Users can have their own visualization means or share a visualization means as shown in FIG. 5.3. In this case, it will require more precise differentiation of not only the speech and gesture input but also feedback of the users’ interaction and visualization of corresponding visual objects among the users.

Fig. 5.3 An application of object verification enabled network (OVEN) in the context based reasoning module (CBRM) for a multi-user multimodal HCI system
As discussed in chapter 4, one of the features in the object verification enabled network (OVEN) is to extend the multimodal HCI paradigm to embrace the domain of web-based applications along with the traditional multimodal HCI applications. The approach uses remote resources in a network environment through a verification process for objects, i.e. information pieces appearing in the network. The OVEN utilizes strengths of the multimodal HCI system in the distributed web environment, such as advantages of accessing to a remote database. Especially, one of its original purposes is to facilitate multiple user interaction with multiple languages in a highly distributed computing environment.

The information sources and their flows in the OVEN are not necessarily context dependent. The OVEN also does not have any assumption, in which the contextual information comes from only human users. Thus, the concept of ‘intention’ or ‘attention’ is not treated any differently from other contextual information. In another words, the OVEN is not necessarily limited to the usage introduced in this chapter. However, the conceptual work in OVEN provides an ideal tool for the computational model of the CBRM, which is used for applying the contextual information in a multimodal HCI system.

5.4 Object Verified Gesture Recognition

The object verified gesture recognition (OVGR) is an approach to utilize the object verification technology in the OVEN for the gesture recognition. In this section, we will discuss the OVGR approach in relation to a module for the application of contextual information, such as the CBRM. Particularly, the two-way roles of CBRM, namely 1) in the fusion process of the multimodal inputs and 2) in the feature recognition are discussed.

The section 5.4.1 discusses an application of the CBRM to a fusion of the multimodal inputs. Previous work for fusion of the multimodality can be found in Nigay, et al. [99]. In the exemplary application of the CBRM, the conceptual part for modeling semantic features in the fusion level process partially uses previous work on collaborative discourse interpretation by Lochbaum, et al. [82, 83], and Grosz, et al. [49].

In our approach, the modeling of semantic features based on the conceptual work by Lochbaum, et al. [82, 83] and Grosz, et al. [49] can be represented in the form of rules in a rule-
base. Then, a polymorphic operator, e.g. the exemplary Boolean logic operator in chapter 4, is used to operate on the verified objects based on the rules.

In the section 5.4.2, the CBRM is applied to the feature recognition processes. In this feature recognition level application of the CBRM, our approach will be a problem domain specific approach for each technology. This approach utilizes feedback for understanding the features in gesture recognition or speech recognition. One of the goals of the utilization is to improve each individual feature recognition performance through a set of predefined software modules, called handlers.

5.4.1 Context based Reasoning Module in Fusion of the Multimodal Input

In this exemplary application of the CBRM to a fusion of the multimodal inputs, the attributes for the discourse interpretation are integrated to the rules in the rule-base for the polymorphic operator, i.e. as a semantic group discussed in chapter 4. The conceptual work for modeling semantic features in this level is partially similar to the previous work on discourse interpretation using attributes for the semantic modeling as in the conceptual work for the “Fusion” in Chai, et al. [24, 25]’s MIND. The approach is different from the “Fusion” in Flippo, et al. [42]’s suggested framework, although the fusion manager in their work can be partially relevant.

One of the differences between the CBRM and the MIND by Chai, et al. [24, 25] can be found in the computational model of the reasoning process. In the CBRM, the computational model for the reasoning process uses a dynamic structure constructed by the information from both 1) the conversation history between the system and the user, and 2) the information from the knowledge base stored in a separate entity of its own. For example, the CBRM communicates with not only the intention, attention, and discourse model, but also the knowledge base of the specific application. How the reasoning module communicates with the knowledge base is mostly subject to the architecture design decision. They can directly communicate with each other or through a dialog manager.

In addition, while Chai, et al. [24, 25]’s initial semantic models and interpretation algorithms were driven by a user study, which leads to a very organized and rigorous set of criteria for the
semantic models, the CBRM does not limit any set of criteria for constructing the semantic models. It is partially due to the fact the CBRM is intended to be used for both conversational and non-conversational contextual information, which may not involve any human discourse, such as using a knowledge based on a remote place as mentioned previously. In this case, the concept of “intention” or “attention” becomes less meaningful. The semantic model construction in the CBRM is based on the object verification, and the definition of the polymorphic operator is based on the defined rules. Therefore, the element in OVEN that works as the computational model for the CBRM does not have any fixed criteria to use for constructing semantic models.

In the exemplary conversational segment of a particular interaction with a multimodal HCI system, shown in Fig. 5.4, the reasoning module partially did not recognize a user’s speech at Time: $t_{k+1}$, e.g. the speech was not defined in the grammar, and it tries to recover the missing part with the help of the reasoning module. Without a reasoning module, the system will, most likely, result in the steps for error handling, such as asking the user to speak again, pausing other processes until the user gives further input command, or providing an alternative instruction for another action. The reasoning module can be structured in such a way that it knows what to do when there is a missing information associated with the keyword ‘highlight’ by looking at its keyword semantic class, contextual type, and format by a pattern matching.
Table 5.1 shows an exemplary structure for semantic grouping of keywords, registered keywords in each semantic group, and the syntactic format. In this example, the deictic gestures are divided into spatial, symbolic, and textual types, depending on the context. The internal process with temporal task in the ‘semantic group 2’ is an example whose contextual reasoning is not directly involved with human conversation. The CBRM can apply the contextual information to any processes in the application as indicated in this example. The speech segment with temporal task in the ‘semantic group 3’ is connected to the associated temporal information rather than a direct temporal keyword or information.

The CBRM looks up the ‘highlight’ in the table of keyword class, and tries to find a match between the speech command and the available format of the keyword. Once there is a match
between them, the reasoning module gives a clue for the missing information. In the exemplary Table 5.1, the available and matching pattern will be

‘Highlight {Spatial Info}’ or ‘Highlight {Symbolic Info}’.

Using the history of the conversation at Time $t_i$, the module extracts the information about the context of the speech command. The module determines the ‘highlight’ command is more appropriate for the format of

‘Highlight {Spatial Info}’,

in this incidence.

The system also determines that the user’s gesture at Time: $t_k$ indicates that the ‘{Spatial Info}’ can be known by the

‘Gesture Type: “Point gesture”, Coordinate(s): $X_k, Y_k$, Time: $t_k$’.

Using these context based information, the reasoning module can propose the missing information as semantically equivalent to ‘this area’ at the position of:

‘Coordinate(s): $X_k, Y_k$, at Time: $t_{k+1} - k+2$’.

The system does not have to know the ‘this area’ speech segment exactly, since

1) the command keyword, ‘highlight’, and the point gesture coordinate, ‘$(X_k, Y_k)$’, are known, and
2) their semantic relationship is defined by the exemplary rule.

At this point, the CBRM notifies the visualization module that takes care of the visualization of the system to show the result at Time: $t_{k+3}$. The CBRM only provides the contextual information to the corresponding visualization module component that handles the visualization.
process. In this exemplary application, the exemplary rules used by the polymorphic operator are as shown in Fig. 5.5.

<table>
<thead>
<tr>
<th>Semantic Class</th>
<th>Registered Keyword</th>
<th>Contextual Type</th>
<th>Syntactic Format</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Semantic Group 1</td>
<td>“Zoom”</td>
<td>Spatial</td>
<td>Keyword (Spatial Info)</td>
<td>“Zoom here”</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>“Zoom this area”</td>
</tr>
<tr>
<td>Semantic Group 2</td>
<td>(Process with</td>
<td>Temporal</td>
<td>(Process) (Temporal Info)</td>
<td>(Point gesture) (At Tn)</td>
</tr>
<tr>
<td></td>
<td>temporal task)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(At time)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Temporal</td>
<td></td>
<td>“How soon can it be done?”</td>
</tr>
<tr>
<td>Semantic Group 4</td>
<td>“Draw”, “Send”,</td>
<td>Spatial Unary</td>
<td>Keyword (Spatial Info)</td>
<td>“Draw a circle”</td>
</tr>
<tr>
<td></td>
<td>“Move”</td>
<td></td>
<td></td>
<td>“Send a fire truck”</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Spatial Binary</td>
<td>Keyword (Spatial Info), (Spatial Info)</td>
<td>“Draw a line from here to</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>there”</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>“Send an ambulance from</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>here to that building”</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>“Move from here to there”</td>
</tr>
<tr>
<td>Semantic Group 5</td>
<td>“Highlight”</td>
<td>Spatial</td>
<td>Keyword (Spatial Info)</td>
<td>“Highlight this area”</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Symbolic</td>
<td>Keyword (Symbolic Info)</td>
<td>“Highlight this”</td>
</tr>
<tr>
<td>Semantic Group 6</td>
<td>“Select”, “Show”</td>
<td>Spatial Unary</td>
<td>Keyword (Spatial Info)</td>
<td>“Select this area”</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>“Show me the area”</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Spatial Binary</td>
<td>Keyword (Spatial Info), (Spatial Info)</td>
<td>“Select from here to there”</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>“Show me from here to there”</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Symbolic</td>
<td>Keyword (Symbolic Info)</td>
<td>“Select the rectangle”,</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>“Select the blue one”</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Textual</td>
<td>Keyword (Textual Info)</td>
<td>“Select this”</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>“Show me this”</td>
</tr>
</tbody>
</table>

Table 5.1 Exemplary classes of keywords for speech inputs in the context based reasoning module.

```xml
...<rule>
  IF 'highlight' [and] 'previous context is spatial'
  THEN 'highlight {Spatial Info}'
  ELSE 'highlight {Symbolic Info}'
</rule>
...
```

```
<rule>
  IF 'highlight {Spatial Info}'; [and] 'point gesture' [and] 'coordinate(s): (X, Y)'
  THEN 'Send a message to gesture recognition' [and] 'highlight coordinate(s): (X, Y)'
      { ELSE else-result }  // optional
</rule>
...```

Fig. 5.5 Exemplary rules for processing a command using spatial/gesture contextual information and conversation contextual information.
5.4.2 Context based Reasoning Module in Feature Recognition

In this theoretical application of the CBRM to a feature recognition process, the key idea is to utilize feedback information from the user interaction and semantics derived from the interaction for the feature level understanding by adjusting the parameters in the individual feature level recognition. The ideas for the conceptual work in this particular step are similar to the “Inference” in Chai, et al. [24, 25]’s MIND and the “Fission” in Flippo, et al. [42]’s suggested framework.

One of the differences over the relevant work is that the CBRM focuses on each problem domain of the sub-technology in the feature recognition, so that the system utilizes the contextual information in a low-level system processing in more detail through a set of software modules, called ‘handlers’. The set of handlers are connected to key feature recognition components, such as gesture recognition or speech recognition. The set of handles also communicate with the components directly and take care of the feature recognition problems in their own problem domain in a more comprehensive way. Fig. 5.11 shows the architecture of the context based reasoning module and an exemplary set of handlers.

Example 1: Resolution Adaptation

In an exemplary situation, the handling of the ‘resolution adaptation’ by the vision handler can automatically signal the system to zoom to the user when the user is far away from the system and the gesture recognition algorithm is having difficulties to track the hand position due to the distance. The CBRM can also detect whether there is a need for more detailed matching between the hand position and the image object on the screen in a finer definition, based on the discourse context.

In an exemplary scenario shown in the Fig. 5.6, the user tries to select an orange object at ‘Time: t_k’ after the system displayed two objects, one with yellow and the other with orange color. However, the ‘Coordinate(s): (X_{k}, Y_{k})’ at ‘Time: t_k’ is not within the boundary of the object’s selectable region, ‘(X_{st-sm}, Y_{st-sn})’. There can be various reasons why this kind of situation could occur, such as jitteriness of the hand movement trajectory or relatively very small
size of the object. Accordingly, the system could not match the ‘coordinate(s): X\textsubscript{k}, Y\textsubscript{k}’ with the correct ‘orange’ object within the boundary of the object location, ‘(X\textsubscript{s1–sm}, Y\textsubscript{s1–sn})’.

In this case, the CBRM can notify the ‘vision handler’ to zoom to the user’s hand by sending a zoom-in command to an active camera or notify the ‘visualization handler’ to enlarge the visualization of the hand gesture area, ‘(X\textsubscript{k}, Y\textsubscript{k})’, and the location of the symbolic object area, ‘(X\textsubscript{s1–sm}, Y\textsubscript{s1–sn})’, on the screen in order to help the matching between a new coordinate of the user’s hand position and the position of the object at ‘Time t\textsubscript{k+2}’.

Fig. 5.6 An exemplary processing sequence and conversational segment in which the vision handler is notified by the CBRM in the feature recognition level for resolution adaptation event.
The CBRM provides the contextual information to the corresponding handler(s) for handling the resolution adaptation. The CBRM model itself does not need to perform the actual processing. As in ‘Error Message: at Time $t_{k+2}$’, the system can give the user appropriate instruction of what to do. The visualization at ‘Time $t_{k+2}$’, does not have any new change from the previous screen, but in the later times, at ‘Time $t_{k+2+\alpha}$’, the notified ‘visualization handler’ will display the enlarged visualization of the corresponding region on the screen.

Fig. 5.7 Exemplary rules for sending messages to handlers based on spatial, gesture-based, symbolic, and conversation contextual information to process a command.
In this exemplary application, the exemplary rules used by the polymorphic operator will be as shown in Fig. 5.7. In order to cover most of the usage cases, it needs to be expanded with more rules in a deeper hierarchical ‘if-then-else’ conditional statement structure. One of the advantages in constructing the rules using the OVEN is that the object verification hides the implementation details behind each verified object. For example, each verified object in the conditional statements is represented just as a string type data. However, the outcome of applying the polymorphic operator provides contextual information to the corresponding handlers or the key components of the multimodal HCI system that perform the actual processing of tasks with the help of the contextual reasoning.

Example 2: Lighting Changes

In another example, the vision handler deals with the issues of the lighting changes in the environment. When the lighting changes in the real world environment, where the multimodal HCI system resides, fundamental algorithms used for the gesture recognition and computer vision image processing often fail. It is because many of the algorithms are based on the assumption that the image signals are captured under relatively constant environmental situation during a period of the time, such as during the interaction time. For example, the skin color based face detection algorithm assumes that the user’s skin color does not change in the middle of interaction. Although an adaptive background subtraction algorithm accommodates the changes in the image over relatively longer period of time, the background subtraction algorithm also often assumes that the intensity in the image do not change at least when the user interacts with the system. However, the assumption could fail in real-world environments, due to various reasons. The lighting could change even in the middle of user interaction. Sudden appearance of other lighting sources, a change of the sunlight direction at different times of the day, or reflections from surrounding objects can be a part of the reasons. Usage of the contextual information could help in this kind of problem by notifying the vision handler to reconfigure the system components. For example, the system can adjust the auto exposure of the camera when a sudden change is detected in the overall intensity values of the processed images in consecutive frames.
Similarly, the handlers can take care of delays in frames and synchronization of the gesture/speech recognition when there is discordance in the conversation. The handlers can be much helped by the semantics from the CBRM to fine-tune the thresholds or templates. In general, it is not ideal to have thresholds for the parameters in an automatic HCI system, as discussed in chapter 5.3. However, if an algorithm in the system cannot avoid the usage of thresholds, automatically getting feedback from the user is one of the best ways to know when to adjust the thresholds. An information to make a decision whether to adjust a threshold or not can be extracted from the conversation or interaction history between the user and the system, using the CBRM. A few preliminary conversation and extracted semantics from the initial conversation with the user can provide useful information for automatically initializing the system when there is a change of users. In addition, handling noise, user independence, shading, and environment control can also be benefited from the CBRM as for the handling of the lighting variation.

The application of the handlers, based on the CBRM, in this section has been mainly focused on improving the gesture recognition. However, the theoretical approach that we discussed for improving the gesture recognition can also be applied to the speech recognition. For example, the feedback with context-based reasoning can be used to switch among multiple voice profiles of the same user, depending on the accuracy rate. When the system detects the accuracy rate of the speech recognition suddenly drops down, the reasoning module can signal the system to change the voice profile of the user with highest probability of increasing the success rate.

5.5 Theoretical Exemplary Experimentation

This section shows how the CBRM can be used for making an improvement for a specific problem in the gesture recognition from the theoretical perspective. Especially, the CBRM collaborate with the OVGR as a complementary approach to achieve the improvement. Jitteriness may be described as a high frequency noise that appears irregularly in the hand trajectory. Reducing the jitteriness is one of the exemplary processes where the semantic reasoning by the CBRM can provide an improvement. It has been observed that jitteriness is caused by one of the innate weaknesses of the hand tracking algorithms developed thus far. For
example, the background subtraction algorithm is one of the key algorithms for segmenting the hand object from other peripheral objects in input images. The background subtraction algorithm is highly sensitive to lighting variation and noise, even without obstacles in the field of view. An adaptive background subtraction algorithm may not be suitable for the real-time hand tracking interaction since it is very difficult to synchronize the adaptation process with the real-time hand tracking.

In addition to this, the human hand constantly moves within the interaction space that can be virtually defined in the 3D physical space. The constant movement includes undesired or unintended movement, such as a slight tremor by the physical phenomenon of the human hand. The visual sensor, i.e. a camera sensor, picks up the movement including the tremor and transfers the signals to the CPU directly as the input, without a differentiation between the intended movement and the unintended movement. This eventually contributes to the jitteriness phenomenon. Filtering algorithms that are applied to the sequence of the coordinates of the hand positions could reduce the jitteriness to a certain degree. However, the filtering technique also has limitations in reducing the jitteriness. For example, the filtering approach is highly dependent on the assumption that only a part of the entire trajectory is an anomaly. Discerning which are normal coordinate segments and which are abnormal coordinate segments from the entire trajectory is not obvious. If the sequence of the coordinates contains a highly irregular sequence for longer than a period of time that the system can handle, it is challenging for the filtering process to reduce the jitteriness. Fundamentally, the root problem is that the coordinates themselves, without higher level contextual information, do not provide enough information for the system to fully handle the unexpected sequence of numerical values. Therefore, we discuss utilization of the contextual information based on the semantic reasoning by the CBRM to reduce the jitteriness in hand trajectories, as a novel and complementary approach to the traditional filtering algorithm in this section.

For real world hand tracking, the system processes a real-time dynamic trajectory of a user’s hand in captured images. The term ‘real-time’ actually means about 1/30 ~ 1/10 seconds of delay from the time when the hand image is captured, although it depends on the variability of the system hardware configuration. The term ‘dynamic’ means that it is not perfectly predictable to locate where the next position of the hand will be in the next a couple of image frames in prior. This prevents any assumptions about the sequence of coordinates in the trajectory. For example,
we cannot assume the next position of the hand will be in the same direction as it is in the previous image frame since the direction of the hand movement can be reversed or paused.

The approach in the CBRM is to use a reasoning process in order to improve the hand tracking and coordinate smoothing. Using the contextual information, the CBRM is able to give higher probability to a candidate position in the next estimation of the hand position. In this process, the reasoning module takes the ownership of the pointer on the screen and consults with the known information gathered by the system at the specific time of incidence. For example, a piece of information from the speech recognition or internal process that matches the information with a knowledge base, such as a GIS database, is forwarded to the reasoning module’s analysis and decision making process.

5.5.1 Ownership could belong to the Mice not Human Hand

A mouse, as an input peripheral device to a computer, is one of the most popular interaction modalities in contemporary human computer interaction. It is worthwhile to observe differences between the user interaction using the mouse and the user interaction using the natural hand gesture. Understanding the differences helps us to redefine the ownership of the cursor on a monitor that is a peripheral output device for a computer.

In the mouse-based interaction paradigm, it needs to be understood that the ownership of the cursor actually belongs to a mouse not the user although the user has an illusion of owning the cursor on the monitor screen. Sometimes, users experience irregular or unintended movements or stoppages of the cursor on the screen while using a mouse. For example, often, an optical mouse on a surface that reflects the light emitted from the mouse itself does not work properly. None the less, it can be observed that the jitteriness of the cursor trajectory on the monitor screen by the mouse occurs significantly less than that of the gesture recognition trajectory by a natural hand gesture. Whether it is an optical mouse or a mechanical mouse, the pattern of movement is detected by the signal processor of the optical mouse or the ball mouse, and the signal from the device is transferred to the visual representation, i.e. cursor, on the screen. In this process, the unintended movement or the tremor by a human hand is naturally and physically filtered. Unless the user tries to cause the jitteriness intentionally by moving the mouse in a shaky fashion, it is
difficult to simulate the jitteriness in a mouse-based interaction. It is because the mouse device stays and moves on a stable physical surface, and the signal generation is not triggered by the hand movement directly. Thus, the device itself is stabilized physically, which results in the filtering of human hand’s unintended movement and tremor.

Therefore, it can be regarded that the user’s hand movement is just an action for initiating the orientation and degree of the movement by giving the inputs to the mouse as a form of physical movement. The mouse is the one that actually controls the cursor on the screen.

In the natural hand gesture based interaction, the hand trajectory is somewhat directly attached to the cursor movement on the screen without a physical filtering process. The intensity of the hand object in the input image changes from a frame to the next. This may be related to the fact that the camera sensor senses the hand image as it is represented in the physical space based on the light reflection, wherein the human hand constantly moves in the physical space. In addition, other factors like ambient light, shade, and interference by adjacent object also contribute to the changes of the intensity in the images. As discussed above, this is one of the reasons why the problem of jitteriness occurs in the hand tracking. Therefore, it can be considered to transfer the ownership of the cursor in the natural hand tracking based interaction from the human hand to an intermediate medium. An intermediate agent can play a role for the filtering process as the physical constraint does it for a mouse device. Although an application of a numerical function-based filtering process to the sequence of the coordinates calculated from captured images is feasible as a software-based intermediate agent, this method alone will suffer from the lack of stability due to the reasons discussed above. Therefore, additional application of contextual information can be useful as a complementary approach that enhances the filtering process.

5.5.2 Interpreted Hand Movement by the Semantic Reasoning of the Context based Reasoning Module

In this section, a semantic reasoning process in the CBRM is applied to the interpretation of hand movement as a concept of viscosity and guided movement. This represents the switch of ownership of the cursor from the human hand to the CBRM. Therefore, the cursor-handling
agent in the CBRM is an exemplary intermediate agent discussed above. The cursor-handling agent interprets the hand trajectory information as an action for initiating the orientation and setting a degree of the next movement rather than directly responding to the actual hand movement coordinates.

The trajectory of hand movement can be divided into

1) *meaningful movement trajectory*,
2) *transitional movement trajectory*, and
3) *arbitrary movement trajectory*.

Meaningful movements are associated with the user’s intention and command action. Examples of meaningful movements comprise point, linear, curvature, and circle gestures. Transitional movements are the movements between meaningful movements and any type of prior movements. Arbitrary movements are any other movements.

The idea discussed in this section is to process the hand movement block-by-block rather than pixel-by-pixel and to treat the movement as semi-curvature movements by limiting the gradient for the next coordinate at each position. A block is a group of pixels in rectangular shape. The goal of the *block-by-block processing* is to allocate a tolerance level in deciding the next position in the movement, as shown in Fig. 5.8 b). The goal of the *semi-curvature treatment* is to narrow the degree of next movement by limiting the gradient according to the previous sequence of coordinates, as shown in Fig. 5.8 c).

The first role of the cursor handling agent in the CBRM in this approach is to take the ownership of the cursor, so that the agent uses the coordinate information from the user’s hand tracking only as the indication for the orientation and the degree of movement not as the direct input value. The agent interprets the indication to make a decision for the next movement. The rules for interpreting the indication are constructed in the OVEN. When the hand tracking process reports the coordinates to the CBRM, the CBRM communicates with the OVEN according to the rules to verify the information and notify the visualization handler and vision handler to adjust the interpreted visualization of the cursor. In this scenario, the agent actually behaves in a similar manner as the hardware mouse does.
The second role is to utilize the semantic reasoning for tuning the tolerance level in deciding the next position in the movement based on contextual information gathered from other input modalities, such as a fragment of word from the conversation segment by the speech recognition.
For example, when the user intends to issue a command like, ‘send a fire truck from here, (pointing P1 in Fig. 5.8 c) at time $t_1)$, to there, (pointing P2 in Fig. 5.8 c) at time $t_2)$’, and as soon as the user issues the command, ‘send a fire truck from here, (pointing P1 in Fig. 5.8 c) at time $t_1$) …’, we can assume that the trajectory of the hand movement will likely be a semi-linear gesture although we do not know the ‘P2’ at ‘time $t_2$’ yet. Therefore, we can lower the tolerance level more for the points during the interval between the ‘time $t_1$’ and the ‘time $t_2$’. There is no reason to represent a shaky or reverse trajectory on the screen for this gesture, and we can assume that the probability of having a reverse trajectory is lower than usual for this particular gesture between the ‘time $t_1$’ and the ‘time $t_2$’.

When the trajectory of a hand movement reverses its direction, it is often unclear whether the reverse trajectory of the hand movement is intended or not. However, the CBRM can improve the accuracy in making a reliable decision whether the reverse trajectory of a hand movement is intended or not, i.e. a meaningful movement trajectory or not, using the contextual information. In the above example, the contextual information by the speech segment, ‘send a fire truck from here, (pointing P1 in Fig. 5.8 c) at time $t_1$) …’, will help the filtering algorithm to filter out the possibility of reverse trajectory with a stronger confidence. The rule for utilizing the contextual information can be defined in the OVEN as not to allow reverse trajectory in this kind of situation.

Since the hand trajectory is a discrete sequence of coordinates in the captured image from a camera and the interpreted sequence of the cursor positions is also a discrete sequence of coordinates, a simple direction information can be gathered from the sequences, in consideration of the nature of a particular command. For example, a command like the ―Send … from … to …‖ will have a semi-linear or curvature trajectory. The Fig. 5.9 shows exemplary rules for helping a filtering algorithm to filter out a reverse trajectory with stronger confidence based on contextual information in this situation.

The reverse trajectory of a hand movement partially suggests additional evidence why a simple filtering algorithm cannot avoid the jitteriness in the hand movement trajectory. In order to predict the next position of the hand movement correctly with the observed position of the hand at ‘time $t_n$’, a filtering algorithm should have enough history of previous coordinate sequences until the ‘time $t_{n-1}$’. This could cause a larger delay of processing the sequence of coordinates. Even if a filtering algorithm is applied, there is a possibility that the actual hand
A trajectory could be correct only in a local area so that the entire trajectory still contains jitteriness in the final trajectory of the hand movement. Therefore, an interpreted hand trajectory could be helpful in the representation of the hand movement on the screen. The CBRM interprets the raw hand trajectory to a smoother hand trajectory using the contextual information as the owner of the cursor on the screen.

```
...<rule> IF 'send' [and] 'location of from: (X₁, Y₁)' [and] 'current location: (Xₖ, Yₖ)'
[and]
'direction information D₁..c-1 from (X₁, Y₁) to (X_{c-1}, Y_{c-1})'
[and]
'direction information D_{c-1..c} from (X_{c-1}, Y_{c-1}) to (Xₖ, Yₖ)'
[and]
'D₁..c-1 match D_{c-1..c}'
THEN 'current coordinate is valid: (Xₖ, Yₖ)'
ELSE 'Notify the 'vision handler' // for possible parameter adjustment
      // in hand tracking and filtering algorithm
      [and]
      'Notify the 'visualization handler' // visualize interpreted hand coordinate
      // rather than the actual coordinate
</rule>
...
```

Fig. 5.9 Exemplary rules for helping a filtering algorithm to filter out a reverse trajectory with stronger confidence based on contextual information.

Depending on the context of current interaction and context at time $t_k$, and after finding out the user’s intention at the current step at the time $t_k$, the CBRM can narrow down the choices of the meaningful movements and guide the hand tracking, using the pre-defined meaningful movement template.

For example, if the user intends to issue a command like, “zoom here, (pointing P₁ in Fig. 5.10 at time $t₁$),” and as soon as the user issues the command at “time $t₁$,” we can assume the trajectory of the hand movement is likely either a pointing gesture or a circle gesture. When there are two possibilities to interpret the initial command and its following gesture, it is difficult to make a decision which gesture was actually intended since we do not know the future...
point $P_n$ at time $t_n$’ and all the sequence of points from the ‘time $t_{k+1}$’ to ‘time $t_n$’ yet, at the current ‘time $t_k$’.

Fig. 5.10 Guided movement by the CBRM

a) A particular trajectory of current hand tracking when jitteriness appears

b) Application of the Guided movement

c) Details of the Guided movement by the CBRM

Fig. 5.10 Guided movement by the CBRM
However, the system can predict and derive the user’s intention from the contextual information based on the history of previous interaction or conversation. For example, the user might have issued a command, “show me the flooded areas” on a map earlier. Based on this contextual information, the CBRM could interpret the deictic command, ‘zoom here’, as designating the concept of area rather than a specific symbolic object on the map, which gives a higher probability in the possibility of a circle gesture.

Furthermore, in the matter of hand tracking, the real-time dynamic trajectory of the hand actually means about 1/30 ~ 1/10 seconds of delay from the time when the hand image is captured as discussed previously. Information for the previous hand positions in a trajectory is available, such as the points from the ‘time \( t_1 \)’ to the current ‘time \( t_k \)’, in the exemplary Fig. 5.10 c). The real-time dynamic trajectory also means we do not know where the next position of the hand will be in the next a couple of image frames, which prevents any fixed assumptions about the next movement, as discussed previously. For example, at ‘time \( t_{k-2} \)’, we do not know where the ‘\( P_{k-1} \)’ will be, and at ‘time \( t_{k-1} \)’ we do not know where the ‘\( P_k \)’ will be, and so on. An observation of users who interact with the systems developed by Krahnstoever, et al. [69] and Rauschert [107] shows that the users’ next hand movement is either ‘holding still’, (actually the hand moves with a tremor within a small range of area even when the user tries to hold the hand still, since the human hand cannot perfectly stay still in the air), or ‘putting down’ to rest, when the gesture is a pointing gesture rather than a circle gesture. Both hand trajectories after a pointing gesture can be classified as the transitional movement trajectories.

However, when the user makes an intentional circle gesture, in the class of meaningful movement trajectory, the sequence of points from the ‘time \( t_1 \)’ to the current ‘time \( t_k \)’ may indicate the circle gesture, especially in the observed sequence of points in the delayed real-time dynamic trajectory. This indication can be helped by the contextual information in the interaction history as discussed above. In this case, the CBRM observes the points from the ‘time \( t_1 \)’ to the current ‘time \( t_k \)’ and apply a guiding template at the ‘time \( t_{k+1} \)’ for the next point. Using the contextual information, the CBRM is able to give higher probability to a certain candidate positions in the next estimation of the hand position.

The guiding template is applied dynamically, which means the template applied at ‘time \( t_{k+1} \)’ does not necessarily have to be the same as the template that will be applied at ‘time \( t_{k+2} \)’. This allows the adaptability to the variable size of the actual trajectory of the hand movement. The
distance ‘Δ’ is approximated as the average distance from the points in the actual hand trajectory to the nearest points in the guiding template. Even if the guiding template could change dynamically, the distance ‘Δ’ does not need to be calculated frequently, since the distance ‘Δ’ is a loosely defined approximation.

5.5.3 Architecture of the Context based Reasoning Module

Flexibility is one of the most important design considerations in the CBRM architecture in order to allow the architecture to work with various computational models. Therefore, the CBRM provides a flexible frame that adapts to different computation models. Regardless of which computational model is used, the frame provides methods for connecting the reasoning module frame to the driver application by an interface, similarly to Flippo, et al. [42].

Based on these ideas, an implementation of the CBRM is currently being developed as a standalone library written in C++. The implementation embodies the flexibility using sub-processes and sub-modules in a hierarchical structure as shown in Fig. 5.11. The sub-processes and sub-modules operate in multiple threads. Virtual functions, based on the object oriented programming principles, are used in order to ensure the polymorphism for adopting any new computational models.

Extensibility is another key design idea in the current system architecture. The implementation allows not only an extension of any reasoning modules on a computational model but also an extension of user handling. In the exemplary application of the CBRM in a multimodal HCI system shown in Fig. 5.3, two reasoning modules and two user handling modules are instantiated from the system. The extension of the reasoning modules helps the system to apply different reasoning rules to different user interactions among the multiple users. The extension of user handlings helps to store user specific information in the system. Storing user specific information helps the system to expedite the user interaction process and provide efficient collaboration among multiple users.

The extensibility for the reasoning modules and user-handling modules allows multiple users to utilize multiple reasoning modules by switching the reasoning modules depending on the context of the interaction. This is implemented using pointers in C++ programming language in
the architecture. The system instantiates a user-handling module for a new user, and it initially assigns a default reasoning module to the user-handling module. Users can switch the reasoning module by requesting the pointer to a new reasoning module. Once the reasoning module is switched, all the consequent processes for dealing with the information, triggered from the interaction, are handled by the new reasoning module.

There are two types of handlers in the CBRM: one is static handler, and the other is dynamic handler. The group of static handlers currently has eight handlers for pre-defined tasks in our system. The handlers in the group of static handlers are ‘vision handler’, ‘speech handler’, ‘multimedia handler’, ‘mobile handler’, ‘semantics handler’, ‘collaboration handler’, ‘dialog handler’, and ‘visualization handler’.

![Fig. 5.11 Architecture of the context based reasoning module and a Set of Handlers](image-url)
The dynamic handlers are not statically dedicated for any pre-defined tasks. They are used when the user needs them before interacting with the system. The handlers in the group of dynamic handlers are ‘enforced rule handler’ and ‘generic handler’. The ‘enforced rule handler’ contains rules defined by the user, and its primary purpose is to simply enforce the user-defined rules to the relevant processing. This is useful for a domain expert to apply certain rules that cannot be categorized as a member of a specific handler. For example, an expert user can define a rule for handling a set of geospatial information in a specific way and apply the rule to the set of geospatial information using the ‘enforced rule handler’. The ‘generic handler’ is like a generic template of the handlers. It can be instantiated to a specific handler for a particular task that is not among the eight pre-defined tasks. The construction of the instantiated ‘generic handler’ is complex. Thus, its primary purpose is to handle more complex tasks. The ‘generic handler’ can be instantiated as many times as needed. This makes the system highly scalable for any application. The implementation of the ‘enforced rule handler’ and ‘generic handler’ is enabled by the usage of the polymorphic operator in our system.

The handlers, especially the static handlers, handle various issues in real-time processing of the information, such as resolution adaptation for different size of the object on the images caused by the various standing point of the user, automatic initialization, issues of not having static boundary such as threshold, issues with noise, lighting variation, user independency (skin color and shape variation), shading, and environment control (obstacles).

In addition, key objectives for the architecture of the CBRM system include the following points:

- A single user or multiple users can use the reasoning module.
- Collaboration among multiple users is facilitated by a collaboration handler.
- Multiple reasoning modules (RMs) can be instantiated in order to process multiple contexts at different levels in various possible scenarios.
- Users can share a common reasoning module, or each user can have its own reasoning module independently from each other.
- Currently, one user can have one instance of a reasoning module at any moment regardless of whether the reasoning module is shared or not. However, if the user
processes multiple jobs in parallel, the system can be extended to adapt to this situation, so that the multiple jobs are processed in different contexts.

- The reasoning module can be dynamically assigned to a user at runtime, thus the user can switch the context of the jobs.
- Reasoning modules are composed of multiple sub-level handlers, and it is easy to add a new handler to the pre-existing set of reasoning modules in the future as it is needed.
- The reasoning module uses an XML based table-like framework, dynamic pattern matching (DPM) in the table, and message passing among the handlers under the reasoning module.
- Reasoning rules are constructed by human expert user(s), in a specific application domain. The system is not intended for self-learning artificial intelligence although the system can be designed to adjust the rules dynamically and interactively, utilizing the base rules.
- The system is designed to be event-driven, in which each pattern matching and reasoning is triggered by event(s) from a driver application. This will make the system dynamically responsive to real-time interactive situations.

The architecture uses an XML based data structure, with which the module allows users to define their own rules (grammar), in the domain specific context. Thus, the computational module can be partially changed according to how the rules are constructed.

5.6 Discussion

This chapter discussed how an information augmentation system like the object verification enabled network (OVEN) can enhance the processes for multimodalities and human computer interaction in relation to a multilingual multimodal HCI system from theoretical perspectives.

There is a possibility that contextual information in user interaction, discourse, and information pieces in a processing unit or a database can improve a multimodal human computer
interaction (HCI) system performance by complementing the key technologies, such as gesture and speech recognition, in the system.

In this chapter, a new method for applying contextual information in a multimodal HCI system is introduced, focusing on the hand gesture recognition input modality. The approach to utilize the contextual information through the object verification in the object verification enabled network (OVEN) for the gesture recognition is called object verified gesture recognition (OVGR). The proposed method is another effort for improving the hand gesture recognition at a higher level than the level of image signals. This makes the proposed method different from the conventional effort.

Our approach for applying the contextual information to a multimodal HCI system is to use an application independent reasoning module called the context based reasoning module (CBRM) in a network that is connected to the multimodal HCI system. The network comprises its own sub-modules, a set of handlers, and the OVEN.

The CBRM is connected to the key processing units in a particular multimodal HCI system, such as gesture recognition or speech recognition modules. The CBRM and the multimodal HCI system communicate with each other through a message passing method. The sub-modules are composed of reasoning modules for handling reasoning process, and user handling modules for handling user interaction and collaboration among the users. The set of handlers processes the contextual information from the related processing units in the system.

Due to the nature of the information processed in a multimodal HCI system, we need a computational model that can handle both quantifiable and non-quantifiable contextual information. The OVEN is a promising information processing framework to develop such a computational model. The OVEN identifies events that occur while processing the tasks and user interaction in the system, and applies pre-defined operation(s) to the events through the polymorphic operators based on the rules.

Typically, an application of the contextual information to the hand gesture recognition process has been introduced either in a low level feature recognition step or in a high level semantics related processing step in a discourse model. Our approach is related to both levels and tries to handle the information flow between the two levels, among several different possible sub-modules in a multimodal HCI system.
For the fusion level application of the CBRM, we can partially use previous work on collaborative discourse interpretation. The CBRM provides more independent and explicitly organized structure in merging the semantics of the user interaction. The CBRM is also designed to be as much generic as possible, so that the module can be easily extended to different application development.

For the feature recognition level application of the CBRM, our approach is a problem domain specific approach. The problem domains include non-trivial technical difficulties that are often found by developers of the multimodal HCI system in the real world environments, while they implement, integrate, and deploy each of such sub-technologies, such as face/hand detection, face/hand tracking, and gesture recognition, in the system. In our system, it is believed that utilizing the feedbacks to each feature recognition level process could improve the feature recognition itself and the system’s overall performance.

When multiple users use a multimodal HCI system, regardless of whether the users interact with the system locally together or remotely through a network, the gesture recognition in the collaborative interaction paradigm requires more complex handling of the information flow. This chapter also briefly discussed how the proposed approach could be deployed in such situations.
Chapter 6

Language Independent Speech Network (LISN)

6.1 Overview

This chapter shows an approach for integrating multiple human language speech recognition and understanding strategies into a multimodal human computer interaction (HCI) system. The approach is called the language independent speech network (LISN). The primary application domain of the approach in this discussion is the geospatial collaborative crisis management situation. However, the approach can be applied to other application domains too.

The LISN is language independent in the sense that the network itself is independent from any specific language although each speech recognition engine (SRE) within the network may take care of a specific language understanding. The LISN can be extended with any number of language-based SREs in a scalable manner. The LISN can be implemented with multilingual SREs, e.g. Fugen, et al. [43] or monolingual SREs. However, the LISN’s primary interest does not lie in any individual SRE, regardless of whether it is multilingual SRE or not. Thus, the ‘multilingualism’ at the SRE level, such as the acoustic modeling of each language among the multiple languages, is not discussed in this chapter.

The LISN is a network in the sense that, not only does it allow each SRE and its processing modules to be located together locally, but also the engines and their processing modules can be distributed remotely. This enables the geospatial collaboration among multiple users in distributed locations to be more dynamic and flexible. The network synchronizes and correlates the semantics from the multiple speech recognition results through a semantic correlation module (SCM). A computational model for the semantic correlation will be discussed later.
6.2 Introduction

As discussed, a multimodal human computer interaction (HCI) and its interface are regarded as an advanced type of interaction paradigm in utilizing a highly complex computerized system and processing information along the interaction. Speech recognition has been known as a major part of the multimodality. It has been integrated into many multimodal HCI system developments. An example can be found in a real-time framework for natural multimodal interaction by Krahnstoeover, et al. [69].

It is possible that the desired human language by users may not be the same as the provided human language for the speech recognition modality in the multimodal HCI system. For example, English may be provided to the users whose primary language preference is not English at the time of interacting with the system. The mismatch could become widened depending on the variation of the application domain, especially when the demographic backgrounds of the users vary in the application domain. Therefore, it will be necessary for the system to have a capability to adjust and switch to the desired language efficiently for an optimal user interaction. In addition to this, it is worthwhile to consider that the system may need to use more than one language for its speech interaction in order to accomplish the goals of given tasks.

The need for multilingual understanding increases in a collaborative environment, such as a geospatial collaborative crisis management situation. In an international geospatial collaborative crisis management environment, the interaction among collaborating users through an emergency management system can be better facilitated through a multilingual speech understanding.

Fig. 6.1 shows an exemplary configuration of the LISN application for an international geospatial collaborative crisis management situation. In the application, the emergency managers in Washington D.C., U.S., Florida, U.S., and Tokyo, Japan try to collaborate for common tasks. In this kind of collaborative environment, the multilingual speech recognition understanding can be particularly beneficial.
Fig. 6.1 An exemplary configuration of the LISN application
Understanding and processing multiple languages is further related to non-real-time multilingual speech audio data, non-speech multilingual textual data, e.g. Chau, et al. [26], Kataoka, et al. [62], and multilingual application content of the system. These could exist in multiple human language formats, especially when the system does not limit the data access to a homogeneous type of database. This requires the system to understand multiple human languages both for the real-time multilingual speech interaction and for the multilingual database information processing.

The LISN suggests an algorithmic approach for handling the discussed issues, by enabling multilingual understanding in a multimodal HCI system in the geospatial collaborative crisis management application domain.

In Fig. 6.1, the exemplary configuration of LISN application is situated in an international collaboration over the Internet for a geospatial collaborative crisis management scenario. In a situation like that in the Fig. 6.1, various interaction patterns can be derived for different scenarios. For example, the emergency manager in Tokyo can communicate with the LISN enabled multimodal HCI application system, i.e. Fig. 6.1 (C), through an independent LISN enabled web application system, i.e. Fig. 6.1 (B). In another interaction pattern, a field agent in Florida, U.S. provides feedback information through a LISN enabled mobile application system, i.e. Fig. 6.1 (D), to the LISN enabled multimodal HCI application system in Washington D.C., i.e. Fig. 6.1 (C). Regardless of the interaction patterns, the LISN system enables emergency managers to speak in their desired languages and then correlates the multilingual information. The types and benefits of the multilingual speech recognition understanding in various situations will be discussed in chapter 8.

Developing multimodal HCI systems is a challenging task. Gesture and speech recognition modalities are a few of many aspects that have been targeted as research topics. Particularly, developing a commercially working multimodal HCI system in a real-world environment poses many challenges. The algorithmic approach of the LISN is mainly derived from the accumulated knowledge for developing a commercially working multimodal HCI system over years. One of the research goals in the LISN is to refine the obstacles that people face in developing a practical real-world HCI system as an embodiment of the LISN and provide an applicable approach for the real-world working system based on scientific discovery and analysis. For this feasibility study, an exemplary embodiment of the LISN is being developed utilizing
commercially available off-the-shelf products. As a result of this experimental study, it is found that the LISN enables a real-world multimodal HCI system to cover a broader range of multilingual application domains and populations in an efficient and scalable way. The details of the approach for the experimental implementation are discussed in this chapter.

6.3 Relevant Work

One of the strengths of a multimodal HCI system lies in its capability to process huge amounts of data in a more efficient way than the conventional interface as discussed in Sharma, et al. [119]. For example, a multimodal HCI system could process a descriptive command in one step user input using a speech recognition input modality, which could otherwise take multiple steps using the conventional interface, such as a keyboard or a mouse. The popularity of automatic speech recognition for a computer-based machine partially originates from this advantage. Unfortunately, the automatic speech recognition modality of a multimodal HCI system could face many obstacles in a real-world working environment, which includes:

- recognition error due to noise in the environment,
- individual speaker dependent characteristics, such as accent,
- training requirement by a particular speech recognition engine, or
- difficulties in speaker identification.

The same obstacles exist even for the monolingual speech recognition with a single user not to mention for the multilingual speech recognition.

Due to this reason, deployment of the system into a real-world working environment usually requires extra engineering techniques in order to alleviate the problems the technology faces and to help the system to perform with a reasonable success level. The application of the extra engineering techniques is commonly requested regardless whether it is a speaker-independent speech recognition system or a speaker-dependent speech recognition system. In general, a speaker-dependent speech recognition system is known to have a higher correct recognition rate than a speaker-independent speech recognition system. However, the speaker-dependent speech
recognition system is also known to require a cumbersome training session for each user prior to the actual interaction with the system. This is very difficult to achieve for a real-world speech recognition system.

When it comes to the multiple user multimodal HCI system, the problem becomes even more complicated. This holds true regardless of whether the system is for monolingual speech recognition or for multilingual speech recognition. Identifying the speaker of a particular speech segment among the multiple users is known to be a non-trivial task. Often, using isolated microphones and a timestamp for the issued speech segment is regarded as a practical solution for this case.

When different languages are used among the multiple users, the problem becomes even more complex. The system’s ability to determine which SRE or language model to use for a particular speech segment becomes a further non-trivial problem to handle while running a set of SREs or a multilingual SRE for the multiple languages.

As noted in chapter 2.1.5., the first step of multilingual understanding is to identify the language the user is speaking before the uttered speech signal is processed by the system if the language is not manually identified.

Navrátil [98] described a phonotactic-acoustic system as a successful approach to the automatic language identification (LID), while introducing different approaches in the topic of language identification. According to Navrátil [98], a compromise between full lexical decoding and a simple phonotactic principle must be found in a LID system in order to make it feasible for real applications.

In the experimental implementation of the LISN, a language switching module (LSM) was used to be responsible for controlling the SREs for multilingual speech input. It could adopt a reliable LID technology, and the automatic language identification feature will increase the value of the LISN. However, currently, the implementation of the LSM uses simpler, non-automatic language switching methods. This is partially due to the fact that the primary interest of the LISN for the multilinguality lies in the multilingual network rather than in an individual multilingual speech recognition engine. This will be clarified when we discuss different types of LISN applications and reasons why the focus of the LISN is directed more toward the network and sub-application components. One of the reasons is that the LISN is developed for a multimodal
HCI system that already provides a much simpler and more reliable controlling input modality for language identification and switching although it may not be automatic.

As Waibel, et al. [141] mentioned, one way to enable multilingual speech recognition in a system is to integrate multiple monolingual recognizers with a language identification module. Another way is to use a multilingual speech recognition engine that combines multiple monolingual recognizers into the single recognizer.

The LISN can work with either of these ways, or both ways simultaneously. Whether the speech recognizer is a multilingual recognizer or a collection of monolingual recognizers, the LISN is designed to be extendible with any number of any types of recognizers.

The mix of a multilingual SRE and a collection of multiple monolingual SREs may not seem to make much sense at first glance. Someone might ask why we need monolingual SREs when there is a multilingual SRE.

An explanation follows: The application domain of the LISN is not limited to a certain geographical location. For a geospatial collaborative crisis management situation, sub-application components of the main application can be distributed remotely. In this case, each sub-application component may need a different approach for the speech recognition. Not all of the sub-application components need a multilingual speech recognition engine for local speech processing. Thus, this situation makes the mix of a multilingual SRE and a collection of multiple monolingual SREs practically more appropriate. For example, in Fig. 6.1, the emergency manager in Florida U.S. may need only an English speech recognition engine, while the collaborating emergency manager in Tokyo, Japan may need a multilingual speech recognition engine for both English and Japanese.

Multilingual studies cover many different application domains, such as multilingual spoken language processing, Lamel, et al. [72], Wang, Z., et al. [145]; multilingual web sites, Huang, et al. [55], Liu, Z., et al. [81], Maeda, et al. [91], Mehta, et al. [94], Morgan, et al. [96], Tonella, et al. [133]; multilingual text, Rösner, et al. [112], Sakaguchi, et al. [114]; multilingual agent, Massaro, et al. [93], Read, et al. [108]; information retrieval, Lin, C. H., et al. [79], Perlman [104], Sheridan, et al. [120]; information access to multilingual database, Hsiao, et al. [54]; indexing multilingual information, Yip, et al. [152]; video text detection, Kim, J. H., et al. [66]; text mining, e.g. Lee, C. H. [77]; text encoding, Mudawwar, et al. [95]; and text-to-speech synthesis, Black, et al. [18], Evans, et al. [41].
Topics for the multilingual multimodal system are found in Lavie, et al. [74, 75]; Burger, et al. [21]; Costantini, et al. [30, 31]; Zue, et al. [154]; Almeida, et al. [3, 4]; Vicsi, et al. [137, 138]; Perugini, et al. [105]; Lyu, et al. [86, 87]; Turunen [134]; and Tanaka [128].

NESPOLE! (Negotiation through SPOken Language in E-commerce) is a multilingual multimodal speech-to-speech machine translation project by a group of researchers, which was funded jointly by the European Commission and the National Science Foundation of U.S., Burger, et al. [21], Costantini, et al. [30, 31], Lavie, et al. [74, 75], Loredana, et al. [84]. Lavie, et al. [74, 75] discussed the architecture and design of the NESPOLE! project. The user interface tries to combine web browsing, real-time sharing of graphical information and multimodal annotations using a shared whiteboard, and real-time multilingual speech communication for Italian, German, French, and English.

Within the NESPOLE! research group, Costantini, et al. [30, 31] and Burger, et al. [21] evaluated the value of multimodality in the NESPOLE! system in the tourism domain for the multilingual communication. Their approach allowed users to interact among each other through the Internet and pen-based gestures. Their studies showed that the multimodality is capable of enhancing dialogue effectiveness while it does not affect the dialogue length. For example, it could aid in the resolution of misunderstanding when ambiguities occur in a dialogue. Especially, they reported that when spatial information is conveyed, the multimodality is clearly better than speech only input in reducing the number of ambiguities and in improving dialogue fluency.

Although the multimodality discussion appears only in a mouse usage, Zue, et al. [154] described a multilingual conversational system. The approach in Zue, et al. [154] is based on the premise that a common semantic representation can be extracted from the input for all languages, at least within the context of restricted domains. They also showed an attempt to separate language dependent information from the system kernel as much as possible. The language dependent information is encoded in external data structures, thus making the internal system manager, discourse and dialogue component, and database language transparent.

In another application domain, the EURESCOM project MUST, (MUltimodal, multilingual information Services for small mobile Terminals) by Almeida, et al. [3, 4], Boves, et al. [20], introduced a multilingual multimodal demonstrator that is a tourist guide to Paris. The demonstrator used speech and pen-based pointing gesture for input, and speech, text, and
graphics for output. Their multilingual question-answering (Q/A) system used a combination of syntactic/semantic parsing and statistical natural language processing techniques to search the Web for the relevant short answers to the questions. Unfortunately, the evaluation for the system by the 12 experts in human machine interaction is mainly concerned about the simultaneous coordinated multimodal interaction.

Vicsi, et al. [137, 138], Perugini, et al. [105], Lyu, et al. [86, 87], Turunen [134], and Tanaka [128] also discussed their work for the issues with regard to a multilingual multimodal HCI system.

This chapter does not discuss the evaluation of the LISN in terms of how effective the multilingual multimodal HCI system is in the geospatial collaborative crisis management context. The evaluation methodology for this kind of complex system requires different approaches than that of conventional methodology for measuring a speech input-based system, and it is beyond the scope of this paper.

As Burger, et al. [21] indicated, classical evaluation methods, such as temporal measurement for task completion, or accuracy rate for word error, will not be appropriate for the evaluation of the efficacy and impact of specific features in a complex communication system, such as their NESPOLE (Negotiation through SPOken Language in E-commerce) in e-commerce application domain or the LISN. In the case of the LISN, for example, a set of measurement criteria for the exemplary scenario discussed in chapter 8 should also include non-classical criteria, such as the evaluation for handling semantic correlation among homonyms in multiple languages. Therefore, a complete evaluation of the LISN using traditional methods is left as a future task. However, it is worthwhile to note that Burger, et al. [21]’s user study reports that multimodal input provides more successful results than speech-only input to the dialogue when the content of a dialogue is related to spatial information.

Although the application domain and the emphasis of the system development are clearly different, Vicsi, et al. [137, 138] introduced a multilingual multimodal speech teaching and training system for 5 to 10-year-old, speech handicapped children, in their SPECO Project. In another separate paper, Vicsi, et al. [137, 138] tried to embrace a qualitative evaluation approach for their system using the answers from speech therapists. One of their conclusions is the effectiveness of the multimodality in their multimodal system was proved, based on the easiness and enjoyment in using the system aided by the multimodality.
Yet, in another application domain for mobile devices, Almeida, et al. [3, 4] presented a work for multimodal multilingual information services for small mobile terminals, called MUST. Given the limitations of the small mobile terminals and having non-professional users as the target user group, Almeida, et al. [3, 4] pointed out that multimodal services are feasible only when it can be built on standard architectures and off-the-shelf components. Another interesting point they addressed is that the mobile networks must be multilingual to allow a customer to speak in a preferred language, such as native language, since the level of performance of the speech recognizer declines for non-native speech. The LIST shares the same idea.

6.4 The Approach in the Language Independent Speech Network

The language independent speech network (LISN) is composed of the LISN core, mobile extension, and a module for the computational model. The LISN core is a collection of the following components:

- Speech information controlling module (SICM)
- Real-time speech recognition processing module (SRPM) and speech information in real-time
- Speech database processing module (SDPM) and a set of database of speech information (which may not need to be processed in real-time)
- Language switching module (LSM)
- Semantic correlation module (SCM)
- A set of speech recognition engines (SREs) for multiple languages in a local area or in remote locations

Fig. 6.2 below shows how these components are connected.
Fig. 6.3 shows an integration of the language independent speech network (LISN) into a multimodal HCI geo-spatial collaborative crisis management application.

The LISN is highly modularized, so that when it is plugged into the main multimodal HCI application, the SICM works as the access point for the main application. A typical multimodal HCI system that utilizes the speech recognition technology has a speech-handling process for controlling the speech recognition. In the exemplary embodiment of the LISN integrated multimodal HCI system, shown in Fig. 6.332, the speech handler is an example of such speech-handling process. The speech handler communicates with the SICM, and the SICM is the main controlling module for the LISN.
From the fact that the LISN is language independent, the speech recognition process in the LISN is also language independent speech recognition. However, it should be re-emphasized that the LISN is not a multilingual SRE. The multilinguality in a SRE level, such as the acoustic modeling of each language among the multiple languages, is not the main interest of the LISN.

Within the network of the LISN, the mobile extension of the LISN also provides a web-based speech recognition application that communicates with a speech server. The web-based speech
recognition application is loaded and used by mobile device(s), such as a laptop or tablet PC. The mobile application for the mobile phone also belongs to the mobile extension of the LISN. The LISN can be extended with any number of language-based speech recognition engine(s), which allows scalability and flexibility in applying the approach to different application domains.

The LISN is also a network of sub-application components. The network synchronizes and correlates the semantics from the multiple speech recognition results of the sub-application components. In the exemplary implementation, this was achieved through a shared message-passing channel in a semantic correlation module (SCM). Overall, the LISN tries to make a multimodal HCI system to cover a broader range of application domains and population in an efficient and scalable way.

6.5 Speech Information Controlling Module

The speech information controlling module (SICM) communicates with a speech-handling process in a multimodal HCI system and its own sub-modules in the LISN. The SICM plays the role of a mediator among the processes. The speech-handling process can also be instantiated from the SICM by inheriting the functionalities in the SICM although this is not necessary.

When the speech-handling process of the main application requests a multilingual speech understanding process to the SICM, the SICM initializes its sub-modules. The sub-modules comprise real-time speech recognition processing module (SRPM), speech database processing module (SDPM), language switching module (LSM), and semantic correlation module (SCM). In our experimental implementation, the SICM is designed to read in a LISN initialization XML file and initializes the language switching module (LSM) by passing the information of the type of SREs, the number of SREs to be used, and the location of the SREs in the initialization XML file.

The SICM controls the processes in the sub-modules according to the request from various sources, including the speech handler, the user, another process of the system, a sub-application component, or any other speech input. This is done through message passing among the modules in our experimental implementation.
6.5.1 Real-time Multilingual Speech Recognition Processing

The real-time SRPM is in charge of recognizing and understanding multilingual speech information in real-time through communicating with a LSM and a SCM. Each SRE operates within the sub-application components of the LISN, as shown in Fig. 6.2, and the results from the SREs are sent to the semantic correlation module (SCM) that synchronizes and correlates the semantics. The correlated semantic information is sent back to the SICM. Once the multilingual speech segments are recognized and understood, the SICM provides appropriate speech recognition results and semantics, which were processed within its own sub-modules, to the main application based on the requests.

6.5.2 Speech Database Processing and Management

The speech database processing module (SDPM) processes the following multilingual speech data, residing in the database, which may not need to be processed in real-time:

- non-real-time multilingual speech audio data,
- non-speech multilingual data, and
- multilingual application content of the system.

These could exist in multiple human languages format, especially when the system does not limit the database access to a certain homogeneous type of database. Some data exist in an audio format and stored in a database for future usage, Heeren, et al. [51], or on the World Wide Web, Rohit, et al. [110], Aone, et al. [8], Basili, et al. [14], Björn, et al. [17], Chen, et al. [27], Gatius, et al. [45].

Non-real-time multilingual speech audio data is the audio data that contains multilingual speech. The SDPM is particularly useful for processing the stored audio data. If any description or attributes for the semantics of the stored speech data is not known, the stored speech data might need to use a speech recognition engine to process the data.
The speech data could exist in a textual format. In this case, a usage of any speech recognition engine is not involved. The textual speech data can be treated as any other textual data. An example of the non-speech multilingual data is multilingual speech segment texts.

The multilingual application content refers to multilingual representation of the application content in a HCI system. An example of the multilingual application content is a multilingual instruction page within a particular screen in an application. The application content is stored close to the processing units, and loaded and unloaded at the execution time dynamically.

6.6 Language Switching Module (LSM)

The language switching module (LSM) takes care of tasks for switching the speech recognition engine among a set of monolingual speech recognition engines (SREs).

Another task of the LSM is to initialize the SREs by communicating with the SICM at the beginning of the LISN initialization. The current embodiment of the LISN uses an initialization XML file at the start of the application execution, which contains information about the structure for the SREs and the LSM. The structural information for the SREs consists of information about the number of SREs and type, name, and location of each SRE, as previously mentioned. Through this information, the LSM comes to know available SREs to switch among them, and where to send a particular speech input segment. By modifying the initialization XML file, the system can extend the set of SREs and dynamically change the settings for them.

As discussed earlier, the current implementation of the LSM does not use any automatic language identification (LID), although Navrátil [98]’s work could be helpful in the future implementation. The LSM can adopt any reliable automatic language identification, and the feature will definitely increase the value of the LISN. However, the primary interest of the LISN for the multilinguality lies in the multilingual network rather than in a multilingual SRE. The application domain of the LISN is geographically distributed in many cases. Therefore, it is more important to handle multiple language information from the entire network point of view rather than a single SRE.

The LISN is embedded into a multimodal HCI system, and it takes advantage of the multimodality of the system by using the other input features as a reliable and accurate way of
switching among the multiple languages. For example, the computer vision based point-and-click gesture or a pen-based point-and-touch gesture enables users to select a graphical icon on a display, which signals to the system to which language it should switch. This simple selection gesture can be robustly implemented, as shown in Krahnstoever, et al. [69].

Although this may not be automatic, it provides a much simpler and more reliable controlling input modality for the language switching. Since the other input modality can be executed in parallel in a multimodal HCI system, the language switching gesture recognition can be done whenever it is desired. This reduces any unnecessary loss of execution time between the language switching intervals.

Unless a single multilingual system is shared by multiple people who try to use different languages for the shared system, there is no significant need of language switching during a session of interaction with the single system. In the case when the sub-application components are run by its own user in a distributed collaboration situation, the users usually set the sub-application components with their desired language, and the language is not usually switched to another during a particular interaction session. However, when a single multilingual system is shared by multiple people, the speech commands should be sequentially executed. If two users utter their own commands to the shared system at the same time, the speech inputs interfere with each other. Even if isolated microphones are used, the speech inputs are processed sequentially. While using a set of monolingual SREs, the speech signals may be processed in parallel in its identified SRE. However, the LISN handles the speech signals in a sequential fashion for such cases when a single multilingual system is shared by multiple people.

Other than the gesture recognition based language switching method, the LISN uses reserved key words to request language switching. For example, a speech command like “Please, switch to Japanese”, requests to switch to a Japanese SRE, and another speech command like “英語（えいご）でください”, (“English, please!”), requests to switch to an English SRE. We are continuously working on the LSM, and a couple of other methods are investigated, including the referenced methods in Navrátil [98]’s work.
6.7 Semantic Correlation Module (SCM)

Fig. 6.4 shows the steps of the multilingual speech understanding process in the LISN. After applying grammar to each speech segment, the meaning of the speech segments can be known. The next step is to correlate the semantics from the recognized multilingual speech segments.

Fig. 6.4 Steps of multilingual speech understanding process in the LISN

The semantic correlation module (SCM) extracts semantics from the recognized speech segments and correlates them according to pre-defined rules. The extraction process is sometimes related to finding common semantics among the speech segments from the multiple languages. For example, when an English speaker and a Japanese speaker collaborate to solve a problem in a specific region on a shared map using their own languages, the English speaker could utter ‘this region’ and the Japanese speaker ‘Gono Chiiki’ in order to represent the specific region. In this case, the SCM helps the system to find common semantics from these speech segments. However, in general, the differences in grammar, vocabulary, idiom, expression, word-formation, and the unique existence of homonym, polysemy, and homophone in each human language make the semantic extraction process very difficult.
In the SCM, the rules for the semantic correlation are defined by the relationship among the verified objects, whose semantics are assigned at the object verification (OV) step. In one aspect, the object verification is a process to make an information segment richer semantically as discussed in chapter 4.

The OVEN provides a file that defines the semantic correlation to the SCM. The SCM constructs a semantic data structure based on the file. Since different languages often have different semantic composition, the semantic assignment in the object verification step transforms the language specific features to the language independent semantic data structure.

The data structure enables the SCM to traverse the nodes in the semantic data structure for mapping the recognized speech segments to the nodes for the semantic units. In case the semantic data structure is implemented based on a tree, i.e. semantic tree, e.g. Angelini, et al. [7], the breadth first search is the primarily search method. Based on the semantic correlation defined in the semantic correlation file, the SCM follows the processing instruction for the mapped semantic units. When two or more semantic tables or semantic trees need to be merged, each tree is regarded as a disjoint set, and the merge follows using the union-find operations of the disjoint set data structure.

The LISN uses a layer, called object representation language (ORL) based on the OVEN, like a meta language as shown in Fig. 6.5. The ORL reduces the complexity of mapping a semantic representation in a language to another semantic representation in another language, among n multiple human languages. An analysis of the semantic correlation processing steps among multiple human languages will be discussed in chapter 8. The ORL is essentially a layer of a set of object representations that defines the semantic relations among the multilingual verified objects.
6.8 Discussion

In this chapter we discussed a novel approach, called the language independent speech network (LISN), for integrating multiple human language speech recognition and understanding into a multimodal human computer interaction (HCI) system based on an information processing approach.

The usefulness and strengths of the LISN are listed as follows:

- **Flexibility**
  - Any type of multilingual object can be verified through the OVEn infrastructure.
  - Using the object representation virtually any computation of the verified objects is possible
- **Linguistic objects are treated as information objects**
  - the information object can be converted to other types of data easily in the information processing system
- **Seamless integration to the information processing system**

Fig. 6.5 Mapping of semantics for a multilingual system and object representation language as a medium for the semantic mapping in the LISN
- Linguistic verified objects are treated in the same way as other verified objects are treated.
- Difference is only in the goal, in which how they are used as specified in their POs.

- User specific interpretation and usage of the information objects are possible in an easier and dynamic way, by modifying the POs (heuristic rules).
- Information mining may be easier
  - Based on the POs, previously unknown facts can be found.
  - Precise machine translation may not be possible. (This is not one of the primary goals in the LISN approach.)

Although the approach has its own unique strengths, due to the nature of the infrastructure that the LISN uses, i.e. OVEN, there are some limitations that we need to further study in the concepts of the LISN. For example, since it is difficult to verify all the multilingual objects in the intended domain, the LISN should be prepared to handle unverified objects. An application of the contextual information could help in this situation as discussed in chapter 5.

Another important fact that we should consider is that the LISN is not intended to be an approach for a precise machine translation (MT) system, e.g. Becker, et al. [15], Walters, et al. [142]. The LISN enables multilingual information processing and collaboration, not necessarily multilingual machine translation. Therefore, due to the fact that LISN is not aimed for getting a precise syntactic translation, the user interaction paradigm and the ensuing interface design need to be understood differently. For example, the user interaction will need to more rely on the richness of the information in the semantics rather than the 1:1 correspondence of the command and feedback. A discussion for this topic appears in chapter 7 with regard to the output presentation of the verified objects.
Chapter 7

Object Verification and Components in Interface Design

7.1 Overview

This chapter discusses three specific components for an interface design in relation to the object verification: 1) database management, 2) output presentation, and 3) web-based collaboration and mobile extension. The goal is to show how the object verification can be utilized in the three components, from the HCI interface design point of view.

7.2 Database Management

7.2.1 Introduction

People are often annoyed when a search engine retrieves data that are irrelevant to what they were looking for. This is particularly aggravated when the size of the database is large. From the user acceptance point of view, retrieved data from a database best satisfies the user when they closely meet the goal for the information retrieval activity, by reducing the gap between the intention of the user who initiated the information retrieval activity and the actually targeted information of the data, Berger, H., et al. [16]. As discussed in chapter 3, the information flow bottleneck can be eased by using a multimodal HCI system. However, the benefit could still be affected by a flood of irrelevant data while interacting with a database.

Therefore, retrieving data sets that are closely relevant to what the user actually wants is crucial from multiple aspects, such as from the point of user satisfaction and acceptance level and from the point of efficient information processing in an information processing system.

In this section, we introduce the VeRification EnAbled Database (VREAD) as a potential solution to help the user to work with closely targeted information. The VREAD designates a database, in which the object verification of the stored data is enabled.
The VREAD adapts the information modeling approach in the object verification approach to the stored data in a database. In the VREAD model, the stored data represent the targeted information of the data through the object verification process since the object verification is an attempt to promote goal-oriented semantics of the stored data.

In relation to a HCI system point of view, the object verification is designed to facilitate a seamless integration of the VREAD into a multimodal HCI interface for handling sophisticated and heterogeneous data and events in the interaction. As Ganta [44] noted, there exist information sources that are highly heterogeneous so that the richness of knowledge extracted from each of the information sources is limited when the information sources are separated. Therefore, seamless integration of the VREAD into a multimodal HCI interface for handling the heterogeneous data is an important task to further facilitate the information processing in the multimodal HCI interface.

With regard to a multilingual information processing system, the multilingual collaborative interaction may use both 1) the multilingual information within a database and 2) the real-time speech information in coordination of the multilingual information in the database. This requires a multilingual information processing system to understand multiple human languages not only for the real-time multilingual speech interaction but also for the multilingual database information processing. A relevant study for a framework for uniform and multilingual access to structured database is found in Xu, et al. [150].

Therefore, as noted in chapter 6, understanding and processing multiple languages in the discussed problem domain of the LISN are further related to non-real-time multilingual speech audio data, non-speech multilingual textual data, and multilingual application content of the system, in a database. The speech database processing module (SDPM) in the LISN is designed to process the multilingual information in a database as it was also discussed in section 6.5.2, through an inter-connected usage with the VREAD. The LISN is primarily concerned with the verified objects in the database that are created for the multilingual information, with regard to the VREAD.
7.2.2 Relevant Work

It is important to note that the object verification and the VREAD is a part of the heuristic information modeling process rather than a database modeling. However, a discussion of the object verification in relation with the features in the conventional database modeling helps us to understand the concepts in the VREAD in an easier manner. Especially the object-orientation model among conventional database modeling methods is relevant to this discussion. In this section, the definitions for the conventional database modeling are based on Ullman, et al. [135, 136].


There have been many works in the object-oriented database and object-oriented database management system (OODBMS) research area. One of the focuses in the object-oriented design of database lies in the feasibility of applying the object-oriented programming paradigm to the database and its management system.

Baroody, et al. [11] discussed an approach where the object-oriented programming paradigm provides structures and facilities for implementing a database. Kim, W., et al. [67] and Banerjee, et al. [10] discussed how to integrate an object-oriented programming system with a database system and data model in ORION. ODE by Agrawal, et al. [2], TIGUKAT by Özsu, et al. [102] and RODAIN by Taina, et al. [127] show exemplary OODBMS among many implementations. HORA by Sutherland, et al. [126] shows an example of object-relational database system as an integration of the object-oriented technology and relational technology. Gottlob, et al. [47] discusses the evolving objects in an OODBMS from an interesting point of view. From the interface design point of view, De Oliveira [34] discussed ten useful features in the user interface design for an OODBMS. Lecluse, et al. [76] also discussed an object-oriented database management system.

The ODL is a language for specifying the structure of a database with the object-oriented design ideas. For example, it is used to develop an object-oriented database management system (OODBMS) in the object-oriented design of a database. Ullman, et al. [135, 136] noted that the
object-oriented database management system (OODBMS) is an object oriented programming system with a capability of database management.

In the object oriented design of a database, the target information for modeling is called ‘object’, an observable entity. In the object-oriented design approach, it is assumed that objects have unique object identities (OIDs), so that one object is distinguished from any other objects. No two objects have the same OID, and no object has two different OIDs.

In the VREAD, the definition of the ‘object’ may be similar to the ‘object’ above, although in chapter 4, the definition of the ‘object’ is introduced in a more restrictive content, i.e. information units based on data or event. In any case, the VREAD is primarily concerned with the verified information units, such as the ‘verified object’ and the OVEN.

In the VREAD, the object verification process may define an object in multiple verified objects. The characteristics of the multiple verified objects for a single object are as follow:

- Each verified object has its own object representation.
- Each object representation may have a unique identification.
- Each verified object is independent.
- If there exist multiple exactly the same object representations in verified objects for a single object, then they are regarded as the same verified object.

There can also be multiple objects whose object verifications may be identical among each other. The characteristics of multiple identical object verifications for multiple objects are as follows:

- If multiple exactly same representations exist for multiple objects, then the objects are regarded as belonging to the same verification group.
- However, each of the verified objects in the same verification group is regarded as the unique verified object.

The ‘entity-relationship’ (E/R) model is a graphical approach to database modeling, which is very similar to the object-oriented approach. In the E/R model, the entity sets are analogous to
classes and the entities are analogous to objects. The relationships are connections among multiple entity sets.

In VREAD, the relationship is defined among the verified objects stored in the database based on their semantics. More precisely, polymorphic operators are defined in such ways that apply the relationships among the verified objects’ representations, based on the predefined rules.

In the VREAD, there is no concept of classes as in the ODL/OODBMS or entities as in the E/R model, which clearly distinguishes the VREAD from the ODL/OODBMS and E/R model.

The relational database management system (RDBMS) represents the data in a two dimensional table called a relation. A schema for the relation is composed of the name of the relation and a set of attributes that defines the structure of a data entry. In this model, a design can consist of one or more relation schemas, so the set of schemas is called a relational database schema. All of the conventional database modeling can be used as a basis to construct the VREAD, but the relational database model is typically a good model for VREAD implementation.

VREAD does not have a rigid relational database schema. The verified objects can be merged and divided. Therefore, a schema within the database for the VREAD can also expand or shrink. Therefore, the entire relational database schema, or the set of schemas can dynamically change.

The operational relationship for the verified objects in the VREAD is defined by the external PO. The data relationship can be defined by the logical relationship between the object and the verified object, such as the verification criteria or guideline mentioned in chapter 4. Therefore, the verification criteria may be used as the relation and the relation name for the table in the VREAD. Since the verification of the object can happen randomly, the VREAD should have a flexible set of schemas to accommodate most of the possible relations. Table 7.1 shows a simple example of the relation by the object verification criteria for the verified objects utilizing the relational database model.
According to Ullman, et al. [135, 136], a database is a collection of information that exists over a long period of time, where the information is often managed by a DBMS. As Ullman, et al. [135, 136] also noted, it is obvious that a database can be designed based on the analysis of the stored target information and the relationships among the components of the information.

A VREAD is a database with verified objects and their object representation as the stored information. The PO can query available verified objects using the object representation as the key in the VREAD. Data operation can be handled through a database management system, i.e. VREAD management system (VMS). For example, the speech database processing module (SDPM) in the LISN stores and retrieves the verified objects in the VREAD through the VMS. Fig. 7.1 shows the architecture of a system that utilizes the VeRification EnAbled Database (VREAD), a VREAD management system (VMS), and an object verification controlling module (OVCM).

As defined in chapter 4, the verified objects are essentially created by information units based on data and event after the object verification step. The verified objects are created

<table>
<thead>
<tr>
<th>Object Representation (OR)</th>
<th>Object</th>
</tr>
</thead>
<tbody>
<tr>
<td>OBJ1 OR</td>
<td>OBJ1</td>
</tr>
<tr>
<td>OBJ2 OR</td>
<td>OBJ2</td>
</tr>
<tr>
<td>OBJ3 OR</td>
<td>OBJ3</td>
</tr>
<tr>
<td>OBJ4 OR</td>
<td>OBJ4</td>
</tr>
</tbody>
</table>

Table 7.1 Relation by the object verification criteria for the verified objects utilizing the relational database model

### 7.2.3 VeRification EnAbled Database (VREAD)

The verified objects are essentially created by information units based on data and event after the object verification step. The verified objects are created
manually by expert users or automatically by internal processes in an information processing system, at the object verification step.

In the case of the manual object verification, the object verification is typically performed by expert users prior to using the system. The expert users essentially enter the verified information units, e.g. the explicit verified objects, to either a pre-existing VREAD or a newly created VREAD, based on the goal of the object verification in the specific domain context.

In the case of the automatic object verification, the result of the object verification by the information processing system can be automatically stored in a VREAD. For example, when an
event is created by the internal processes, such as speech recognition or gesture recognition, the verification module can verify the event as a verified object according to predefined rules, and store the verified object in a connected VREAD.

One of the key differences between the VREAD and the conventional database, e.g. Demurjian, et al. [36], is that the VREAD can be organized based on the object verification information, i.e. object representation, rather than the data itself. This is because the semantics of the information that describes the object distinctively in each context is one of the key factors of the stored data in the VREAD.

The object verification controlling module (OVCM) is an approach to help the construction and storage of object representations in a VREAD. The OVCM can also communicate with another remotely located VREAD to find external verified objects.

There is no upper limit for the number of object representations for an object that is verified. However, if we limit the maximum number of object representations for an object that is verified with a constant ‘k’ and the average size of the object representation as ‘r’, then the size of each verified object stored in the VREAD is as big as ‘k*r’. Therefore, the total size of the VREAD can be estimated as follows:

\[
\text{size (VREAD)} \leq k * r * n_v
\]  

(7.1)

where ‘k’ is a constant for the maximum number of object representations for an object that is verified, ‘r’ is an average size of the object representation, and \( n_v \) is the number of verified objects stored in the VREAD.

For example, if each unit in the ‘r’ has a size of one byte, i.e., ‘r bytes’ per object representation, the VREAD will have a size of approximately ‘k*r*n_v’ bytes.

Since an object can be verified to multiple verified objects and thus represented by multiple object representations,

- the size of the VREAD could be larger than the conventional database, and
- there could exist multiple ‘many-to-1’ relationships and ‘1-to-many’ relationships, such as
- ‘many-to-1’ relationship:
  o multiple object representations \(\rightarrow\) a single object
  o this relationship is formulated when a single object is verified to multiple verified objects that are not identical to each other, thus creating multiple non-identical object representations

- ‘1-to-many’ relationship:
  o identical object representation \(\rightarrow\) two or more objects
  o this relationship is formulated when multiple objects are verified to multiple verified objects that are identical to each other, thus creating identical multiple object representations

The object verification process may need to define an object in multiple concepts, depending on the goals of the target processes in the system. In this case, the object can be represented in multiple object representations, not a single object representation. This creates multiple verified objects derived from the single object. Each of the multiple verified objects will have its own verification information through the object verification process. Table 7.2 a) shows an exemplary ‘many to 1’ relationships for ‘object 1’ and ‘object 3’ in a VREAD.

When multiple object representations exist for an object, each of the verified objects is treated differently as a separate verified object, depending on the object representation. Accordingly, their embodiments exist separately in order to avoid ambiguity.

A pointer can be attached from each object representation of the verified object to the original object for the ease of data access. If an object has ‘m’ number of object representations and verified objects, then ‘m’ pointers exist from each object representation to the original object. The aggregation of the ‘many-to-1’ relationships and pointers between a set of object representations and its corresponding object create overall many ‘many-to-1’ relationships in a VREAD.
### Table 7.2 Exemplary “many to 1” and “1 to many” relationship in VREAD

<table>
<thead>
<tr>
<th>Object Representation</th>
<th>Object</th>
</tr>
</thead>
<tbody>
<tr>
<td>OBJ1 OR1</td>
<td>OBJ1</td>
</tr>
<tr>
<td>OBJ1 OR2</td>
<td>OBJ1 (same as above)</td>
</tr>
<tr>
<td>OBJ1 OR3</td>
<td>OBJ1 (same as above)</td>
</tr>
<tr>
<td>OBJ2 OR1</td>
<td>OBJ2</td>
</tr>
<tr>
<td>OBJ3 OR1</td>
<td>OBJ3</td>
</tr>
<tr>
<td>OBJ3 OR2</td>
<td>OBJ3 (same as above)</td>
</tr>
<tr>
<td>OBJ3 OR3</td>
<td>OBJ3 (same as above)</td>
</tr>
<tr>
<td>OBJ3 OR4</td>
<td>OBJ3 (same as above)</td>
</tr>
<tr>
<td>OBJ4 OR1</td>
<td>OBJ4</td>
</tr>
</tbody>
</table>

#### a) “many to 1” relationship in a VREAD

<table>
<thead>
<tr>
<th>Object Representation</th>
<th>Object</th>
</tr>
</thead>
<tbody>
<tr>
<td>OBJ1 OR1</td>
<td>OBJ1</td>
</tr>
<tr>
<td>OBJ1 OR3</td>
<td>OBJ1 (same as above)</td>
</tr>
<tr>
<td>OBJ1 OR1</td>
<td>OBJ1</td>
</tr>
<tr>
<td>OBJ3 OR2</td>
<td>OBJ3 (same as above)</td>
</tr>
<tr>
<td>OBJ3 OR3</td>
<td>OBJ3 (same as above)</td>
</tr>
<tr>
<td>OBJ3 OR4</td>
<td>OBJ3 (same as above)</td>
</tr>
<tr>
<td>OBJ4 OR1</td>
<td>OBJ4</td>
</tr>
</tbody>
</table>

#### b) “1 to many” relationship in a VREAD

<table>
<thead>
<tr>
<th>Object Representation</th>
<th>Object</th>
</tr>
</thead>
<tbody>
<tr>
<td>OBJ1 OR1</td>
<td>OBJ1</td>
</tr>
<tr>
<td>OBJ1 OR2</td>
<td>OBJ1 (same as above)</td>
</tr>
<tr>
<td>OBJ1 OR1</td>
<td>OBJ1</td>
</tr>
<tr>
<td>OBJ3 OR2</td>
<td>OBJ3 (same as above)</td>
</tr>
<tr>
<td>OBJ3 OR3</td>
<td>OBJ3 (same as above)</td>
</tr>
<tr>
<td>OBJ3 OR4</td>
<td>OBJ3 (same as above)</td>
</tr>
<tr>
<td>OBJ4 OR1</td>
<td>OBJ4</td>
</tr>
</tbody>
</table>
For example, if there are ‘m’ object representations for an object, then there will be ‘m-to-1’ relationships between the ‘m’ pointers of the object representations and the object. If the ‘m’ is the maximum number of the object representations an object can have and if there are k number of such objects, the overall number of the object representations is O(k*m).

If there exists an object representation that is mapped to multiple objects, the semantic relationship is ‘1-To-Many’ relationship. The values of the multiple objects are still duplicated in multiple relations, i.e. rows, in the relational database table. Multiple pointers can be assigned from an object representation to multiple objects. Table 7.2 b) shows an exemplary ‘1 to many’ relationship in a VREAD.

Even if multiple object representations for different verified objects are the same, each object representation of the verified objects is treated differently as for representing a separate verified object, as defined previously. Therefore, the size of the VREAD will not change regardless of whether there are repeated object representations or not.

In the process of using the queried data, the PO can utilize the pointer from the object representation to the many objects in order to apply the relationship to the objects and find best matching result. The method to find the best matching result will be defined in the rules for the PO. Therefore, if the rule for the PO is written in such a way that it allows multiple results, then the operation could also produce multiple results.

The format and data type of the object representation can be based on numerical, literal, symbolical, graphical, or any form of data representation. The polymorphic operator interprets the relationship among the heterogeneous object representations through the predefined relationship among the representations in the rule-base.

The literal form of object representation can be expressed in any human languages. For example, the object representation can contain “name” and “type” elements, using English words, i.e. ASCII coded character data, or any different human languages, i.e. Unicode based character data, as the representation for a verified object.

On the other hand, multiple human languages can be simplified by a unified object representation, since a single object representation method can be applied to the multiple human languages. The object representation language (ORL) is based on this concept.

The VREAD management system (VMS) is concerned with the management of multiple VREADs in a network of VREADs, not just a single VREAD. One of the goals of the VMS is to
provide a capability for the host system to efficiently organize, manage, and access the verified objects in a database.

It needs to be emphasized that the VMS can seamlessly utilize the verified objects in a database in relation to an OVEN based human computer interaction (HCI) system because the object verification unifies the information model in both systems. For example, an object verification enabled database controlling module (VREAD CM) in a host information processing system can communicate with the VMS in order to access the verified objects in a VREAD seamlessly.

Fig. 7.2 shows a network of the object verification enabled databases (VREADs) and the VREAD management system (VMS)
The speech database processing and management is a good exemplary application where the process particularly benefits from the VREAD model. For example, certain audio data are stored in a database for future usage, Heeren, et al. [51]. The verification information in the stored audio data based on the VREAD model can facilitate the processing and management of the stored audio data.

The non-real-time multilingual speech audio data, non-speech multilingual data, and multilingual application content of the system in chapter 6.5.2 are converted to verified objects and stored in a VREAD.

In addition to the speech database processing and management, the VREAD can be utilized in various applications of a database.

Using a search engine to retrieve a data element in a database or a network of databases could often become frustrating when the search engine does not correlate with the semantics or context of the information. However, using the VREAD, the system can utilize the semantic and contextual information in searching data in a complex and non-uniform environment, such as the World Wide Web (WWW).

It is difficult to expect a uniform and homogeneous data structure and information from the numerous web contents, such as the user-created content (UCC), on the WWW. Depending on the perception of the author who created the UCC, the information assigned to the content can vary greatly. Therefore, finding right UCC among the numerous postings is a non-trivial task.

The VREAD can help in retrieving non-uniformly organized or un-organized data, such as the UCC, on the web. Once the information objects are processed through a verification process and stored in a VREAD, the PO can query available verified objects using the object representation as the key in the VREAD.

7.3 Output Presentation

7.3.1 Introduction

This section discusses an approach for presenting the verified objects as output. In a relevant work, Kim, H., et al. [65] discusses a presentation agent for multilingual services.
The object verification process results in emphasizing the primary semantics of the objects into the verified objects. Therefore, in order to express the verified objects according to their intended semantics and fulfill the purpose of the object verification, a novel and intelligent approach for the presentation of the verified objects is needed. The proposed approach is called verified object and multimedia output (VOMO).

The VOMO includes methods and ideas for:

1) the visualization of verified graphical objects, and
2) the presentation of other types of multimedia output on relevant devices, such as playing an audio-type verified object on a speaker.

It is important to point out that the representations of the verified objects themselves, i.e. the object representations, are not necessarily the targets of the presentation. The goal in the VOMO lies in the method of how to best utilize the object representations in an application according to the object verification purposes rather than the method of how to visualize or play the object representations. Therefore, although the object representations can also be explicitly expressed as output, the primary usage of the object representations by the VOMO is to provide information to related processes in achieving the goal of the object verification by the POs, underneath the superficial presentation of the verified object.

One of the applications of the approach can be found in the presentation of the multilingual information. This section will discuss the approaches of presenting the multilingual information through the VOMO.

### 7.3.2 Relevant Work

Output presentation of the object-oriented user interfaces (OOUI) is relevant to the discussion in this section. Therefore, we will revisit the relevant literatures in the OOUI research area from the point of output presentation.

Mandel [92] discussed that the OOUI is a means for taking the graphical interface environment beyond the simple representation of applications by truly focusing on the objects.
Tibbetts [130] pointed out that the OOUI utilizes a GUI for its set of components but adds to it a significant semantic content on the meaning of objects and user interaction.

It is interesting to see the similarities between the OOUI and the VOMO, in the sense that both emphasize the importance of assigning the semantic content to the objects. In both approaches the drawn objects on a screen are not just simple representation of the content.

However, they are also clearly distinguished in the following important points.

- In the case of OOUI, the graphical object is a simulation. In the case of VOMO, the graphical object is a visualized graphical object element (GOEL), which is verified according to the object verification step, thus containing augmented semantics.
- VOMO defines a concept to represent the verified objects in output means, such as monitor screen or speaker, according to the verification information and to facilitate the processes with an interface to optimize the usage of other aspects of an OVEN embedded system, whereas the OOUI is an interface design method. For example, the VOMO focuses on how well and seamlessly an interface that follows the concepts of VOMO works with the features and functionalities, such as the object verification, in the OVEN, whereas the OOUI focuses on the object orientation of the visualized object.
- The VOMO also embraces non-graphical output of the verified objects, whereas OOUI is primarily concerned with the visualized interface.
- In VOMO, the verified objects themselves do not have functionalities. In VOMO, the verified objects and their functionalities are completely separated, and the functionalities are governed by the PO. In OOUI, each object on the screen can contain predefined functionalities.

Regardless of the differences, the relationship between them can be understood as complementary because an OOUI can be a suitable exemplary interface and useful tool that can embody the concepts in the VOMO relatively easily with the object orientation characteristics.

The output presentation in a HCI system causes user responses and feedbacks as ensuing results. From a point of view, the output presentation further initiates next user interactions. Therefore, the output of the verified objects is not regarded as the final result in the VOMO.
An information processing system that utilizes the VOMO defines a set of POs in order to communicate with the user response with regard to the output presentation by the VOMO. It is noteworthy that the POs can be constructed in such a way that the response and feedback information is tightly coupled with the verified object utilizing the verification information. For example, a PO can gather the user’s response specifically in relation to the defined verification information in the object representation, while explicitly expecting to receive a feedback. Then, the feedback can be passed to the process of handling multimodal input in a multimodal HCI system to adjust the parameters in the input modalities.

7.3.3 Verified Object and Multimedia Output (VOMO)

Fig. 7.3 shows a model for the verified object and multimedia output (VOMO).

In an information processing system like a multimodal HCI system, the processes of presenting the output of the processing results can be highly complicated. Especially, when the output is used as a feedback to the users, incorrect output presentation can cause ambiguity and misrepresentation of the information. Therefore, precise presentation based on the verified semantics will help the system to provide more correct information and possibly to prevent any confusion in the interaction, such as the multilingual collaboration.

In addition, it is also expected that the VOMO can further help the user interaction process with an information processing system due to its capability for dynamic presentation. In the dynamic presentation, the presented data is not just a final static report to the users. It is a dynamic output that could contribute to a reproduction of the contextual information in the collaboration. It also helps the system to better assess the users’ intention and goal in the consequent interactions. A graphical object is not just a group of pixels on a display, but a verified information unit that can dynamically be manipulated by the users in the VOMO.

So far, the dissertation has focused on the multimodal input aspects in relation to the multilingual information. However, another role of the VOMO is to help in the multimodal output side of a multimodal HCI system, in addition to the input side of the system.
As discussed in chapter 3, the multimodal HCI system provides more powerful input modalities for complex tasks. The VOMO is designed to provide methods for the visualization and output presentation of various types of multimedia in the advanced multimodal HCI system.

From one point of view, the VOMO is a framework that connects the strengths between the verified objects in the multimodal input and the verified objects in the multimodal output. The approach in the verified object and multimedia output (VOMO) allows the system to display hybrid and complex information dynamically.

In VOMO, the visualized graphical object element (GOEL) on the display is not just an image but contains relevant information to the graphical object. The association of the relevant
information to the graphical object makes the element more intelligent. The audio object element (AOEL) designates both the audio data in a database or in a real-time audio processing. Using the information related to the graphical object element and the audio object element, dynamic manipulation of the data is possible. For example, a polymorphic operator can be defined to utilize the object verification information in the GOEL to achieve a certain goal, such as prioritizing the GOELs on a screen based on its relevance rankings to a search criterion.

It is important to point out that the VOMO is not related to how to convert a data among different human languages. For example, VOMO is not related to how the Unicode representation of a certain language. VOMO is related to how to represent the verified objects on output devices and utilize them intelligently as intended in predefined rules.

### 7.3.4 Output of the Multilingual Information

A multilingual processing system that utilizes the LISN is a good example where an intelligent output presentation approach like the VOMO is needed. Due to the complex nature of the multilingual information, the processes of presenting the output of the multilingual information can be highly complicated. In VOMO, the multilingual verified objects from the LISN can be intelligently presented. Exemplary results include multilingual image data on a screen or multilingual audio data played on a speaker.

As discussed, the dynamic presentation capability of the VOMO can further help the multilingual collaboration process. In the dynamic presentation, the presented multilingual data is not just a final static report to the users, but a dynamic output that contributes to the reproduction of the contextual information in the collaboration and helps the system to better assess the intention and goal in the consequent interactions.

For example, a multilingual graphical object with verified information on a display can dynamically help the collaborators to understand the object better and manipulate it according to the goal of a task with regard to the object without language barriers. Fig. 7.4 shows a screen shot of an exemplary output for the verified objects in the VOMO. The GOEL is an exemplary verified object with Japanese letters that was semantically correlated with a task in English.
7.4 Web-based Collaboration and Mobile Extension

7.4.1 Introduction

Web-based interfaces and mobile devices are important parts of the computer supported multilingual collaboration. Therefore, this section discusses the web-based collaboration and mobile extension with regard to the LISN model and interface design.
### 7.4.2 Relevant Work

Ubiquitous computing refers to the study of pervasive computing devices and the distributed interaction paradigm into people’s daily computing interaction beyond the desktop oriented interaction paradigm, e.g. Tolmie, et al. [131]. The topics in this dissertation can naturally be extended to the ubiquitous computing study in the sense that the paradigm of the computer supported multilingual collaboration can be extended to the ubiquitous devices, such as mobile computing devices.

Abowd, et al. [1] explained three themes in relation to the broad discussion for the ubiquitous computing. According to Abowd, et al. [1], natural interfaces facilitate richer communication between human and computing devices, and the ubiquitous applications need to be context-aware. With regard to the automated capture and access, Abowd, et al. [1] noted that it is more useful to derive additional information from a raw input stream.

In its investigative method, Lumsden, et al. [85] noted that the paradigm shift is necessary in interaction techniques with the mobile devices. One of the reasons is based on the difficulties in designing a mobile interface under the constant movement of the user. The multilingual collaboration on mobile devices will also have to adapt to the paradigm change although the speech recognition based interface is less susceptible to the motion change by the user. For example, under noisy environment, the speech recognition based interface may not work as well as it does under a less noisy environment. Under a noisy environment, the parsed information segment may be fragmented from the time of utterance of voice to the time it reaches the processor, which could make the machine translation to perform less optimally.

Since, the object verification in the LISN can be constructed in such a way that it focuses on the semantic correlation among the lexical segments, it is less susceptible to the error caused by the fragmentation. As a group, the lexical segments may provide contextual information to recover the fragment. Therefore, it will be one of the advantages of the object verification in a multilingual ubiquitous computing with mobile devices.
7.4.3 Web Services and Collaboration

Scribner, et al. [117] describes the Web Services as any functionality that is accessible over the Internet. In general, the Web Services are implemented with eXtensible Markup Language (XML) messaging method, Simple Object Access Protocol (SOAP) specification protocol, and Web Services Description Language (WSDL) interface on a HTTP network protocol although this is not required. Along with web page creation tools on the WWW, the Web Services can be a powerful tool to process and share information in a distributed collaborative environment, Wang, M., et al. [143].

The goal of the research with regard to this area is to investigate the possibilities of creating Web Services for the information in the collaborative environment based on the approaches in the OVEN. Since the OVEN model is designed to be a network, the approaches can seamlessly fit to the Web Services. For example, any verified objects can be written in the XML message formats, and they can be transferred to the other collaborators on the net, using the SOAP protocol. The WSDL will provide detailed information of the verified objects that ease the process of formulating the PO since the WSDL will give the information consumer the scope and characteristics of the semantics in the verified objects.

7.4.4 Extension to Mobile Devices

The web based speech recognition and collaboration paradigm can be extended to mobile devices. Our initial experimentation in the LISN contained a mobile extension. The mobile extension of the LISN handles both:

- a web-based speech recognition application, and
- a speech recognition using mobile devices.

In both cases, the mobile extension communicates with a speech server. The web-based speech recognition application can be loaded and used by mobile devices, such as a laptop, netbook or tablet PC. The mobile application for the mobile phone also belongs to the mobile extension of the LISN.
Although our experimentation showed the possibility of implementing the multilingual speech recognition on mobile devices, the technology seems to be still limited in the sense that not all the mobile devices have the computing power for handling the tasks at the current moment. For example, in the case of a laptop computer, the speech recognition was similarly performed as in desktop PCs. However, when the task was run in the Pocket PC, the performance level was deteriorated due to the low computing power in the CPU and memory, and the inconsistent network connection. As noted earlier, the LISN can be extended with any number of any language-based speech recognition engines, which allows scalability and flexibility in applying the approach to different application domains from the perspective of mobile devices too. Therefore, we hope the further study in this topic overcomes the hardware limitation along with the increased computing power in the future mobile devices.

7.5 Discussion

1) Database management, 2) output presentation, and 3) web-based collaboration and mobile extension are important components in designing an interface for computer supported multilingual collaboration. This chapter discussed the three components in relation to the OVEN.

The VREAD is introduced as a novel approach for the database with verified objects and their object representation as the stored information. In the VREAD, the object verification promotes goal-oriented semantics of the stored data by leveraging the strengths of the semantic augmentation. The speech database processing and management and information retrieval for characteristic data are the a few exemplary application areas where the VREAD can apply the semantic oriented features.

In order to express the verified objects according to their intended semantics and truly fulfill the purpose of the object verification, the verified object and multimedia output (VOMO) was suggested as a novel and intelligent approach for the presentation of the verified objects. The VOMO is not only for the visualization of the verified graphical object, but also for the presentation of other types of multimedia output on relevant devices, such as the play of an audio-type verified object on a speaker.
One key point in the VOMO is that the goal in the VOMO lies in how to best utilize the object representations according to the object verification purposes rather than the method of how to visualize or play the object representations. The primary usage of the object representations by the VOMO is to provide the best semantic information in achieving the goal of the object verification by the POs underneath the superficial presentation of the verified object.

The computer supported multilingual collaboration paradigm in the LISN is extended to the web-based interfaces and mobile devices. A discussion based on our current implementation, on laptop PC or netbook, for the web-based collaboration and mobile extension with regard to the LISN model was briefly noted.
Chapter 8
Experimentation

8.1 Overview

The discussion of this chapter is based on a theoretical work and an experimentation that was carried out through an ongoing implementation of the language independent speech network (LISN). The implementation is an exemplary embodiment of the concepts in the LISN, and designed to support a simulation scenario for a geo-spatial collaborative crisis management situation. The experimentation is intended to show how the proposed OVEN can handle the multilingual collaboration in a host system, such as the LISN, for the simulated situation. This chapter also includes an evaluation of the experimentation based on the execution of the LISN.

As discussed in chapters 4 and 6, one of the goals of the experimentation is to investigate how the information processing model by augmenting the information units, rather than a language model, can handle the multilingual information in a multilingual HCI system like the LISN. Therefore, it is construed that the usage of a limited number of vocabularies is sufficient for the simulation scenario. In addition, the applied grammar is also constructed with a limited syntactic structure in this experimentation. It needs to be re-emphasized that the LISN is not a natural language processing system, as discussed in chapter 6, although the LISN can extend the number of vocabularies and the grammar structure easily.

8.2 Evaluation Methodology

Overall, the evaluation in this section is performed based on a couple of known standard evaluation methods. One method is based on a theoretical evaluation method, namely predictive evaluation. Another method is based on the exemplary implementation and its performance estimation with a simulation, showing the feasibility of the ideas in an embodiment as discussed above.
Therefore, the evaluation methodology in this dissertation may be called a hybrid evaluation method, similar to that of discount usability evaluation, as it is also noted in Preece, et al. [106]. According to Preece, et al. [106], the discount usability evaluation consists of scenarios, simplified thinking aloud, and heuristic evaluation. The discount usability evaluation is a hybrid of empirical usability testing and heuristic evaluation. The traditional evaluation method for usability is replaced with the discount usability evaluation in this chapter. One of the major motivations in this hybrid evaluation method is due to the desire to lower the evaluation cost.

Preece, et al. [106], listed five different kinds of evaluation methods for a HCI system, which are as follows:

- Observing and monitoring
- Users’ opinions
- Experiments and benchmarks
- Interpretive evaluation
- Predictive evaluation

The evaluation method in this dissertation is primarily based on the methods for the predictive evaluation. The predictive evaluation predicts aspects of usage rather than observing it directly. The predictive evaluation utilizes either inspection or modeling.

Well-known predictive evaluation methods comprise the following methods:

- Usage simulations (Expert reviews)
- Heuristic evaluation
- Discount usability evaluation
- Walkthroughs
- Modeling

Preece, et al. [106] noted that the main goal of all inspection methods is to generate a list of usability problems. From a viewpoint, the usability problems are tied to the scientific challenges that were discussed in chapters 2, 5, and 6.
As a part of the hybrid evaluation approach, this chapter will introduce exemplary experimentation with a simulation in section 8.3 and an exemplary implementation that was used for the simulation in section 8.4. In the section 8.5, we will attempt an algorithmic analysis and evaluation of the approaches in both the simulation and the implementation from a heuristic perspective.

8.3 Exemplary Experimentation

The simulated situation in the scenario is as follows:

“Multiple weather forecast experts in Japan predicted a possible massive Tsunami disaster along the south coastline of Japan. An emergency manager in an emergency operation center (EOC) in Tokyo, Japan and another in Florida, U.S. are collaborating to share the information of how to handle the emergency situation. In the exemplary scenario, the manager in Tokyo asks a help from the manager in Florida, U.S. based on his previous experience and domain knowledge from recent incidences that he had.”

In the short exemplary discourse and interaction segments that we used, the two managers are particularly interested in sharing expert knowledge, relevant information, and experience of the field agents in handling the prevention of a levee breakage. The emergency manager in Tokyo asks the emergency manager in Florida for a help to get the information, rather than finding the information by himself, in order to get better context-dependent information in a quicker way. This is because the emergency manager in Florida is more knowledgeable and experienced for the subject matter through the particular incidences of the disaster he recently had in this scenario.

The scenario contains various initiative patterns in order to show how the semantic correlation and reasoning process in the LISN can handle different situations with the multilingual information. Mixed initiatives among participants are made

1) between a LISN-enabled system and the manager in Florida
2) between a LISN-enabled system and the manager in Tokyo
   (Japanese is the primary information conveyance tool.), and
3) between the LISN-enabled system in Florida and the other LISN-enabled system in Tokyo
   (Verified objects created by the LISN in the systems are the primary information conveyance tool.).

Fig. 8.1 shows the languages and information-conveyance tools that are used among the conversers.

![Fig. 8.1 Languages and information conveyance tools](image)

Fig. 8.1 Languages and information conveyance tools used in the language independent speech network (LISN) for an exemplary scenario

Table 8.1 shows one of the interaction segments in the scenario. In the relatively simple exemplary interaction segment, Tanaka, an emergency manager in Tokyo, asks David, another emergency manager in Florida, which area was most severely damaged on the map that they have been sharing until the time of the request. Their own LISN-enabled multimodal HCI system displays a synchronized shared map for a specific region on which they are collaborating. The LISN-enabled system in Japan converts the Japanese request (‘DAVID: 最もっとも

(English is the primary information conveyance tool.),

154
大おおきなひがい[被害]を受うけるちいき[地域]はどこですか – David: Which area was most severely damaged?") in a verified object format and sends it to the other LISN-enabled system in U.S. The verified object format has information about the issued command, such as the command’s main goal is to find a location, “which area”, parametric values for the goal are “most severely” and “damaged”, and the destination of the command is ‘David’ not the LISN-enabled system in Japan.

<table>
<thead>
<tr>
<th>No</th>
<th>Initiator/Responder</th>
<th>Communication Medium</th>
<th>Command/Action</th>
</tr>
</thead>
</table>
| 1  | Tanaka              | Japanese             | "DAVID: 最ももっと大おおきなひがい[被害]を受うけるちいき[地域]はどこですか"  
                                      "David: Which area was most severely damaged?" |
| 2  | System in Japan     | LISN                 | [Verified Object of the request is sent to David.] |
| 3  | David               | English              | "Zoom here" |
| 4  | System in U.S.      | LISN                 | [Map is displayed.] |
| 5  | David               | English              | "Send the map to Tanaka" |
| 6  | System in U.S.      | LISN                 | [Verified Object of the Map is sent to Tanaka.] |

Table 8.1 An exemplary interaction segment from the scenario

The system in the U.S. simply displays Tanaka’s request to David after receiving the object verified command, unless the object verified information has any other processing instruction. In response to Tanaka’s request, David asks his own LISN-enabled system for the relevant information and sends back the result to Tanaka.

Language translation is not one of the main goals in the proposed LISN. In a loosely defined concept of the translation, the semantic correlation may look like the language translation. However, the LISN is not concerned with the exact representation of a translated speech segment in the target language in a display for the correlated semantics. Therefore, in this example, the syntactic structure of the displayed Tanaka’s request to David may not be exactly equivalent to how it could be done if it were translated. In response to Tanaka’s request, David asks his own LISN-enabled system for the relevant information and sends back the result to Tanaka.
Table 8.2 shows another exemplary interaction segment in the scenario. In the interaction segment, Tanaka wants to find out the main reason that caused the breakage (Table 8.2, command no. 1). He first asks his LISN-enabled system whether there exist any relevant information.

<table>
<thead>
<tr>
<th>No</th>
<th>Initiator/Responder</th>
<th>Communication Medium</th>
<th>Command/Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Tanaka</td>
<td>Japanese</td>
<td>“ていぼう [堤防] が 切れる りゆう [理由] は なんですか* &quot;What was the main reason of the breakage?&quot;</td>
</tr>
<tr>
<td>2</td>
<td>System in Japan</td>
<td>LISN</td>
<td>“Verified Object of the Information is not available.”</td>
</tr>
<tr>
<td>3</td>
<td>System in Japan</td>
<td>LISN</td>
<td>(“Do you want to get the information remotely?”) [Verified Object of the request is propagated to David by the system, because the original data belonged to David.]</td>
</tr>
<tr>
<td>4</td>
<td>David</td>
<td>English</td>
<td>“Tanaka: Which specific area are you interested in?”</td>
</tr>
<tr>
<td>5</td>
<td>System in U.S.</td>
<td>LISN</td>
<td>[Verified Object of the Information is shown.]</td>
</tr>
<tr>
<td>6</td>
<td>Tanaka</td>
<td>Japanese</td>
<td>“DAVID: ここから あそこまでです”  “David: I am interested in the area from here to there.”</td>
</tr>
<tr>
<td>7</td>
<td>System in Japan</td>
<td>LISN</td>
<td>[Verified Object of the Information is shown.]</td>
</tr>
<tr>
<td>8</td>
<td>David</td>
<td>English</td>
<td>“Retrieve Experts' opinion on this breakage area.”</td>
</tr>
<tr>
<td>9</td>
<td>System in U.S.</td>
<td>LISN</td>
<td>[Verified Object of the Information is shown.]</td>
</tr>
<tr>
<td>10</td>
<td>David</td>
<td>English</td>
<td>“Find relevant information gathered from field agents from this point to that point.”</td>
</tr>
<tr>
<td>11</td>
<td>System in U.S.</td>
<td>LISN</td>
<td>[Verified Object of the Information is shown.]</td>
</tr>
<tr>
<td>12</td>
<td>David</td>
<td>English</td>
<td>“Any other additional information?”</td>
</tr>
<tr>
<td>13</td>
<td>System in U.S.</td>
<td>LISN</td>
<td>“No additional information is available”. [Verified Object of the request is NOT propagated to anybody by the system, unless the ownership of the original data is known to the system.]</td>
</tr>
<tr>
<td>14</td>
<td>David</td>
<td>English</td>
<td>“Tanaka: From our database, it appears that the main reason of the breakage was a structural problem in the old levee.”</td>
</tr>
<tr>
<td>15</td>
<td>David</td>
<td>English</td>
<td>“Send verified information 1 and 2 to Tanaka.”</td>
</tr>
<tr>
<td>16</td>
<td>System in U.S.</td>
<td>LISN</td>
<td>[Verified Object of the Information is sent.]</td>
</tr>
<tr>
<td>17</td>
<td>Tanaka</td>
<td>Japanese</td>
<td>“かんれん [関連] した じょうほう [情報] を 見みて ください”  “Find the relevant information (in Japan with regard to the information 1 and 2)”</td>
</tr>
<tr>
<td>18</td>
<td>System in Japan</td>
<td>LISN</td>
<td>[Verified Object of the Information is shown.]</td>
</tr>
</tbody>
</table>

Table 8.2 Another exemplary interaction segment from the scenario
It is important to understand that the kind of command Tanaka used in (Table 8.2, command no. 1) is different from that of a simple ‘get and show’ command. In a simple ‘get and show’ kind of command, the outcome is simply to fetch the predefined result, such as either a result of a query or a default result.

However, the system can be designed in such a way that the behavior of the system can produce one step deeper outcome. In order to produce a more intelligent outcome in response to the command rather than simply fetch the result, the system can execute a reasoning process that examines the command’s semantics.

The reasoning process of the system can be initiated for special incidences of commands rather than for all commands. Therefore, whether the reasoning process is needed or not is defined in the verification process for a set of verified objects. For example, the reasoning processing can only be initiated when the command is a query that asks a reason why the breakage occurred for that particular incidence. Therefore, the first step of executing the reasoning process is to distinguish between the commands that need a reasoning process and the other commands that do not need a reasoning process. The object verification in the OVEN helps the system to sense the need for the reasoning process for the first group of commands and initiate the system for the reasoning process.

The OVEN can set up conditional situations to an object with a polymorphic operator, which will also have a processing instruction when the conditional situations are met, in the verification process. Therefore, when the system receives a command that requires a reasoning process, the LISN-enabled system can sense the requirement through the conditional situations and respond to the command as follows:

1) The LISN-enable system checks the verified object to find whether any conditional situation information with a polymorphic operator(s) exists associated with the verified object,

2) If the conditional situation information with a polymorphic operator(s) exists associated with the verified object, the LISN-enabled system determines if the conditions are met by applying the polymorphic operator(s), and

3) The system executes the processing instruction according to the result from applying the polymorphic operator(s).
The processing instruction may require the LISN-enabled system to communicate with another LISN-enabled system. Therefore, a backend communication channel for both sending and receiving messages can be embedded into every LISN-enable system.

In this particular incidence, if the data, which Tanaka had acquired in previous interactions with the system or David, had such object verified information, it would have returned a result based on the information in the object-verified data after following the aforementioned steps in the LISN. However, the system could not find any relevant information, and instead it asks Tanaka whether he wants the system to search for the information remotely (Table 8.2, command no. 3).

Another useful capability of the OVEN is the automatic propagation of the information search to a certain location. In Table 8.2, command no. 3, the system also propagates the fact that Tanaka could not get what he wanted, to the original owner, which is David, by displaying the same information “Do you want to get the information remotely?” on David’s system. This is like a callback function from David’s point of view, which he can turn on before sending the information, as the owner of the information.

This functionality increases collaboration speed and quality, by enabling them to share a result from a party’s command request. If the origin of the information source is not clearly known, such as in the case of Table 8.2 command no. 13, this functionality is safely disabled.

Throughout the interaction in the Table 8.2, the initiatives are changing from Tanaka (no. 1) to the LISN-enabled system in Japan (no. 3), to David (no. 4), and finally to Tanaka (no. 17). It was possible for the LISN-enabled system in Japan to initiate a dialog process because of the verified object that it received from Tanaka that triggered a reasoning process. In the end, Tanaka wanted to assess what options have to be carefully looked at for making decisions based on the specific and useful information from David. As discussed, although David and Tanaka speak in their own language, the semantic correlation module (SCM) in the LISN, along with the OVEN, extracts appropriate semantics and correlates them as intended in the scenario. Fig. 8.2, 8.3, and 8.4 show exemplary grammars that can be used in the scenario.
Fig. 8.2 Exemplary grammar for “get” command in English

Fig. 8.3 Exemplary grammar for image type in English especially for a web application
Fig. 8.4 Exemplary Grammar for “get” command in Japanese
8.4 Implementation of the LISN

Among the components that are discussed in this dissertation, preliminary implementation has been made in progress for the OVEN, LISN, VOMO, CBRM, and a part of the mobile extension, especially in support of the experimentation scenario in section 8.2. This section will focus on a discussion about the implementation of the LISN.

The current LISN implementation is composed of two sub-application components:

1) a standalone application component, and
2) a web-based speech recognition application component.

The LISN uses a collection of monolingual SREs, whose target recognition languages are English, Japanese, and Chinese for the standalone sub-application component. Another type of monolingual SRE for English is used for the web-based speech recognition sub-application component.

The reason that the standalone sub-application component uses a collection of monolingual SREs instead of a multilingual SRE is due to the following facts:

- As Waibel, et al. [141] mentioned, usage of a multilingual speech recognition engine requires a combined acoustic model, which may not be scalable for our application domain. According to this approach, the multilingual speech recognition engine has to be reconstructed after retraining the individual language acoustic model with a reasonable amount of data whenever there is a non-trivial change in the individual language acoustic model, which could be costly.
- Viikki, et al. [139] reported there is marginal recognition rate degradation in the multilingual acoustic model, in a constrained test for the recognition rate comparison, between multilingual acoustic models and monolingual acoustic models. Therefore, there were no particularly strong attraction points in using a multilingual speech recognition engine with the could-be-negligible but still degrading recognition rate.
- Since the LISN is not intended to be dependent on any specific speech recognition engine and the approach provides a framework extendible with any number of any types of
speech recognition engines, the integration of the collection of the multiple monolingual recognizers was easily done within the heuristic approach.

- One of the most important reasons was the availability of already commercialized monolingual recognizers, such as the IBM ViaVoice, Microsoft Speech SDK, or Microsoft SASDK. In our research lab, we have been using these tools for several years, and we have developed non-trivial frameworks, which efficiently work with these SDKs.

The standalone application component is written in C++, and it uses the Microsoft Speech SDK 5.1 as the SRE. The Microsoft Speech SDK 5.1 provides speech recognition for English, Japanese, and Chinese. The web-based speech recognition application component is written in C# and JavaScript in ASP.NET, and it uses the Microsoft Speech Application SDK (SASDK) 1.0 as its SRE in a speech server. The SASDK supports English.

These two application components run their own speech recognition engines independently. Each application component should run in its own computer, because both Microsoft Speech SDK 5.1 and SASDK 1.0 cannot run on the same computer.

Various combinations of the application components are possible. For example, the LISN can use two or multiple standalone application components in a network. It can also use a mix of a standalone application component and a web-based speech recognition application component. The web-based speech recognition application component is intended to be used in the web-environment. Fig. 8.5 shows various configurations of the LISN applications, where different combinations of the application components are used. The application components communicate with each other through a socket communication on a TCP/IP network. The components are tested on Windows XP, Windows 2003 Server, and Windows 2000 computers.

Both a LISN initialization file and an OVEN-enabled semantic correlation rule file are written in XML. The LISN initialization XML file is fed to the speech information controlling module (SICM) that uses a Simple API for XML (SAX) parser. The parsing of the initialization file is usually performed once at the beginning of the system interaction. The SCM parses the OVEN-enabled semantic correlation rule XML file with a Document Object Model (DOM) parser. The DOM API provides a convenient way for traversing and manipulating the nodes in the semantic data structure, i.e. semantic table or semantic tree. Using the API, the SCM dynamically applies the rules to the speech segments. The character encoding for the recognized speech segments are
Unicode based. Thus, the internationalization of coding is practiced. C++ wide-character APIs are used for string processes.

Fig. 8.5 Various configurations of the LISN applications
8.4.1 Technical Aspects and Configurations of the LISN Application

The modularized components of the LISN enable various configurations of the LISN application. Fig. 6.1 in chapter 6 showed an exemplary configuration of the LISN application, in which

- a multimodal HCI system,
- a web application,
- a mobile application,
- a desktop application, and
- Web Services of the LISN

run in distributed geographic locations. Other configurations of the LISN applications are possible, as shown in Fig. 8.5.

Fig. 8.5 (A) shows an English speaker is collaborating with a Japanese speaker using a multimodal HCI system in the same geographic location. Fig. 8.5 (B) shows an emergency manager in Washington D.C. who is collaborating with another emergency manager in Tokyo, Japan, using a multimodal HCI system, web application/desktop application in the distributed geographic location. Fig. 8.5 (C) shows an emergency manager in Washington D.C. who is collaborating with another emergency manager in Florida and with a field agent in Florida, using a multimodal HCI system, web application, mobile application, and Web Services in the distributed Geographic Location.

Fig. 8.6 shows a diagram of processes in a typical configuration of the LISN application, such as the exemplary configuration in Fig. 8.5 (C). In the figure, a SICM and a Web-based speech recognition server are running in parallel for multilingual speech inputs, where the Web-based speech recognition server in the mobile extension of the presented implementation sends recognition results to the SICM. The SICM gathers the recognized speech segments and applies the semantic correlation rule through the SCM. Communication between the SICM and the Web-based speech recognition server is performed through a socket communication.
8.4.2 Contribution of the Implementation

The implementation of the LISN in our experimentation uses publicly available codes and executables for the speech recognition engine, XML parser, OpenGL and image codec, web-browser on the Microsoft’s Windows based operating system. However, a major part of the preliminary stage of the programming implementation that conveys the ideas in the LISN, VOMO, and CBRM has been developed based on the research in this dissertation. The modules that are written based on the approaches comprise:

![Diagram of processes in a configuration of the LISN application](image)
- speech information controlling module (SICM)
- real-time speech recognition processing module (SRPM)
- language switching module (LSM)
- semantic correlation module (SCM)
- XML processing module
- Network processing module
- Image visualization module
- Miscellaneous sub-modules and libraries for supporting the whole architecture

Although the implementation is at the preliminary stage, it embodies the ideas in the research, and shows a proof of the concepts in which the ideas in the research are feasible and practically applicable in the context of the restricted domain discussed in this dissertation.

8.5 Analysis and Evaluation

The analysis and evaluation of the approaches in the exemplary experimentation in section 8.3 and the implementation in section 8.4 will be discussed below. Algorithmic analysis for the approaches will also be discussed in this section.

8.5.1 Simulation

Our focus and goals of the simulation test are to see:

1) whether the proposed deterministic information processing steps in the OVEN could provide an applicable mechanism in handling the multilingual collaboration, by analyzing the computability of the semantic correlation among keywords,
2) how a verified object triggers a reasoning process by discussing how the verified object generates the output in a different way from that of the regular information units, and
3) whether the proposed approach can process the semantic correlation with more contextual semantics by discussing how the initially parsed verified objects are mapped to corresponding semantics based on the contextually predefined rules.

In the study of formal language, the phrase-structure grammar is defined as follows, by Rosen [111]:

\[ G = (V, T, S, P) \]  

(8.1)

where

\( V \) is vocabulary,

\( T \) is a subset of \( V \), \( T \) consists of terminal elements,

\( S \) is a start symbol, and

\( P \) is a set of productions.

Noam Chomsky classified the phrase-structure grammar into four different types of grammars according to the types of productions that are allowed, wherein type 3 grammar is most restrictive and type 0 grammar is least restrictive. Among the four types, the type 1 grammar is called context-sensitive grammar because this type of grammar includes strings that can be derived to next strings only when the surrounding strings for both left side and right side of the production are the same in this type of grammar. In other words, the derived symbols are dependent on the surrounding strings.

Since the LISN is not a grammar, the LISN does not belong to any of the four types of grammars by Noam Chomsky. As discussed, the LISN is context-sensitive in the information correlation process. Therefore, it is important not to be confused between the LISN’s context-sensitive information correlation process and the context-sensitive grammar, i.e. type 1 grammar, in formal language study by Noam Chomsky.

Often the grammar is modeled using finite state machines, such as a Moore machine or a finite state automaton, for language recognition. A deterministic finite state automaton assigns a single next state for each pair of input value and state. Unlike the deterministic finite state
automaton, the nondeterministic finite state automaton can assign multiple next states for each pair of input value and state.

Similar to the deterministic finite state automaton discussed in Rosen [111], the LISN can be modeled using the finite state automata, wherein each state represents a semantic correlation process rather than language understanding process. The modeling of the abstract machine validates the computability of the LISN system from theoretical point of view using the framework that is already used for other study areas such as the formal language study. For example, the LISN can be modeled with a rule-based deterministic finite state automaton as follows:

\[ M = (S, I_{vo}, f_r, s_0, F) \]  \hspace{1cm} (8.2)

where

- \( S \) is a set of states,
- \( I_{vo} \) is a finite set of input verified objects,
- \( f_r \) is a transition function that assigns a next state to each pair of state and input verified object, based on predefined rules,
- \( s_0 \) is a starting state, and
- \( F \) is a subset of \( S \) consisting of the final states.

In this modeling, the transition function \( f_r \) is essentially equivalent to a polymorphic operator. The polymorphic operator maps input verified objects to predefined outcomes in a deterministic way in this example. It is possible to model the LISN using nondeterministic finite state automaton by defining the behavior of the polymorphic operator, i.e. operation rules, for certain special cases, such as parallel processing. However, we primarily discuss the deterministic polymorphic operators in this dissertation.

Fig. 8.7 shows an exemplary state diagram for the information correlation of a homonym ‘bank’ in the LISN, and Fig. 8.8 shows exemplary rules for state transition conditions in a state diagram for the information correlation of a homonym ‘bank’.
where
$s_0$ is a starting state,
$s_1$ is a state that represents ‘a word “bank” exists’,
$s_2$ is a state that represents ‘a financial institution’,
$s_3$ is a state that represents ‘land area alongside a river’,
$s_4$ is a state that represents ‘“a bank of” structure exists’,
$s_5$ is a state that represents ‘a set of similar things arranged in a row’,
$s_6$ is a state that represents ‘find a bank with the lowest APR in the location’,

rule 1: if “bank” and “savings” then ‘a financial institution’,
rule 2: if “bank” and “river” then ‘land area alongside a river’,
rule 3: if “bank” and “a bank of” structure then ‘“a bank of” structure exists’,
rule 4: if ‘“a bank of” structure’ and ‘noun’ then ‘a set of similar things arranged in a row’, and
rule 5: if “bank” and ‘pointing gesture to a location’ then ‘find a bank with the lowest APR in the location’.

Fig. 8.7 An exemplary state diagram for the information correlation of a homonym ‘bank’ in the LISN
The process that involves multiple human languages inherits a group of topics that are nontrivial to overcome from the linguistics. For example, the topics include the followings:

- Differences in grammars, vocabularies, idioms, expression, and word-formation in the multiple human languages need to be re-conciliated to a satisfactory level for reasonable semantic mapping among the multiple human languages.
- Lexical and semantic ambiguities that exist in certain types of words need to be handled with a more complex reasoning process. The types of words that cause the ambiguities include homonyms, polysemes, and homophones.

1) Homonyms are defined as words that are written and pronounced the same way as another, but have different meaning and different etymology, such as ‘bank’. The word ‘bank’ has different meanings with different etymology in phrases like ‘a river

---

**Fig. 8.8 Exemplary rules for state transition conditions in a state diagram for the information correlation of a homonym ‘bank’**
bank’, ‘a savings bank’, and ‘a bank of switches’ although the word ‘bank’ in each phrase is written in exactly the same way.

2) Polysemes are defined as words that have two or more similar meanings with same etymology, such as ‘box’. The word ‘box’ has multiple meanings, such as ‘a type of tree’, ‘a container’, or ‘a seating area’.

3) Homophones are defined as words that sound the same but have different meanings, such as ‘right’ and ‘write’, ‘to’, ‘too’, and ‘two’, or ‘rain’, ‘rein’ and ‘reign’.

The reconciliation of the linguistic differences in multiple human languages is particularly important for the LISN system. The semantic correlation in the LISN is not the same as language translation. Rigorous translation among the multiple human languages is not the primary function of the system. One of the goals in the LISN system is to map the semantic match and represent the interpretation based on the object verification to the human users with their own mother language as best as it can. Therefore, the correctness of the semantic mapping can be measured by measuring how closely the semantic correlation conveys the intended usage of the verified objects at the verification stage. This is one of the features that distinguish the LISN system from the traditional language interpreter.

The LISN system utilizes the object verification process to clarify the semantics of words in homonyms, polysemes, and homophones by defining the meanings in multiple layers at the object verification stage. Each layer can contain contextual information that aids a correct semantic correlation. For example, the word ‘bank’ can be primarily defined as an institution for financial transaction when the word ‘bank’ appears with the word ‘savings’. Here, the word ‘savings’ provides the contextual information for the word ‘bank’ to be correlated with financial institution. Likewise, it can be primarily defined as the land area alongside a river when it appears with relevant words like ‘river’, and as a set or pile of similar things arranged in a row when it appears in a structure of [‘a bank of’ + noun].

However, it is important to note that the object verification process can augment the word ‘bank’ beyond the boundary of dictionary definition. For example, although the primary verification of the word ‘bank’ can be an institution for financial transaction when the word ‘bank’ appears with the word ‘savings’, the secondary verification of the word ‘bank’ can be a special function or procedure that performs a whole different task, while being associated with
the ‘bank’ and ‘savings’. For example, the procedure can be defined as ‘a procedure to find bank with the highest or lowest APR in the area the user is using the LISN system’. As discussed in chapter 4, the user-defined function or procedure, at the object verification stage, can be any task that the user wishes to achieve with the verified object, i.e., in this case the word ‘bank’. This characteristic in the LISN system opens a way for a potentially dynamic system behavior in the information processing.

8.5.2 Algorithm Analysis

This section discusses computational time in the algorithms through an algorithm analysis. Traditionally, one of the strategies for the evaluation is to measure system performance in terms of processing speed. In this evaluation, a set of test data is passed to one of the independent core components with temporal constraints. The processing speed inside a component can be measured utilizing timestamps between the input time of a set of test data into the particular component and the output time when the component produces a result. However, this evaluation approach is hardware dependent and highly relative to the hardware performance. Therefore, instead of discussing a measurement of the processing speed for the components in the implementation, we will discuss the computational time in the algorithms through an algorithm analysis. Since the architecture is highly modularized for the approaches in the LISN, it is feasible to isolate each component and discuss the computational time in each component independently.

The goal of the efficiency measurement in computational time is similar to the goal of finding optimal search algorithm for a key in a data structure. It is because the performance of the semantic correlation depends on how efficiently the system finds the matching keyword in a semantic data structure of a language corpus. Therefore, this part of the evaluation can be replaced by the theoretical analysis using the techniques known in the algorithm analysis for the semantic data structure and operation.

Since the behavior of the nodes in the semantic data structure in the LISN approach has a similarity to the behavior of the disjoint set data structure, the algorithm analysis will analyze the disjoint set data structure operations in the semantic data structure.
As an information conveyance tool, the object representation language (ORL) reduces the complexity of mapping the key words from a group of keywords to a group of semantic units. For example, if there are \( l \) total human languages that are used in a multilingual multi-user multimodal human computer interaction system, then the processing time, \( T \), for mapping a keyword from a language to another keyword in another language, without the ORL, can be calculated as follows:

\[
T(l)=C_1 = l
\]

\[
T(n_i) = O(n_i)
\]

\[
T_i = T(l) \times \left( \frac{\sum_{j=1}^{l} T(n_j) }{l} \right) = l \times \left( \frac{\sum_{j=1}^{l} T(n_j) }{l} \right)
\]

\[
T_j = T(l-1) \times \left( \frac{\sum_{j=1}^{l-1} T(n_j) - T(n_j) }{l-1} \right) = (l-1) \times \left( \frac{\sum_{j=1}^{l-1} T(n_j) - T(n_j) }{l-1} \right)
\]

\[
T = T_i \times T_j
\]

\[
O(n^2)
\]

where

- \( l \) is total number of languages,
- \( T(l) \) is the time to identify the language of the input keyword,
- \( n_i \) is total number of keywords in a language \( l_i \),
- \( T(n_i) \) is the time to search the input keyword in a language \( l_i \),
- \( T_i \) is the time to identify the semantics of the input keyword in a language \( l_i \),
- \( T_j \) is the time to identify another keyword in another language \( l_j \) that has the same semantics of the input keyword in a language \( l_i \), and
\( T \) is the time to map the input keyword in a language \( l_i \) to another keyword in another language \( l_j \).

The processing time, \( T(l) \), is equivalent to a selection of an element from the set of languages in the LISN, as discussed in chapter 6.3 for language identification. Therefore a combination notation is used. \( T(n_i) \) is based on worst case running time for search in a hash table structure. The term, \( T_j \), assumes that the language \( l_j \) is different from the language \( l_i \).

However, with the ORL, the processing time can be reduced as follows:

\[
T(l) = \sum_{i=1}^{l} T(n_i) = O(n_i)
\]

\[
T_i = T(l) \sum_{i=1}^{l} T(n_i) = l \sum_{i=1}^{l} T(n_i)
\]

\[
T_o = c_o
\]

\[
T = T_i + T_o = 2c_o l O(n_i)
\]

where

- \( l \) is total number of languages,
- \( T(l) \) is the time to identify the language of the input keyword,
- \( n_i \) is total number of keywords in a language \( l_i \),
- \( T(n_i) \) is the time to search the input keyword in a language \( l_i \),
- \( T_i \) is the time to identify the semantics of the input keyword in a language \( l_i \),
- \( T_o \) is the time to map the input keyword in a language \( l_i \) to \( c_o \) pointer(s) to object representation(s), and
- \( T \) is the time to map the input keyword in a language \( l_i \) to another keyword in another language \( l_j \) using the \( c_o \) pointer(s) to object representation(s).
The processing time, $T(l)$, is equivalent to a selection of an element from the set of languages, same as above. $T(n)$ is based on worst case running time for search in a hash table structure. The term, $T(l)$, from the previous calculation is not necessary in this approach because the process of identifying another keyword in another language $l_j$ that has the same semantics of the input keyword in a language $l_i$ is replaced by the $c_o$ pointer(s) to object representation(s) that have been already calculated at the $T_i$.

In the exemplary embodiment of the LISN, the verified objects of the multilingual information can be organized in a hash table structure or a tree-like data structure, i.e. semantic tree, according to the semantics of the object representation. The semantic relationship of the verified objects is expressed in these structures. The semantic data structure is used to explain the relationships and analyze the operations among the verified objects in the specific experimentation. In the case of a semantic tree, a single node may contain a pointer to a verified object, regardless whether the verified object is a simple verified object or a composite verified object. The multiple nodes in the data structure are organized according to their relationships in the hierarchy.

In essence, the semantic data structure provides the semantic relationship for the ORL from the LISN perspective. The verified objects from the multilingual data objects can form meaningful relationships from the lexical point of view.

However, it is important to note that the data structure for the relationships among the verified objects in the LISN does not need be a tree-like structure always. The verified objects even do not need to be in a rigid data structure in case the relationships do not exist. It is because dynamic data processing is one of the key strengths in the LISN supported by a deterministic information processing framework. Thus, limitations for a data structure in the LISN can be loosely applied.

Fig. 8.9 shows an exemplary semantic data structure that is constructed using multiple hash tables in multiple layers. For example, the keys in the table $T$ can be associated with secondary tables $(T_1, T_2, ..., T_i, ..., T_n)$ in the next level of layers through the pointers that point to memory spaces for the tables $(T_1, T_2, ..., T_i, ..., T_n)$. 
Fig. 8.9 Exemplary semantic data structure using hash tables
$U$: universe of keys for a first layer definition,
$K$: actual keys for the first layer definition,
$\Theta(|K|)$: storage requirement for a first layer definition,
$O(1 + \alpha)$: average search time for the hash table $T$,
$h: U \rightarrow \{0, 1, \ldots, m - 1\}$

where
$h$ is a hash function that maps the universe $U$ of keys into the slots of a hash table $T[0..m - 1]$,
$n$ is the number of elements,
$m$ is the number of slots,
$\alpha = \frac{n}{m}$ is the load factor,

$U_i$: universe of keys for a semantic group $SG_i$ in a second layer definition,
$K_i$: actual keys for the semantic group $SG_i$,
$\Theta(|K_i|)$: storage requirement for the semantic group $SG_i$,
$O(1 + \alpha_i)$: average search time for the hash table $T_i$,
$h_i: U_i \rightarrow \{0, 1, \ldots, m_i - 1\}$

where
$h_i$ is a hash function that maps the universe $U_i$ of keys into the slots of a hash table $T_i[0..m_i - 1]$,
$n_i$ is the number of elements,
$m_i$ is the number of slots,
$\alpha_i = \frac{n_i}{m_i}$ is the load factor,
$T_j$ is another secondary table $T_j[0..m_j - 1]$,
There can be at most $K$ number of secondary tables $(T_1, T_2, ..., T_i, ..., T_k)$ in the next level of layers. The size of $U$, is essentially the same as the size of $K_i$ since the secondary tables are used for specific object verification of each keyword in a predefined manner. The average search time for each of the secondary tables is also bound by $O(1 + \alpha_i)$ as the primary table. The sizes of $n_i$ and $m_i$ are dependent on the object verification process. In practice, the sizes of $n_i$ and $m_i$ can be controlled so that they can be within predefined limits. Therefore, the search time to find a keyword and its object verification in the exemplary data structure is as follows:

\[
\frac{1}{n} \sum_{i=1}^{n} \left(1 + \frac{i-1}{m}\right) + \frac{1}{n_i} \sum_{j=1}^{n_i} \left(1 + \frac{j-1}{m_i}\right) \\
= \left(1 + \frac{1}{nm} \sum_{i=1}^{n} (i-1)\right) + \left(1 + \frac{1}{n_i m_i} \sum_{j=1}^{n_i} (j-1)\right) \\
= \left(1 + \frac{1}{nm} \frac{(n-1)n}{2}\right) + \left(1 + \frac{1}{n_i m_i} \frac{(n_i-1)n_i}{2}\right) \\
= \left(1 + \frac{n-1}{2m}\right) + \left(1 + \frac{n_i-1}{2m_i}\right) \\
= \left(2 + \frac{\alpha + \alpha_i}{2} - \frac{1}{2m} - \frac{1}{2m_i}\right)
\]

which is $\Theta\left(4 + \alpha - \frac{1}{m}\right)$, including the time for two hash function executions for the primary and secondary tables. (8.5)

Therefore, a successful search time takes $\Theta(1 + \alpha)$ in the exemplary semantic data structure. In essence, this is equivalent to executing the search operation twice, i.e. once for the primary table and once for a linked secondary table.

Additionally, the semantic data structure can be implemented using a semantic tree that includes sub-trees as the members of the parent node’s disjoint set, inclusively and recursively. The value of a single node is a pointer to a simple verified object in a memory space. The sub-trees and their parent node form a disjoint set of multiple nodes, which is a composite verified
object as a whole. Each of the sub-trees is recursively a simple or composite verified object. The links that connect the nodes represent tight bindings.

The value of each node in the semantic data structure can also be implemented to be a pointer to an external disjoint set. In this case, the pointer in a node will point to the representative of the external disjoint set. The members of the disjoint set exist in the pointed external disjoint set rather than a sub-structure. For example, each node in a tree or a hash table can point to an associated disjoint set of a composite or a simple verified object.

In this case, the links that connect the nodes represent loose bindings. The relationship between a parent node and a child node is semantically loosely bound, in which the child is not a member of the parent’s disjoint set, i.e. not part of the parent verified object. The child node itself points to a disjoint set that is an independent verified object. The child node is simply a loosely relevant verified object to another verified object pointed by the parent node. Therefore, organization of these nodes from the entire semantic data structure point of view can be organized according to a loosely defined semantic relation.

The characteristic of the behavior of each node is similar to the functional behavior in the disjoint set data structure. Each node in a semantic data structure can be created as a disjoint set. When two nodes need to be combined, a union operation between the two disjoint sets can be performed.

Therefore, the type of a node in a semantic data structure may not be homogeneous. The size of each node in a semantic data structure may not be fixed either.

Cormen, et al. [29] notes that, in the disjoint set data structures,

- finding a disjoint set to which an element belongs and
- uniting two disjoint sets

are important operations. These two operations are also equally important for the semantic correlation processes.

Depending on how the polymorphic operator (PO) is designed, the execution of the PO may require two disjoint data structures or sub-structures to be merged together. For example, this could happen in order to make a disjoint data structure contain richer information by increasing the number of connected semantics and having more complicated semantic structure in a single
disjoint set data structure. This is where the merge of the two or multiple disjoint data structures can play an important role.

For example, each linked list in the exemplary semantic data structure using hash tables shown in Fig. 8.9 can be regarded as a disjoint set. The union operation among the disjoint sets in the data structure can be performed by merging the relevant disjoint sets using union operation as shown in Fig. 8.10. In Fig. 8.10, two linked lists from table $T_j$ are appended to a linked list that is associated with the slot $s_i$ in the table $T_i$. In this approach, the representative of the set can be the relevant slot in the secondary table. Therefore, the operation for finding a set can also be performed by simply returning a pointer to the relevant slot for the linked list in the table $T_i$.

As discussed in Cormen, et al. [29], the total number of objects updated by all the union operations is

$$
\sum_{i=1}^{q-1} i = \Theta(q^2)
$$

where

$m$ = the total number of operations for making new sets, union, and finding a set,

$n$ = the number of operations for making new sets,

$n = \lceil m/2 \rceil + 1$,

$q = m - n = \lceil m/2 \rceil - 1$, (the number of operations for union and finding a set),

an amortized time of an operation is $\Theta(m)$.

In the example shown in Fig. 8.10, the operations for making new sets may not be needed since only pointer updates are essentially needed for the union operations. Therefore, the first step for an union operation is to search and find disjoint sets, i.e. linked lists, in the semantic data structure. As discussed above, a successful search time takes $\Theta(1 + \alpha)$ in the exemplary semantic data structure. The operation for finding a disjoint set returns a pointer to the slot that is associated for the linked list in the hash table.
Fig. 8.10 Union of disjoint sets in a semantic data structure implemented in a hash table
Cormen, et al. [29] also discusses the union operation by applying a weighted-union heuristic, in which smaller linked lists are appended to longer linked lists. This approach may be applied to the exemplary semantic data structure.

In order to correlate semantically equivalent vocabularies in multiple languages, the process needs to find matching verification information in each corresponding language corpus in a specific application domain. This is where finding the element’s location in an entire collection of sets of disjoint sets becomes significantly relevant. Once the matching verification information is found, the verified object is passed to another collaborator’s LISN system. The system executes the search operation in its own set of verified objects according to the semantic data structure. In this operation scenario, the verification information in the verified object is commonly shared between the collaborators’ systems.

The process of constructing an object representation might look similar to the process of finding a key in a disjoint set data structure. However, the object representation could comprise heterogeneous data types in a higher-level data structure. For example, entire key concepts of the object can be represented in one composite object representation. On the other hand, only a partial concept of the object can also be used as the object representation.
Chapter 9

Conclusion

9.1 Summary

This dissertation discussed the following topics:

1) Complexity in multilingual computer supported collaboration,
2) Needs for a multilingual multimodal HCI system to enable the multilingual computer supported collaboration,
3) Scientific challenges in implementing the multilingual multimodal HCI system,
4) Previous work for such systems,
5) Suggested approach, and
6) Evaluation of the approach,

in a geo-spatial collaborative crisis management domain.

9.1.1 Suggested Approach

An experimental multilingual HCI system, called language independent speech network (LISN) was discussed. The LISN was introduced as a potential solution to enable the multilingual computer supported collaboration. Studies for the detailed modules and processes in an exemplary embodiment of the LISN were also discussed.

The LISN utilizes a deterministic information processing framework for the correlation of the semantics among the recognized multilingual speech segments. The framework is called object verification enabled network (OVEN). The speech segments are processed in sub-application components. The recognized multilingual speech segments are converted into verified objects, and their semantics are correlated by the execution of polymorphic operators in the semantic
correlation module (SCM). The relationships of the semantics are pre-defined based on the rules designed by expert users in a specific application domain.

Many multimodal HCI systems can benefit from the approaches in the LISN for collaborative applications. Geospatial collaborative crisis management is one of the application domains that are well suited for the strengths of the LISN, and it is the primary application domain in this dissertation.

In the experimentation, a scenario that simulates a geospatial collaborative crisis management situation was developed. The simulated scenario showed how the LISN-embedded multimodal system facilitates the user interactions particularly for handling the complex situations and processes the multilingual speech inputs in the context of a geospatial collaborative crisis management situation.

The LISN is composed of a core, a computational model, and a mobile extension. The core is built with a collection of processing modules. The processing modules enable efficient and scalable usage of multiple speech recognition engines (SREs). The processing modules also synchronize semantic correlation among the recognized speech segments from the multiple SREs.

The modularity of the LISN allows the system to be utilized in various types of collaboration cases. For example, the sub-application components and the collection of processing modules can be re-aligned case by case, accommodating special requirements in a specific collaboration case. Since geospatial collaborative crisis management may require a spatially distributed collaboration among the participating users, flexibility that can accommodate various types of collaboration will be one of the most important features in the LISN.

The scalability of the LISN provides more flexibility in the utilization of the set of SREs. This also means the LISN-enabled system can reach to a broader range of multilingual application domains and populations with various demographic backgrounds, in a scalable and resource-efficient way.

From one aspect, the OVEN is a way to augment the information unit, called the ‘object’, with user-defined intrinsic semantics. In the OVEN, the object originates from two classes of information units: 1) data of various types or 2) events that are created as an outcome of interworking in a multimodal HCI system. The OVEN applies a syntactic structure to the information units, i.e. verified objects, where the syntactic structure executes a function between
the information units based on the relationship defined in a rule base. The syntactic structure is called as ‘polymorphic operator’.

9.1.2 Evaluation of the Approach

The evaluation of the proposed approach was performed based on a predictive evaluation method and experimentation of a developmental system. We learned the following points through the evaluation:

- Intrinsic attributes of a multilingual multimodal HCI system are complex due to the following factors:
  1) Linguistic differences exist among the multiple human languages used in the system. The differences are caused by both semantic and syntactic aspects of the languages. Therefore, the multilingual multimodal HCI system should address both semantic and syntactic aspects of the languages.
  2) A multimodal HCI system consists of multiple input modalities. Each low-level signal processing still has challenges for better accuracy although improvement in the relevant technologies has been made for the last decades. The inaccuracy is partially caused by imperfect real-world data that the system acquires. This is a fundamental limitation of approaches that try to improve the accuracy of the low-level signal processing. Thus, advancement in the low-level signal processing alone is not enough to enable proper human computer interaction in a multilingual multimodal HCI system. Therefore, the multilingual multimodal HCI system should address high-level information processing also.

- The idea of information augmentation and an algorithmic framework that operates on the augmented information are a suitable approach to address the issues above in a multilingual multimodal HCI system, further including the following reasons:
  1) Information augmentation in the suggested information processing algorithm facilitates semantic correlation and provides an efficient strategy for utilizing the contextual information in a multilingual multimodal HCI system.
2) Usage of polymorphic operators along with the augmented information processing algorithm, such as the OVEN, in a multilingual multimodal HCI system facilitates an efficient handling of the linguistic differences among the multiple human languages and unique characteristics in each language.

- Usage of a meta language improves processing time in the task of mapping words among the multiple languages from $O(n^2)$ to $O(n)$ asymptotically. The meta language used in this research is called as object representation language (ORL).
- Multilingual collaboration can be facilitated by a multilingual multimodal HCI system. For example, multimodality improves semantic correlation process in a multilingual collaboration.
- Matching keywords in the target language in the semantic correlation process can be operated with the same computational complexity as that of an application of a fast searching algorithm, such as a binary search tree. A semantic data structure was introduced using multiple hash tables in two layers, and it was discussed that a successful search time to find a keyword in the exemplary data structure takes $\Theta(1 + \alpha)$, where $\alpha$ is the load factor.

The suggested approach in this dissertation also covers multilingual output and language identification of the speech input. The research limited the discussion within the context of the geo-spatial collaborative crisis management. Therefore, the corpus of the multilingual data in the research, especially both for developing the experimental multilingual HCI system and evaluating the system was restricted within a limited range.

The research also discussed various viewpoints for a highly efficient and feasible multilingual communication and interaction method among the collaborators for handling complex multilingual data and information. It is hoped that the viewpoints contribute to the tasks that attempt to maximize the multilingual collaboration results in the specified domain.
9.2 Future Work

The approaches introduced in this dissertation cover a broad range of components in a multimodal HCI system. A further study in each component will be needed in the future. A few points for the future work in each component are discussed in the following sections.

9.2.1 User Collaboration Handling Based on the Information Processing Scheme

The further relationship between the computer supported collaboration and the information processing approach introduced in this dissertation will remain as a part of the future study.

The implementation of the LISN is still in the developmental stage although a major part of the framework has been implemented. As part of the future work, more efficient algorithms for the semantic correlation will be tested. For the implementation of a working system, it is hoped that a sub-application component for mobile devices, such as mobile phone, smartphone, and tablet PC, is implemented for the LISN. A continuous improvement of the current implementation for the laptop PC and tablet PC will be made. Another important future work is to investigate the usage of automatic language identification (LID) and multilingual SRE and compare the approach with the current approach of using a multiple single language SREs, in a real-world application situation.

9.2.2 Usability and Human Factor Study

A user study will be needed among the multilingual collaborators using the system developed for the experimentation. In general, it is difficult to sample unbiased group of people for the human usability study of the HCI interface. Especially, selecting multilingual user groups will not be a trivial task. Therefore, the evaluation with the human subjects will remain as a part of the future work. However, with the increasing number of users from diverse language
backgrounds and multilingual contents in the contemporary computing environment, including the WWW on the Internet, it is hoped that the user study and formulating a sample group among the multilingual collaborators becomes a more feasible task than it is now.

9.2.3 Measurement and Evaluation for Other Aspects of the Approach

Evaluation of a complex HCI system can be performed in various aspects of the system. Chapter 8 showed a part of the aspects. The measurement and evaluation for other aspects of the approach, such as evaluation of the VREAD, will also remain as a part of the future study.

9.3 Related Contribution

Master’s thesis

Bibliography


197


Appendix

Referenced Technical Books

XML


Web Services and Server


**Mobile, Wireless, WAP, PocketPC**


**ASP, .NET, C#, and Script**


OpenGL and COM

Database, Oracle, SQL, and ADO


Vita

Namsoon Jung

2137 Quail Run Rd.
State College PA 16801
academy.eyapa@gmail.com

Education:
- Bachelor of Arts in German Language and Literature, Korea University, Seoul, Korea, February 1992
- Master of Arts in Computer Science, City University of New York (Queens College), Flushing, NY, September 1999
- Master of Science in Computer Science and Engineering, Pennsylvania State University, University Park, PA, August 2002
- Thesis Title: Real Time Superimposition of Virtual Objects over the 2D Human Face Image: Facial Enhancement Technology (F.E.T.) System
- Thesis Advisor: Dr. Rajeev Sharma

Related Experience:
- Researcher and Senior Software Engineer, Advanced Interfaces, Inc., State College, PA, September 2000 – August 2005
- Research Assistant, Pennsylvania State University, Computer Science and Engineering Department, University Park, PA, January 2001 - July 2002 and September 2003 - December 2005
- VP of Innovation and Research / Senior Researcher, VideoMining Corporation, State College, PA, January 2006 – September 2012

Awards:
- ‘Mayor's Award’ as the best student of the year for the academic excellence, Anyang City, Korea, 1988
- Departmental scholarship, Korea University, Seoul, Korea, Spring 1989
- 21 US Patents and 29 US Patents Pending, September 2012