Proceedings of the Third IASTED International Conference
Human-Computer Interaction
March 17-19, 2008 Innsbruck, Austria
ISBN Hardcopy: 978-0-88986-725-3 / CD: 978-0-88986-726-0

OPERATING NETWORKED APPLIANCES USING GAZE INFORMATION AND VOICE RECOGNITION

Naohiro YUASA † , Kohei MITSUI † , Hiroki SAKAKIBARA † ,
Hiroshi IGAKI ‡ , Masahide NAKAMURA ‡ and Ken-ichi MATSUMOTO †

† Graduate School of Information Science, Nara Institute of Science and Technology
8916-5 Takayama, Ikoma, Nara, 630-0192 Japan
email: {naohiro-y, kohei-m, hiroki-s, matumoto}@is.naist.jp
‡ Graduate School of Engineering, Kobe University
1-1 Rokkodai, Nada, Kobe, Hyogo, 657-8501 Japan
email: {igaki, masa-n}@cs.kobe-u.ac.jp

ABSTRACT
In the emerging home network systems (HNS), household appliances and sensors are connected to the network to provide various value-added services. However, the cost taken for users to operate the HNS would increase significantly, due to the diversification of appliances and services. Thus, intuitive and simple human interfaces are strongly required. This paper presents a method that operates the networked appliances using user's voice and gaze information together. The conventional voice command interface cannot achieve high precision, when the number of target operations is large. Therefore, the gaze information is used to reduce the number of operations, so that the system handles only appliances that the user is currently gazing at. We have implemented the proposed method and evaluated its effectiveness.

KEY WORDS
home network system, voice recognition, gaze information, multimodal interface, appliances controller, precision

1 Introduction

Along with the progress of ubiquitous network technologies, many objects become networked to provide various communication services. The *home network system* (HNS, for short) is a typical application of such ubiquitous technologies, where various household appliances are connected to a network at home (called, *networked appliances*). Research and development of the HNS are currently a hot topic [1]. There are standard protocols for HNS (e.g., [2][3][4]). Several companies have released commercial products [5] [6]. As the technologies mature in the near future, more and more services and appliances will be developed and provided for home users.

However, the cost taken for users to operate the HNS would increase significantly, due to the diversification of appliances and services. For example, every appliance has a remote controller in general. As the number of appliances increases, it will be troublesome for the user to search a right controller from many, and to find a correct button for the desired operation. Also a new controller with many buttons imposes a significant learning cost to the user. Thus, intuitive and simple human interfaces for the HNS are strongly required.

As for the new human interface for the HNS, Mitsui et al. recently proposed to use the *eye gaze information* for operating networked appliances [7]. However, the gazing appliance can only be used in quite simple operations not mapping complex operations. Also, it is difficult for the system to know exactly if the user is gazing for the operations, or just glancing.

Another typical method is to use the *voice recognition* [6], where the user commands the system via the voice. However, if the numbers of appliances and operations become large, the voice recognition cannot achieve high precision, which may lead to wrong operations. Also, every instance of appliance must be *named*. Even for appliances of the same kind, different names must be assigned (e.g., light1, light2, light3...), which decreases the intuition.

Thus, the conventional methods with the gaze and the voice have certain limitations if they are used separately. However, we consider that there is enough room to combine them to complement their advantage and limitations. In this paper, we propose a multimodal interface that integrates the gaze information and the voice recognition together for more efficient operation of the HNS. In the proposed method, the user can choose appliances to be operated by looking at the appliances. Then, the system dynamically re-constructs and optimizes the voice recognition model for the appliances in the scope. Thus, using the gaze information, the system narrows down the appliances to be operated. This improves the precision of the voice recognition, significantly. Also, the cost for learning every appliance name is reduced. Thus, the gaze is used to improve the precision of the voice recognition, whereas the voice is used to enrich the usability of appliance operations. We have also developed a system called GAVAC to evaluate the proposed method. The experimental results show that the proposed method achieves higher precision and better

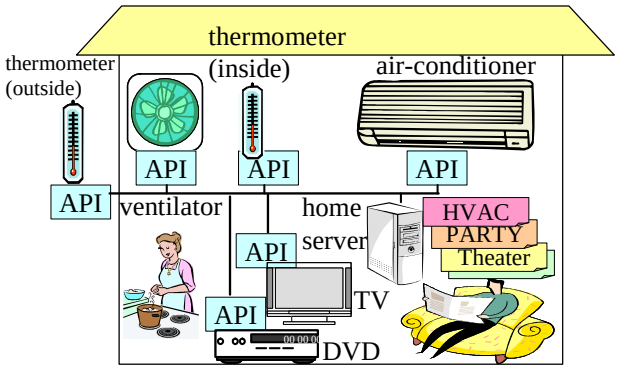

Figure 1. Example of Home Network System

memorability under the environment of many appliances. It is also shown that the precision was robust in noisy circumstances.

2 Preliminaries

2.1 Home Network System

As shown in Figure 1, a home network system (HNS) consists of one or more networked appliances connected to a local area network at home. Every appliance has a set of APIs (Application Program Interface), by which users or external software agents can control the appliance via network. For example, a TV has APIs like ON(), OFF(), selectChannel(), setVolume(), getState(), etc. For convenience, we denote A.m() to represent an API invocation m() of an appliance A (e.g., TV.setVolume(20), Curtain.open(), DVD.getState()). Typically, a HNS has a *home server*, which manages all the appliances in a centralized manner. The appliances may be controlled individually by the user with the remote controller, or by the software applications installed in the home server. Such applications include *integrated services* [8], which control multiple appliances together. For instance, integrating a TV, a DVD player, speakers, lights and curtains would implement *Theater Services*, where a user can watch movies in a theater-like atmosphere through a single point of operation. As the technology matures, it is expected that more and more appliances and services for the HNS will be developed and provided for the home users. The most current HNS technologies are for the ease of machine-to-machine interactions. However, for the feasibility of human operations, intuitive human interface is required, which must be beyond the conventional remote controllers.

2.2 Conventional Human Interface for HNS

According to our review of the conventional human interfaces, we consider that *voice recognition* and *gaze information* are promising means for implementing the intuitive interface of the HNS.

2.2.1 Voice Recognition

As a related work, [9][10] presented a voice command interface for operating electric appliances. For example, if the user says “turn on TV”, then the system interprets the command and the TV is switched on. In general, to improve the precision of the voice recognition, the *voice recognition model* must be tuned for individual applications. For instance, the syntax of the commands and the vocabulary for appliances and operations are registered to the recognition model beforehand. The advantage and limitations of the voice recognition within the context of the HNS interface are summarized as follows:

Advantage:

(a) The user can speak the command even when he/she is occupied with other tasks.

(b) The interface is more intuitive than the remote controller, since a few words can be mapped on a complex task.

Limitations:

(c) Multiple appliances must have different names, even though they are of the same kind (e.g., light1, light2,...).

(d) The precision declines significantly, when the numbers of appliances and operations grow.

(e) The recognition fails under noisy circumstances, especially when operating appliances yield sound and noise.

2.2.2 Gaze Information

Mitsui et al. [7] proposed a system called AXELLA. AXELLA captures the user’s *eye gaze* information with an eye camera, and performs an action for the appliance that the user is currently looking at. The service “See and Know” speaks the current status of every networked appliances, when the user looks at the appliance. We also summarize its advantage and limitations:

Advantage:

(f) Gazing is quite simple and intuitive for selection (or simple operation) tasks for appliances.

Limitations:

(g) Gazing is so simple that it cannot be varied for complex operations.

(h) When appliances are close to each other, it is difficult for the system to recognize which appliance is gazed.

(i) It is difficult for the system to distinguish the gaze for the command from just the *glance*.

3 Using Gaze and Voice for HNS Control

3.1 System Requirements

To cope with the limitations, we integrate the voice and gaze methods to implement efficient interface for the HNS. We first summarize the requirements to be achieved by the proposed method.

Requirement R1: The system must be able to identify the *command mode* clearly, where the user issues the command to the appliances. This is to cope with the limitation (i) for

the gaze interface.

Requirement R2: The interface must have a natural means to distinguish individual appliances (even those in the same kind). This is to eliminate limitation (c) of the voice recognition.

Requirement R3: Even if the numbers of appliances and operations grow, the system must be able to interpret the user's command precisely, specifically. Note that the precision is required for both the selection of an appliance and the execution of an operation.

3.2 Proposed Method

We propose a new human interface system *GAVAC* (GAze and Voice Appliance Controller), which integrates voice recognition and gaze information for efficient operations for the HNS. To meet the requirement, the GAVAC implements the following special features F1, F2 and F3, which are intended to contribute to achieving Requirements R1, R2 and R3, respectively.

F1: Changing Mode Feature

The GAVAC interprets a special word for entering the command mode. Specifically, when the user says "gaze mode", the GAVAC changes the mode to wait for the user's command in the gaze or the voice. Without the word, the GAVAC ignores any gaze and voice, and produces no feedback.

F2: Gaze-based Appliance Selection and Feedback

The GAVAC allows the user to choose the target appliance just by *looking at* the appliance. Hence, we do not need artificial names even for appliances in the same kind. To improve the reliability, the GAVAC feedbacks the selection via voice. Specifically, as the user looks at an appliance, the GAVAC *speaks* which appliance the user is gazing.

F3: Voice-based Operation with Dynamic Model Reconfiguration

Once an appliance is selected by F2, the GAVAC waits for the operation for the appliance via *voice*. As the user says a word, the system tries to interpret the word as an operation for the appliance. For this, the GAVAC dynamically reconstructs the voice recognition model, and optimizes it for the selected appliance only. It significantly improves the precision of the voice recognition.

Figure 2 depicts an overview of the proposed system. We will explain how the user interacts with the GAVAC.

(0) The user says "gaze mode".

(1) The user gazes at an appliance that the user wants to operate. Let the appliance be app.

(2) *Eye Gaze Analyzer (EGA)* captures the gaze information and identifies app.

(3) EGA tells ID of app to *Voice Recognizer (VR)*.

(4) VR reconstructs the voice recognition model for app.

(5) The speech engine responses "You are choosing app, please say operation...".

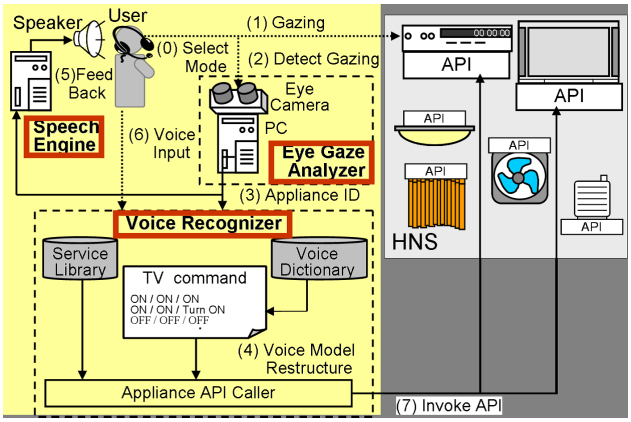

Figure 2. Overall System Architecture of GAVAC

(6) The user tells the desired operation op by voice, e.g. "Turn on". If app is the wrong selection, the user says "None" to go back to (1).

(7) The system interprets op and *Appliance API Caller* invokes the API $app.op$ within the HNS.

3.3 Implementation

We have implemented the GAVAC based on the proposed method. The GAVAC basically consists of three components, (a) *Eye Gaze Analyzer*, (b) *Voice Recognizer* and (c) *Speech Engine*. The speech engine is just a generic Text-to-Speech system, which synthesizes voice from a given text string. So, we omit the detailed explanation for it due to limited space.

3.3.1 Eye Gaze Analyzer (EGA)

The eye gaze analyzer captures the user's eye gaze information. It identifies which appliance the user is currently looking at. To implement the analyzer, we employ the existing face-and-gaze measurement system [7]. The measurement system consists of a three-dimensional stereo camera and a PC. It captures the user's face image, and performs an image processing to identify position of every facial organ. Then, the system calculates the direction and angle of eyes. Finally, the system identifies the object (i.e, appliance) by detecting the intersection of the eye direction and the object based on floor plan data prepared in advance. Then, the name (i.e., identifier) of the appliance is passed to the voice recognizer. To cover a wider range of the gaze measurement, we assume to deploy multiple eye gaze analyzers in the HNS. Figure 3 shows the EGA, where the user is looking at a speaker.

3.3.2 Voice Recognizer (VR)

On receiving an appliance ID (say it app), the voice recognizer (VR) processes the voice command as the operation of app. To achieve the high recognition precision, the VR

Figure 3. Screen Shot Of Eye Gaze Analyzer

```
<RULE name="TV_POWER" toplevel="ACTIVE">
  <L propname="TV_POWER">
    <P valstr="TV.ON">ON;</P>
    <P valstr="TV.ON">TURN ON;</P>
    <P valstr="TV.ON">SWITCH ON;</P>
    <P valstr="TV.OFF">OFF;</P>
    <P valstr="TV.OFF">TURN OFF;</P>
    <P valstr="TV.OFF">SWITCH OFF;</P>
  </L>
</RULE>

<RULE name="TV_VOLUME" toplevel="ACTIVE">
  <L propname="TV_VOLUME">
    <P valstr="TV.VOLUME_UP">VOLUME UP;</P>
    <P valstr="TV.VOLUME_UP">LOUD;</P>
    <P valstr="TV.VOLUME_DOWN">VOLUME DOWN;</P>
    <P valstr="TV.VOLUME_DOWN">LOW;</P>
  </L>
</RULE>
                    ...
```

Figure 4. Voice Recognition Model

dynamically reconstructs the recognition model, optimized for app.

To implement the mechanism, we have employed an open source speech recognition engine, *Julius* [11]. The great advantage of Julius is in its open architecture. In the Julius system, the speech engine itself and the surrounding models (e.g., vocabulary, syntax definitions, acoustic models, etc.) are separately defined. So, the user can easily tune the recognition model for any specific application.

To achieve our goal, we first prepare a *voice dictionary*, in which the operation syntax and vocabulary are defined for *every* appliance. Figure 4 shows an example of the voice dictionary, which is designated for operations of a TV. There are two rules for recognizing operations for power and sound volume of the TV. To achieve the intuitive interactions, multiple vocabulary are defined for a single operation. For instance, the voice "On", "turn on", "switch on" are all assigned for the operation TV.ON. Next, when the VR receives user's voice successively, it interprets the voice based on the current model. An operation op for app is derived from the model. Finally, the VR invokes the API $app.op()$ within the HNS.

Let us recall the above example. If a user looks at the TV, the EGA captures the gaze and sends the ID of the TV to the VR. Then the VR updates the model with the dictionary in Figure 4. If the user says "switch on", the voice is bound with the operation TV.ON. Finally, the API TV.ON() is invoked. Note that in this example saying these words affects the TV only, since the recognition models for any other appliances are overwritten in this moment. Thus, we can keep the good precision even if the number of appliances grows.

3.3.3 Deploying GAVAC in Home Network

We have deployed the GAVAC in the existing home network system, called *NAIST-HNS* [12]. The NAIST-HNS consists of legacy household appliances, each of which is networked with the technology of the Web services.

We have developed a testbed of the GAVAC by connecting it to the NAIST-HNS. The testbed consists of the following appliances and devices.

- Eye camera for EGA: Pointgrey Research – Frea x 2
- Bluetooth Wirelss Microphone : Princeton PTM-BEM3
- Voice Recognition Engine : Julius for SAPI ver 2.3
- TV: NEC PX-50XM2
- DVD Recorder : Toshiba RF-XS46
- 5.1ch Surround Speeker: Pioneer HTZ-535DV
- Electronic Curtain : Navio PowerTrack RS-555CT
- Light : Kishima KFF-1708
- AirCirculator : MORITA DENKO MCF-257NR

4 Experiments

4.1 Experiments Overview

In the experiments, we evaluated the proposed method from two viewpoints: *precision* and *usability*. Using the testbed implemented, we let subjects operate the networked appliances with the conventional method (voice only) and with the proposed method (gaze and voice). To examine the effect of noise on the voice recognition, we made two environment settings: *silent* (average 8dB) and *noisy* (42dB).

Within the testbed, we instructed subjects to perform certain tasks. For each task, we prepared a detailed manual, describing how to operate the appliance with the GAVAC and the voice interface. All the operations for the appliances were logged, and activities of the subjects were taped for the subsequent analysis. The order of tasks for each subject were balanced using the Latin square, to minimize the learning effects caused by individual difference. After completion of all tasks, we asked each subject to answer a questionnaire to evaluate the usability.

4.2 Evaluation Metrics

Precision

The precision refers to the degree of how accurately the

system recognized the user's commands. Specifically, we have defined the metrics as follows: the proportion of the number of operations recognized correctly to the total number of gazes or voices issued during the command mode. The criteria of failed recognitions were defined as follows:

- The system selected wrong appliances in the gaze-based appliance selection.
- The system could not capture the voice since the user's voice was too quiet.
- The system recognized wrong appliances or other operations than what the user said.

Usability
To evaluate the usability, we have chosen the following five metrics from the reference book [13].
Learnability: so that the user can rapidly begin working with the system.
Efficiency of Use: enabling a user who has learned the system to attain a high level of productivity.
Memorability: allowing the casual user to return to the system after a period of non-use without having to re-learn everything.
Few and Noncatastrophic Errors: low error rate, so that users make fewer and easily rectifiable errors while using the system. Further, catastrophic errors must not occur; and finally.
Subjective Satisfaction: pleasant to use, so that users are subjectively satisfied when using it.

For the conventional and the proposed methods, each of the above metrics was evaluated by the 4-point Likert scale: 4:High, 3:A little high, 2:A little low, 1:Low.

4.3 Subjects and Tasks

A total of 10 graduate students participated in the experiment. They were all in their 20's. The subjects were used to operating electrical appliances at home. We prepared the tasks which were quite common operations in daily life. The following shows a typical task assigned for a subject:
Task with the conventional interface (voice only)
(1) Say "TV Turn On".
(2) Say "TV Volume Down".
(3) Say "DVD Stop".
(4) Say "Curtain Close".
(5) Say "Air Circulator Medium".
(6) Say "Right-side light Turn On".
Task with the proposed method (voice and gaze)
(1) Gaze at TV and say "Turn On".
(2) Gaze at TV and say "Volume Down"
(3) Gaze at DVD and say "Stop"
(4) Gaze at Curtain and say "Close"
(5) Gaze at AirCirculator Gaze say "Medium"
(6) Gaze at Right-side light and say "Turn On"

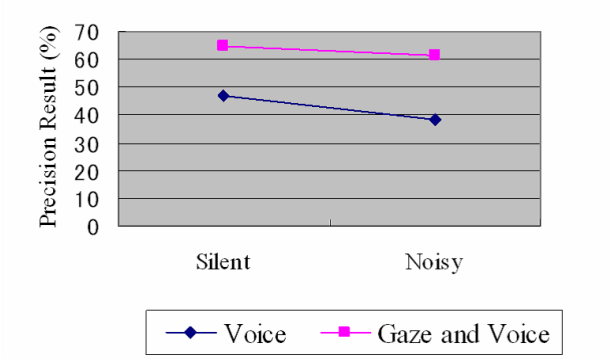

Figure 5. Precision Result

4.4 Results

Figure 5 shows the result of the precision. The horizontal axis plots the values under the different environment settings. The vertical axis represents the average value of the precision of all subjects.

In the silent environment setting, the average precision was 46.7% for the conventional method (cases varied from 16.7% to 83.3%). On the other hand, the average for the proposed method was 65.0% (cases varied from 33.3% to 83.3%). In the noisy setting, the average precision of the conventional method declined to 38.3%, while the proposed method achieved 61.7%. Thus, it can be seen that the proposed method achieved better precision than the conventional method, and that the proposed method was more *robust* for the noise. This is because of the dynamic model reconfiguration of the proposed method. For every time the user gazes at an appliance, the voice recognition model is optimized for the appliance. Therefore, there was less room for the noise to be misrecognized as the command.

As for the usability evaluation with the questionnaire, we found no significant difference between the conventional and the proposed methods. For both methods, the average value for each of the five metrics was just around 2.0 point, which was "*A little Low*". The proposed method was a little bit superior with respect to the memorability, since the subjects did not have to remember the names of all appliances. Through the interview after the questionnaire, the main reason was that the subjects were not used to handle these emerging interfaces efficiently. Things may change as the subjects get familiar with the interface. However, there should be rooms to enhance the usability of GAVAC as well.

4.5 Discussion

We here discuss how the limitations of the conventional methods are addressed by the proposed method. Table1 summarizes the comparison among the two conventional methods and the proposed method, with respect to the six limitations mentioned in Section 2.2. In the table,

Table 1. Comparison of Methods

Conventional Limitations	Voice	Gaze	Proposed
(c) Each Distinction	△1	◯	◯
(d) Precision for Many Apps	×	◯	◯
(e) Noisy Circumstance	×	◯	△2
(g) Detailed Operation	◯	×	◯
(h) Neighbored Appliances	◯	×	△3
(i) Operation Intent	◯	×	◯

◯, × and △ respectively represent "strong", "weak" and "medium" characteristics. As was supposed in the design principle, the proposed method would complement the limitations of the conventional methods. Each item with △ means that the method would be feasible with some limitations, explained as follows.

△1:The voice recognition can distinguish multiple appliances individually by putting a unique name to every appliance. However, the intuition is decreased and the learning cost is increased.

△2:In noisy environment, the recognition rate of the proposed method would decline somewhat, since the voice recognition is used. However, the proposed method is yet more robust than the voice only, since the dynamic model reconstruction can reduce the number of words to be recognized.

△3:The appliance selection with the gaze interface may choose a wrong appliance, if multiple appliances are close to each other. However, the proposed method provides the user a means to confirm which appliance is currently gazed, by using the speech response.

5 Conclusion

In this paper, we presented a human interface that integrates the voice recognition and the gaze information for efficient operation of networked appliances. Based on requirements, we have implemented the system called GAVAC, and evaluated the feasibility through experiments. Finally, we summarize our future work as follows. The precision of the EGA should be enhanced. In the experiment, we observed a few cases that the system cannot recognize where the user is now gazing, despite the user gazing some appliance. The reason was that the user's face moved beyond the range of calculation. As a result, the EGA went out of the ready state, and didn't recover autonomously. We need features that can detect *continuous non-ready states* and recover from them autonomously. Additionally, integrating multiple EGAs for capturing the eye gaze from multiple direction would improve the precision.

Acknowledgment

This research was partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Young Scientists (B) (No. 18700062), Start Up (No.18800060), and Scientific Research (B) (No. 17300007), and by JSPS and MAE under the Japan-France Integrated Action Program (SAKURA).

References

[1] C. Hsiang and R. Hong, System integration of WAP and SMS for home network system, *Computer Networks*, 42(4), 2003, 493-502

[2] Digital Living Network Alliance http://www.dlna.org

[3] UPnP Forum http://www.upnp.org

[4] ECHONET Consortium, http://www.echonet.or.jp

[5] Hitachi Appliances, Inc., Horaso Network Service, http://www.hitachi-ap.co.jp/

[6] Bandai Robot Laboratory, Kikino, http://www.roboken.channel.or.jp/

[7] K. Mitsui, H. Igaki, K. Takemura, M. Nakamura, and K. Matsumoto, Exploiting Eye Gaze Information for Operating Services in Home Network System, *2006 International Symposium on Ubiquitous Computing Systems (UCS2006)*, Seoul, Korea, 2006, 13-27.

[8] H. Igaki, M. Nakamura and K. Matsumoto, A Service-Oriented Framework for Networked Appliances to Achieve Appliance Interoperability and Evolution in Home Network System, *Proc. of International Workshop on Principles of Software Evolution (IWPSE 2005)*, Lisbon, Portugal, 2005, 61-64.

[9] S. Yamamoto, J.-M. Valin, K. Nakadai, H. Tsujino, J. Rouat, F. Michaud, T. Ogata, K. Komatani and H. G. Okuno, Enhanced Robot Speech Recognition Based on Microphone Array Source Separation and Missing Feature Theory, *Proc. IEEE-RAS International Conference on Robots and Automation (ICRA-2005)*, Barcelona, Spain, 2005, 1489-1494.

[10] K. Ishiwatari, S. Wakisaka, K. Ito, T. Toge and M. Tanaka, Recognition Dictionary System Structure and Changeover Method of Speech Recognition System for Car Navigation, *United States Patent*, 2000, 6112174

[11] Open-Source Large Vocabulary CSR Engine – Julius, http://julius.sourceforge.jp/en_index.php

[12] M. Nakamura, A. Tanaka, H. Igaki, H. Tamada, and K. Matsumoto, Adapting Legacy Home Appliances to Home Network Systems Using Web Services, *Proc. of International Conference on Web Services (ICWS2006)*, Chicago, USA, 2006, 849-858.

[13] J. Nielsen, *Usability Engineering* (San Francisco: Morgan Kaufmann, 1994)