Design of Physically Grounded Communication System

実世界指向コミュニケーション特論

2006 年度

Michita Imai

今井　倫太

1.Introduction

1.1 Topics

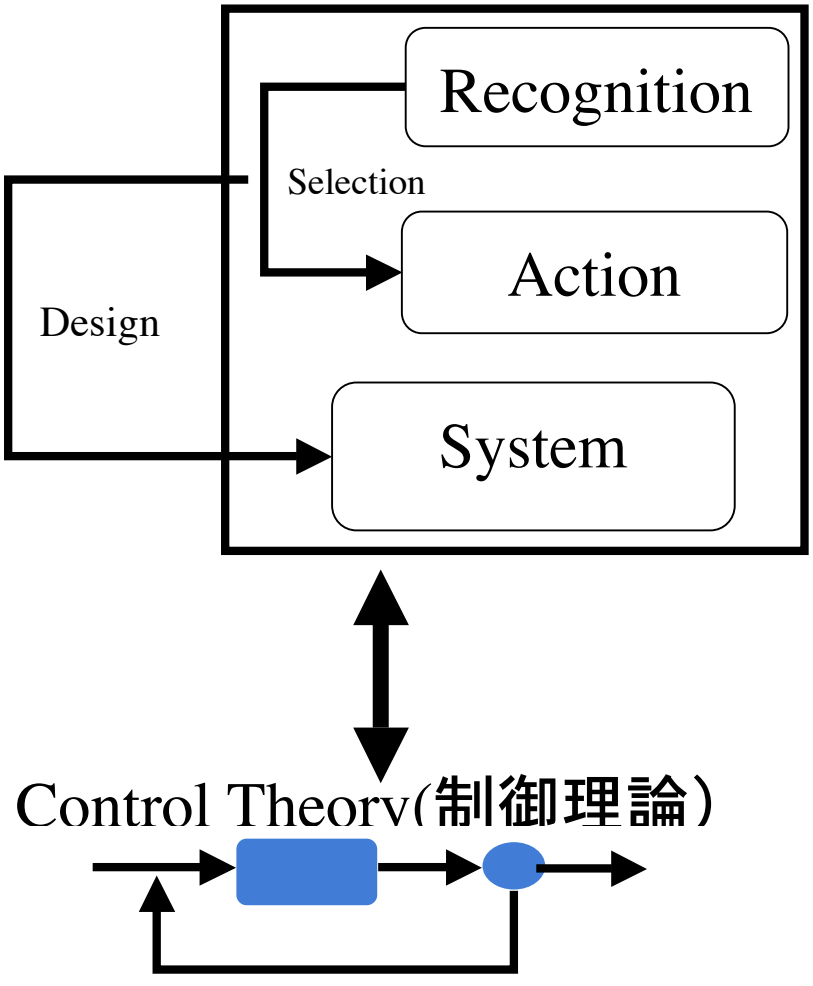
- Systems dealing with real world interaction
- Communication mechanisms from person's viewpoint
  - Psychological aspect
- How to develop a system dealing with the psychological aspect of the interaction or dynamics of human-machine interaction
  - Computation model and strategy

1.2 Target of This Lecture

- For designing a system interacting with people
  - Communication Robot
  - Intelligent Room
  - Multimodal-Interface
  - All intelligent human-interface

1.3. Contents of Lecture

Design of Communication System



- System Structure
  - Vertical System
  - Horizontal System
- Contents (human-factor)
  - Sociality of System (社会性)
  - Joint Attention (共同注意)
  - Theory of Mind (心の理論)
- Effect of Contents
  - Ecological Approach (生態学的アプローチ)

The list of actual contents which this lecture explains.

1. Vertical System: Systems in virtual world:
   - SHRDLU
2. Horizontal System:
   - Subsumption Architecture for Robustness of Robot
   - Systems dealing with real world interaction
3. Joint Attention:
   - Multi-modal Interface and Attention Mechanism and Joint Attention and Communication Robot
4. Theory of Mind Model
5. Sociality of System:
   - Infant Robot and Social Agent
6. Ecological Approach

1.4 Feature of Real World

- Events occur regardless of systems' behaviors
  - Predictable event vs. Unpredictable event
- Unpredictability of human's behavior
  - Difficulty to make model
    - Difference among individuals
  - Effect of psychological state
  - Absence of attention toward system

1.5 Contrast between Real World and Virtual World

- Real world
  - Infinite numbers of entities exist regardless of systems.

- Uncertainty: someone or something may induce an event anytime.

・ Virtual world
  - Entities are given when a system is developed.
  - Almost events are predictable.
  - Only user's behavior is unpredictable.

1.6 Example Problems of Real World

・ Fault in planned actions
・ Insect behavior for dynamic world
・ Reference to an object in a conversation
・ Mind-reading in a conversation
・ The effect of Sociality
・ Immersion (没入) in interaction
・ Difficulty to deal with physical objects
・ Ecological Approach

1.6.1 Fault in planned actions

・ Planner software generates a sequence toward a goal state prior to the execution of a system.
・ Generated plan cannot be used under dynamic world.
・ Planner is used to generate intermediate plan in actual systems.

The following explanation is an example of the fault in planning actions.
Imagine that there are four blocks B, C, R, and Y on the table. What plan is generated when a human ask a arm robot to put R on the other block.
The initial state of the block world is expressed as the following with lisp expressions.
(ON B TABLE) (ON Y R) (ON R TABLE) (ON C TABLE)
The following sequence of robot's actions leads to the goal situation.
(HOLD Y) (PUT Y TABLE) (HOLD R) (PUT R B)
The state of the blocks changes like the following by the actions.
(ON B TABLE) (ON R B) (ON Y TALBE) (ON C TABLE)
However, if B is taken away by someone before the robot carries out (PUT R B), what happens? Planner must generate alternative plan.

1.6.2 Insect behavior such as ants

・ Vast behaviors are selected in response to environmental events.

- Easy to adapt to dynamic environment
- Swarm intelligence (群知能) emerges.
- It cannot select actions intentionally.

However, Intentional behavior is important for communication. Model based control mechanism required to enable the intelligence to behave intentionally.

An ant path is one of example of swarm intelligence. Imagine there are two ways A and B from the nest of ants to food. Also, A is shorter than B. When ants move, they leave pheromones along their paths. And the pheromones attract other ants.

Since it takes a shorter time to go via A, the density of the pheromones increases on the path A. As a result, path A emerges.

However, if scenery along B is beautiful and the ants like beautiful thing, they must select path B intentionally to see the sight. Unfortunately, it is impossible for them to go along B as far as they obey the procedure.

1.6.3 Reference to an object in a conversation

- Reference can be achieved under cooperation between a speaker and a hearer.
- Person's attention
- Joint attention
- Difficult for the system to identify targets.

A speaker and a hearer must pay attention to the same situation to understand an utterance of the speaker. This phenomenon is called joint attention in the social psychology. For example, if a speaker says "this is cool!" pointing to a car, a hearer must interpret the word "this" by noticing what the speaker refers to. However, since it is difficult for a system to notice a target where a human directs his/her attention, the establishment of joint attention is significant theme in interaction in the real world.

1.6.4 Mind-reading in a conversation

- Mind-reading in communication
- Human's action and utterance include hidden intention to take the action or to give the utterance.
- Human does not always read other's mind.
- What trigger these phenomena?

Mind-read is also important factor when we consider a communication in the real

world. Humans always read other's intention (other's mind) when communicating. For example, why did the other say so, why did he/she behave so. For example, imagine a person who carries heavy burdens tries to go through a closed door. Also, there is another person who is his friend. At the situation, the person will open the door even though he does not say anything. This scene occurs because the person reads the intention of him.

However, mind-reading does not always happen. It needs a relationship between humans when someone reads the other's intention. The fact indicates that if a system makes a person read the intention of the system, it must establish a relationship with the person.

1.6.5 The effect of Sociality

- Sociality must be also considered to develop an interactive system.
- How to give a robot social appearance?

The sociality of a system must be considered when we design a communication system. Sociality means not only how to give a system a skill to interact with humans socially but also how to make humans deal with the system socially in interacting with it. I will give an example of a social effect. The example system is a dialogue system which ask a human's name at first.

System: Welcome, sir!
System: Could you input your name?
Human: My name is Imai.
System: Mr. Imai, I'm glad to navigate…….

However, if this system is set at a public space, the following interaction frequently occurs.

System: Welcome, sir!
System: Could you input your name?
Human: My name is Thief.
System: Mr. Thief, I'm glad to navigate…….

When this type of interaction occurs, all people think the system is incomplete. However, does the fault depend on only the ability of dialogue system? Consider carefully human-human interaction. There is no person who says my name is thief to a person who he/she meets for the first time. The phenomenon indicates that the sociality of a system is too low to interact in human's society. Giving sociality to a system is also one of significant theme in a communication system in the real world.

1.6.6 Immersion (没入) in interaction

- Immersion in interaction
  - Mind-reading
  - Joint Attention
- What factor makes a person immerse him/herself in interaction?

When a human interacts with a system or a robot, he/she does not always immerse him/herself in the interaction. Some people observe its behaviors to know what function it has. Because of the observation, they frequently ignore messages from the system. To establish a communication with humans, a system or a robot must have a mechanism to engage them in the interaction.

1.6.7 Difficulty to deal with physical objects

- The state of physical object is obscure from system's viewpoint.
  - Ownership of objects
    - Easy to identify ownership of files in a computer system
    - Difficult to identify it in real world
  - Person's intention toward objects

In the real world, it is difficult for a system to deal with the obscure features of objects. Ownership is an example of the obscure feature. Computer system easily manages the ownership of files. However, the ownership of the objects in the real world cannot be identified from the system because mind-reading is required to identify it.

1.6.8 Ecological Approach

- Ecological Approach
  - When you drive a car backward, relationship between movement and recognition provide significant information to control the car.
- No formal way to implement the approach

Biological systems such as humans, other animals, insects are said that they have the other type of behavior model to deal with the complexity of the real world. One of proposed models is ecological approach. The approach consists of a connection between actions and recognitions. For example, when you drive a car backward, you are not conscious of the angle of a handle. You notice the direction of the car by moving it. The

control is established on the recognition in the motion. We can see the same control strategies in behaviors of living beings.

1.6.9 Example Problems of Real World

- The some of the problems are not solved.
- They are important for the design of the system in real world.

2. Intelligent system in virtual world

- System dealing with toy world
  - Extension of thought experiment
  - Static world
  - Lack of effect of time
  - Sufficient to solve a problem logically
  - Interaction depends on discrete turn-taking.
  - It cannot deal with dynamic changes.
- Vertical System was frequently used for the system used in virtual world.
  - R. Brooks named it.
  - Each module is executed in response to the execution results of former module.
  - The output is determined through all modules.

2.1 SHRDLU

- 1968, T. Winograd developed it.
- Dialogue system on a virtual world (block world)

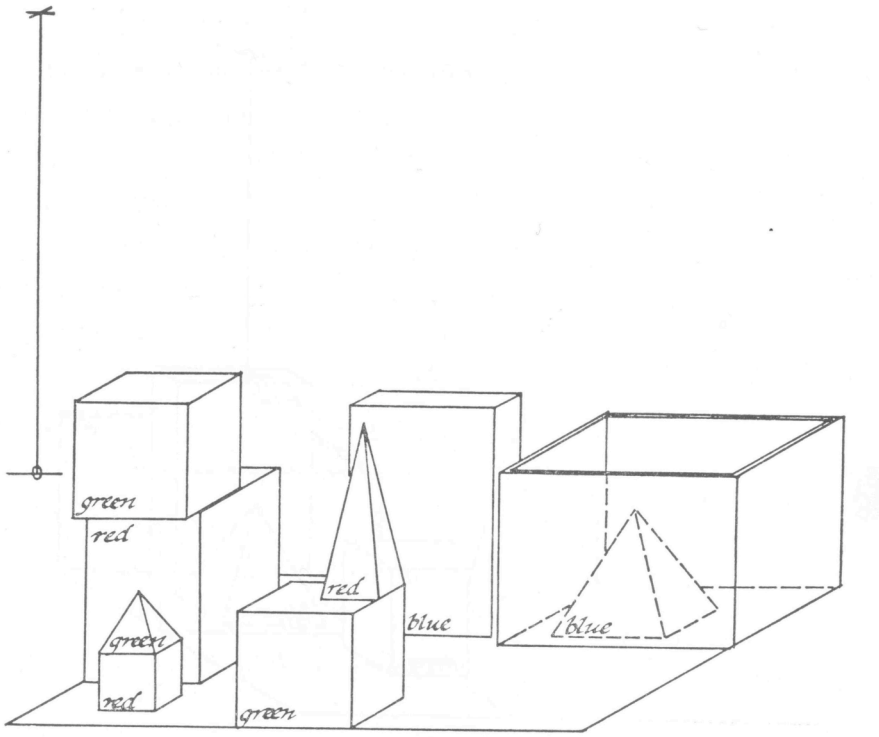
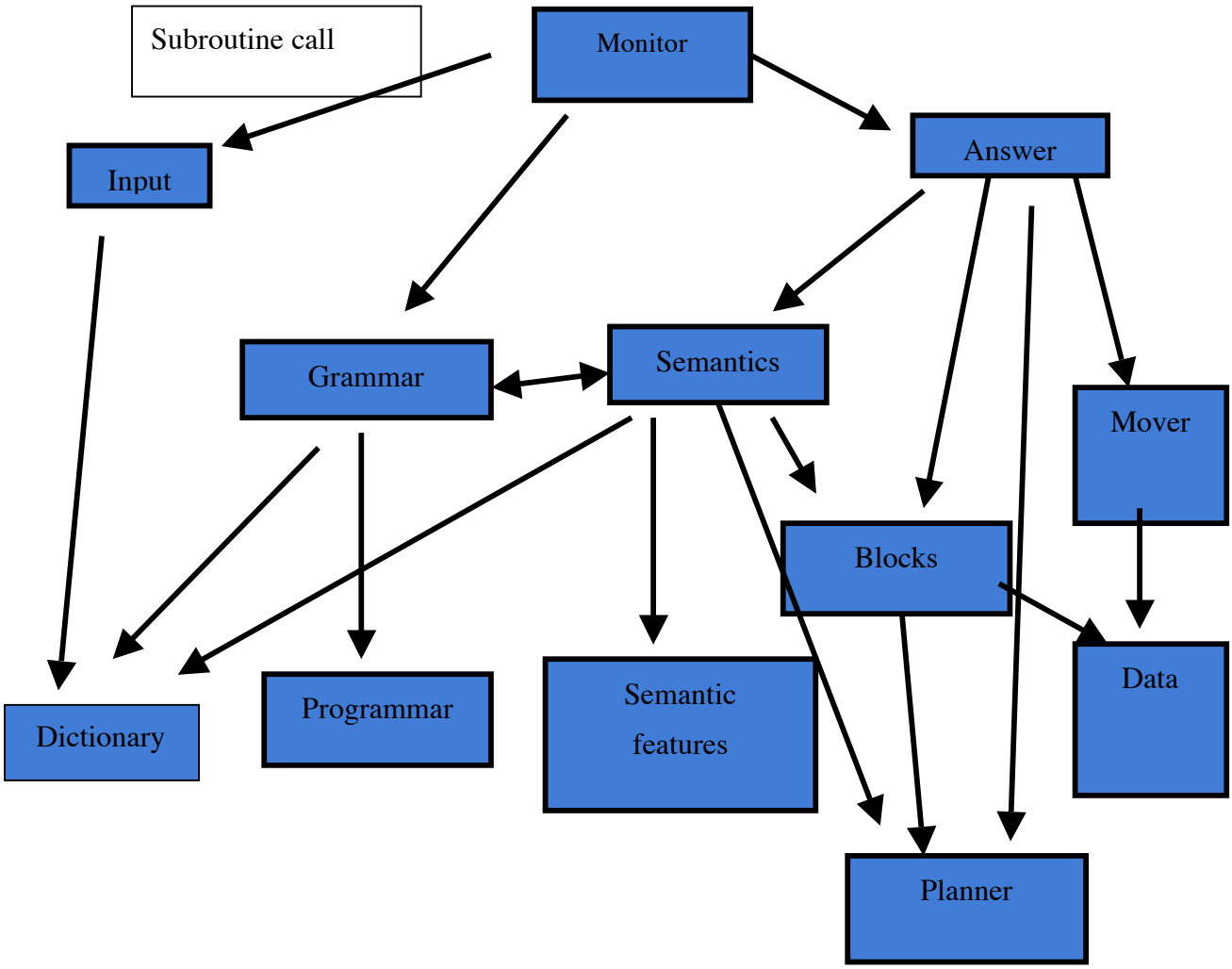


図 3 "Pick up a big red block."

2.2 Overview of SHRDLU



- Monitor
  - Main routine which calls main routines: Input, Grammar, Semantics, and Answer.
- Input
  - This routine obtains text string typed in English and carries out morphological analysis.
    - running -> run + ing
- Grammar
  - This routine analyzes the input texts with grammatical information and generates a parse tree.
  - Example of grammar
    - SENTENCE →NP, VP
    - VP→VERB,NP | VP and VP
    - NP→DET, NOUN | NOUN
    - NOUN →NOUN and NOUN

- He ate the apples and peaches.
- He ate the apples and drank the beer.

- Semantics
  - This routine interprets the input texts cooperating with Grammar and infers an appropriate meaning with Planner.
  - All humans are foolish.
    - (For ALL (X) (IMPLIES (#HUMAN X) (#FOOLISH X))
  - A thesis is acceptable if either it is long or it contains a persuasive argment.
    - (For ALL (X) (IMPLIES
      (AND (#THSIS X)
      (OR (#LONG X)
      (EXISTS (Y)
      (AND (#PERSUASIVE Y)
      (#ARGUMENT Y)
      (#CONTAINS X Y)))))
      (#ACCEPTABLE X)))
  - Is Sam's thesis acceptable?
    - (#ACCEPTABLE :SAM-THESIS)
    - Theorem proving
  - Pick up the block and put it in the box.
    - (AND (GOAL (#PICKUP :BLOCK23))
      (GOAL (#PUTIN :BOLOCK23 :BOX7)))
  - Find red blocks.
    - (FIND ALL $?X
      (GOAL (#BLOCK $?X))
      (GOAL (#COLOR $?X RED)))

- Answer
  - This routine controls the responses of SHRDLU by generating a natural language expression and storing conversation contexts.
- Programmar
  - This routine is a set of parser programs executed by Grammar. The programs produce a parse tree.
- Dictionary
  - This routine has the definitions of words used by Input, Grammar, and Semantics.
- Semantics features

- This routine has the feature lists of the block world. The features are used to categorize items in the block world and the action of the robot. Ex. “in”
- How many blocks are in the box?
- The red block in the tall stack

• Blocks
- This routine is the knowledge of the block world. Physical constraints.
- Example of expressions in SHRDLU
  - (#SUPPORT :B2 :B1)
  - (#SIZE :B5 (100 100 300))
  - (#COLOR :B5 #WHITE)
  - (#SHAPE :B5 #ROUNDED)

• Mover
- This routine shows the block world and the robot on a computer display.
- Example of the expression of an action.
  - (GOAL (#MOVEHAND (600 200 300))

• Planner
- This routine infers relations between the items in the block world to generate information for parsing texts and to confirm the statement of the text.
- Example
  - Turing is a human.
  - All humans are foolish. → Turing is foolish.
- Example of the inference on SHRDLU
  - Insert knowledge to DB.
    - (ASSERT (#HUMAN :TURING))
    - (DEF-THEOREM
      ((CONSEQUENT (X) (#FOOLISH $?X)
      (GOAL (#HUMAN $?X)))))
  - Question1: Is Turing foolish?
    - (GOAL (#FOOLISH :TRUING))
  - SHRDLU searches on DB for the statement. → Fail
  - It uses theorem to prove the statement.
  - Question2: Is anything foolish?
    - (GOAL (#FOOLISH $?Y))

• Data
- This routine keeps information about a current state of the world.

SHRDLU has the following knowledge.

- Current situation: Knowledge of the world.
- Context : Knowledge of the conversation
- Grammar: Lexical Knowledge
- Semantics: Knowledge of physical constraints

Recognition skills of SHRDLU

- It is able to identify the size, figure, color, and position of the items in the block world.
- The feature of the virtual world: it is easy for a system to obtain the information in the virtual world.

2.3 Terminology used by SHRDLU

- Objects used by SHRDLU
  (:SHRDLU) The name of the robot (:HAND) The hand or the robot
  (:FRIND) Human (:BOX) A box which can contain something.
  :B1, :B2, :B3, Objects
- The conceptual structure of the world

```
|--#PHYSOB-------------------------- |--#TABLE
|--#ROBOT                           |--#BOX        |--#BLOCK
|--#PERSON                          |--#MANIP---   |--#BALL
|--#PROPERTY---  |--#COLOR     |--#HAND      |--#PYLAMID
                 |--#SHAPE     |--#STACK
```

#PHYSOB denotes PHYsical OBject

#MANIP denotes an object which a robot can move.

- Examples of object types
  (#IS :SHRDLU #ROBOT) (#IS :TABLE #TABLE) (#IS :HAND #HAND)
  (#IS :BOX #BOX) (IS :FRIEND #PERSON) (#IS :B5 #BLOCK) (#IS :B5 #MANIP)
- Examples of physical properties
  (#COLOR :BOX #WHITE) {#BLACK, #RED,#GREEN,#BLUE}
  (#SHAPE :B5 #POINTED){#ROUNDED, #RECTANGULAR, #BLOCK}
  (#AT :B5 (400 600 200))
  (#SIZE :B5 (100 100 300))
  (#CONTAIN :BOX :B5)
  (#GRASPING :SHRDLU :B2)
  (#SUPPORT :B1 :B2) (#SUPPORT :B2 :B3)

(#ON :B3 :B1)

#HEIGHT, #WIDTH, and #LENGH of an object are calculated from its size.

- :B1 is shorter than :B2 is expressed as (#MORE #HEIGHT :B2 :B1)
- Greater than or equal is expressed as #MORE,#ASMUCH

・ Variations of actions

(#MOVEHAND (X Y Z)) (#UNGRASP) (#GRASP X) (#PUT W (X Y Z))
(#RAISEHAND) (#PICKUP X) (#PUTON X Y) (#PUTIN X Y)
(#GET-RID-OF X) (#CLEAR-TOP X) ($STACKUP (X Y …))
(#FINDSPACE A (X Y Z) B $ _C)

This rule finds space B on A. When SHRDLU must left a space (X Y Z), (X Y Z) must be set.

2.4 The definitions of action rules

・ (DEF (#PICKUP $?X)
((GOAL (#MANIP $?X))
(COND ((GOAL (#GRASPING $?X)))
((GOAL (#GRASPING $ _Y))
(GOAL (#GET-RID-OF $?Y)))
(T))
(GOAL (#CLEARTOP $?X))
(GOAL (#MOVEHAND $?X))
(ASSERT (#GRASPING $?X))))

COND is the set of conditional branches. In the following example, if CONDITION1 is true, program01, 02, … are executed and COND returns value of the result the value of the last program. If it is false, CONDITION2 is evaluated. If CONDITION2 is true, program11,12… are executed.

(COND ((CONDITION1) Program01 Program02,….)
((CONDITION2) Program11 Program12…..)
…..
((CONDITIONn) programn1 Programn2…))

・ (DEF (#GRASP $?X)
((GOAL (#MANIP $?X))
(COND ((GOAL (#GRASPING $?X)))
((GOAL (#GRASPING $_Y))
(GOAL (#GET-RID-OF $?Y)))

```
                (T))
        (GOAL (#CLEARTOP    $?X))
        (GOAL (#MOVEHAND    $?X))
        (ASSERT (#GRASPING $?X))))
```

・
```
(DEF (#GET-RID-OF $?X)
        ((OR (GOAL (#PUTON $?X :TABLE))
             (GOAL (#PUTON $?X $?Y)))))
```

・
```
(DEF (#CLEARTOP $?X)
        ((COND ((GOAL (#SUPPORT $?X $_Y))
                  (GOAL (#GET-RID-OF $?Y)))
               ((ASSERT (#CLEARTOP $?X))))))
```

・
```
(DEF (#PUTON $?X $?Y)
        ((NOT (EQ $?X $?Y))
         (GOAL (#FINDSPACE $?Y $E (SIZE $?X) $_Z))
         (GOAL (#PUT $?X $?Z))))
```

・
```
(DEF (#PUT $?X    $?Y)
        ((CLEAR $?Y (SIZE $?X) $?X)
         (GOAL (#GRASP $?X))
         (GOAL (#MOVEHAND $?Y))
         (GOAL (#UNGRASP))))
```

2.5 Example of interaction on SHRDLU

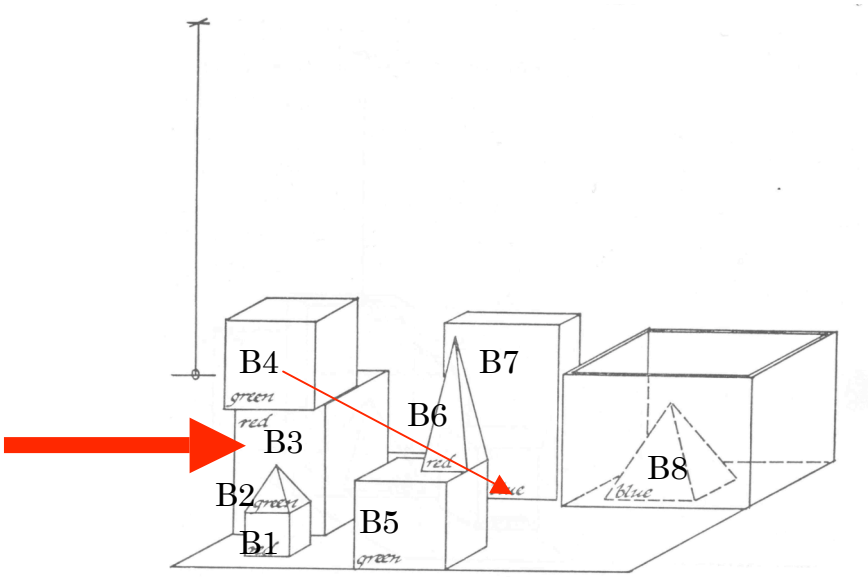


図 3 "Pick up a big red block."

Exaple1.

F: Pick up a big red block

S: OK

The contents of Database at this situation

Initial State of DB

(#IS :B1 #BLOCK)

(#IS :B2 #PYRAMID)

(#IS :B3 #BLOCK)

*

*

(#IS :B7 #BLOCK)

(#IS :B8 #PYRAMID)

(#COLOR :B1 #RED)

(#COLOR :B2 #GREEN)

(#COLOR :B3 #RED)

*

*

(#COLOR :B7 #BLUE)

(#COLOR :B8 #BLUE)

*

*

(#MANIP :B3)

*

*

(#MANIP :B7)

*

*

(#SUPPORT :B3 :B4)

*

*

(#CONTAIN :BOX :B8)

(#SUPPORT :BOX :B8)

(#CLEARTOP :B4)

(#CLEARTOP :B7)

The command “pick up a big red block” is expressed as the following,

```
(GOAL (FIND ALL $?X $?Y
            (GOAL (#IS $?X #BLOCK))
            (GOAL (#COLOR $?X #RED))
            (NOT (AND (GOAL (#IS $?Y #BLOCK))
                      (GOAL (#COLOR $?Y #RED))
                      (GOAL (#MORE #SIZE,#ASUMCH Y X))))))
(GOAL (#PICKUUP $?X))
```

SHRDLU finds :B3 as the target object. Then, it carries out the following command.

(GOAL (#PICKUP :B3))

From the definition of the #PICKUP, the following sequences are carried out.

```
(#PICKUP :B3)
    (GOAL (#MANIP :B3)) → TRUE
    (COND ((GOAL (#GRASPING :B3))) → FALSE
           ((GOAL (#GRASPING $ _Y)) → FALSE
             (GOAL (#GET-RID-OF $?Y)))→skip
           (T))
              (GOAL (#CLEARTOP   :B3)) → Check DB. If it is not,
                                                  #CLEARTOP is executed
              (GOAL (#MOVEHAND   :B3))→ execued after #CLEARTOP
              (ASSERT (#GRASPING :B3))))→It inserts this fact to DB.
```

Since the initial DB does not include (#CLEARTOP :B3), the rule of (#CLEARTOP :B3) is carries out. The following sequences is the execution of #CLEARTOP.

(#CLEARTOP :B3)

(COND ((GOAL (#SUPPORT :B3 $_Y)) → It finds :B4.

(GOAL (#GET-RID-OF $?Y))) → It tries to get rid of :B4

((ASSERT (#CLEARTOP :B3))))

The following is the sequences of #GET-RID-OF.

(#GET-RID-OF :B4)

(OR (GOAL (#PUT ON :B4 :TALBE))→The command is executed.

(GOAL (#PUTON :B4 $?Y)))

The following is the sequences of #PUTON.

(#PUTON :B4 :TABLE)

(NOT (EQ :B4 :TABLE)) → TRUE

(GOAL (#FINDSPACE :TABLE $E (SIZE :B4) $_Z)) → $?Z is set to the target position.

(GOAL (#PUT :B4 $?Z))) → Then executed

The following is the sequences of #PUT.

(PUT :B4 $?Z)

(#CLEAR $?Z (SIZE :B4) :B4) → Confirms the target location.

(GOAL (#GRASP :B4)) → Then executed.

(GOAL (#MOVEHAND $?Z)) → Moves the robot hand to $?Z

(GOAL (#UNGRASP)) → Releases its hold. Also, delete #GRASPING from DB.

The following is the sequences of #GRASP.

(#GRASP :B4)

(GOAL (#MANIP :B4))→TRUE

(COND ((GOAL (#GRASPING :B4)))→FALSE

((GOAL (#GRASPING $_Y))→FALSE

(GOAL (#GET-RID-OF $?Y)))→skip

(T))

(GOAL (#CLEARTOP :B4))→Checking DB.

(GOAL (#MOVEHAND :B4))→Moves hand to :B4.

(ASSERT (#GRASPING :B4))→Grasps :B4.

The result of the sequence of commands consists of

(#PIKUP :B3)

(#CLEARTOP :B3)

(#GET-RID-OF :B4)

(#PUTON :B4 :TABLE)
(#PUT :B4 :TABLE)
(#GRASP :B4)
(#MOVEHAND :B4)
(#GRASPING :B4)
(#MOVEHAND :TABLE)
(#UNGRASP)
(#MOVEHAND :B3)
(#GRASPING :B3).

Throughout the sequences, “pick up a big red block” is executed. After the sequences, the database changes as the following,

(#IS :B1 #BLOCK)
(#IS :B2 #PYRAMID)
(#IS :B3 #BLOCK)
*
*
(#IS :B7 #BLOCK)
(#IS :B8 #PYRAMID)
(#COLOR :B1 #RED)
(#COLOR :B2 #GREEN)
(#COLOR :B3 #RED)
*
*
(#COLOR :B7 #BLUE)
(#COLOR :B8 #BLUE)
*
*
(#MANIP :B3)
*
*
(#MANIP :B7)
*
*
(#CONTAIN :BOX :B8)
(#SUPPORT :BOX :B8)
(#CLEARTOP :B4)

(#CLEARTOP :B7)
(#CLEARTOP :B3)→new one
(#GRASIPING :B3)→new one

Exercise 1.1 The contents of DB change several times throughout the execution of each command. Write down the changes in the contents and follow the behaviors.

Example 2.
F: Find a block which is taller than the one you are holding and put it into the box.
S: But "IT," I ASSUME YOU MEAN THE BLOCK WHICH IS TALLER THAN THE ONE I AM HOLDING. OK

The question of the human is expressed as the following,
(GOAL (FIND ALL $?X
(GOAL (#IS $?X #BLOCK))
(GOAL (#GRASPING $_Y))→It finds :B3 which is a big red block.
(GOAL (#MORE #HEIGHT $?X $?Y))))→:B7 satisfies the constraint.
(GOAL (#PUT $?X :BOX))→There is an ambiguity when interpreting the word "IT."
The referent of it corresponds to $?X. There are two potential referents: :B3 or :B7. In this example, SHRDLU determines the referents is :B7 and generates an utterance to confirm the determination of the referent.
The following sequence is the result of the execution of (#PUT :B7 :BOX).
(#PUT :B7 :BOX)
(#GRASP :B7)
(#GET-RID-OF :B3)
(#PUTON :B3 :TABLE)
(#PUT :B3 :TABLE)
(#MOVEHAND :TABLE)
(#UNGRASP)
(#MOVEHAND :B7)
(#GRASPING :B7)
(#MOVEHAND :BOX)
(#UNGRASP)
Exercise 1.2 Follow the sequence of the command by referring to the action rules.

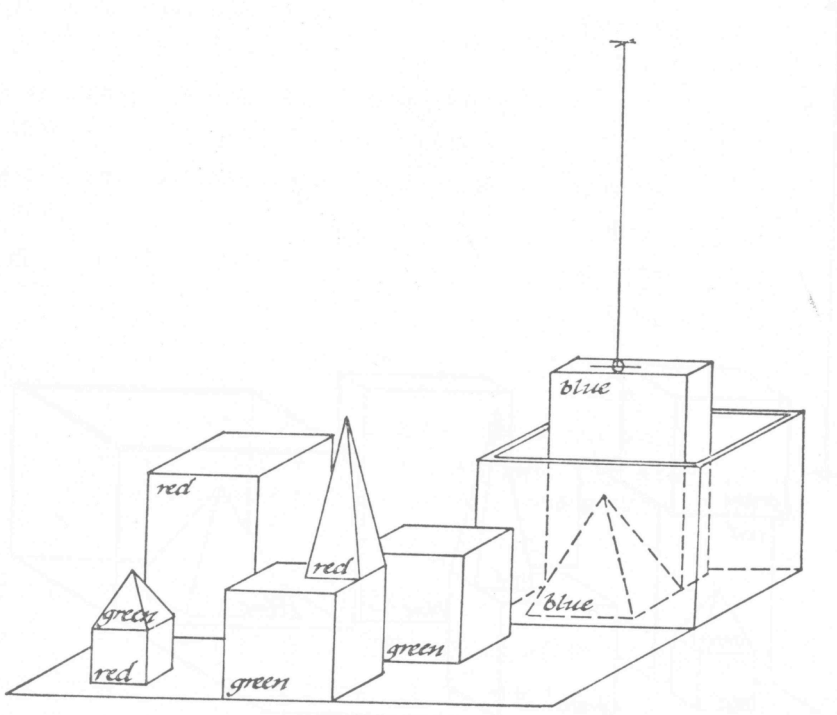


図 4 "Find a block which is bigger than the one you are holding and put it into the box.

This figure shows the result of Example 2.

Example 3.
F: What does the box contains?
S: THE BLUE PYRAMID AND THE BLUE BLOCK.

The question from the human is expressed as the following,
(FIND ALL $?X
    (GOAL (#CONTAIN :BOX $?X)))
SHRDLU obtains the answer X = (:B8 :B7). Then it generates the answer by referring to the features of :B8 and B7 in DB.

Example 4.
F: What is the pyramid supported by?
S: THE BOX

To answer the question, SHRDLU must refer to the context of the conversation. That is it must determine which pyramid the human mentions. The following is the expressions of the question.
(FIND ALL $?X

(GOAL (#IS $?Y #PYRAMID))

(GOAL (#SUPPORT $?Y $?X)))

In DB, there are three pyramids B2, B6, and B7. However, since SHRDLU manages the context of a conversation, it holds :B8 and :B7 in the list of the context. By using the context, it determines the referent of “the pyramid” is :B7.

Other examples of the interaction on SHRDLU

The following interactions are question and answer about the current situation.

F: How many blocks are not in the box?

S:FOUR OF THEM.

F: Is at least one of them narrower than the one which I told you to pick up?

S: YES, THE RED CUBE

Also, SHRDLU can answer the question about the physical constraint.

F: Can the table pick up blocks?

S: NO

F: Can a pyramid be supported by a block?

S: YES

2.6 Toward real world application

- There are vast (infinite) numbers of items or events in the real world.
- A part of them relates to the meaning of conversation. All items are not used.
- The items are already selected in the virtual world.
- A system must selects the items related to communication or interaction in real world.

3. Intelligent System in Real World

- The system must deal with
  - □ Dynamic changes in environments
  - □ Unpredictable events
  - □ Vast number of entities
    - ✧ Difficulty to make a model
- Vertical System and Horizontal System

3.1 Typical design of a system

- Horizontal decomposition with vertical slices
  - □ A traditional decomposition of a system is based on functional modules
    - ✧ Functional modules: Sensing, Mapping sensor data into a world representation, Planning, Task execution, and Motor control
    - ✧ Modeling and planning has an expression corresponding to the virtual expressions of SHRDLU.

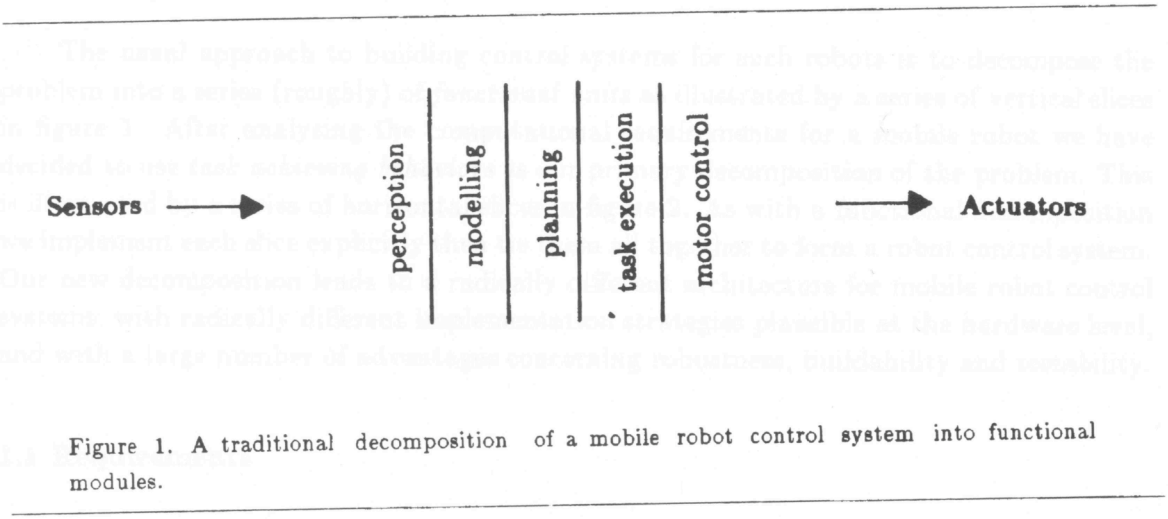


Figure 1. A traditional decomposition of a mobile robot control system into functional modules.

- A designer must consider the overall performance of a robot.
- It is difficult for a designer to modify the system.
- Disadvantage of the traditional decomposition is that the system cannot react the changes in an environment immediately.

3.2 Horizontal System

- A decomposition of a control system which is based on achieving behaviors
- Subsumption Architecture (SSA)
- Complex behaviors are the reflection of a complex environment.
  - ➢ On the other hand, the complex behaviors of traditional systems come from the complexity of the internal computations.
- Easy to react the changes in an environment immediately

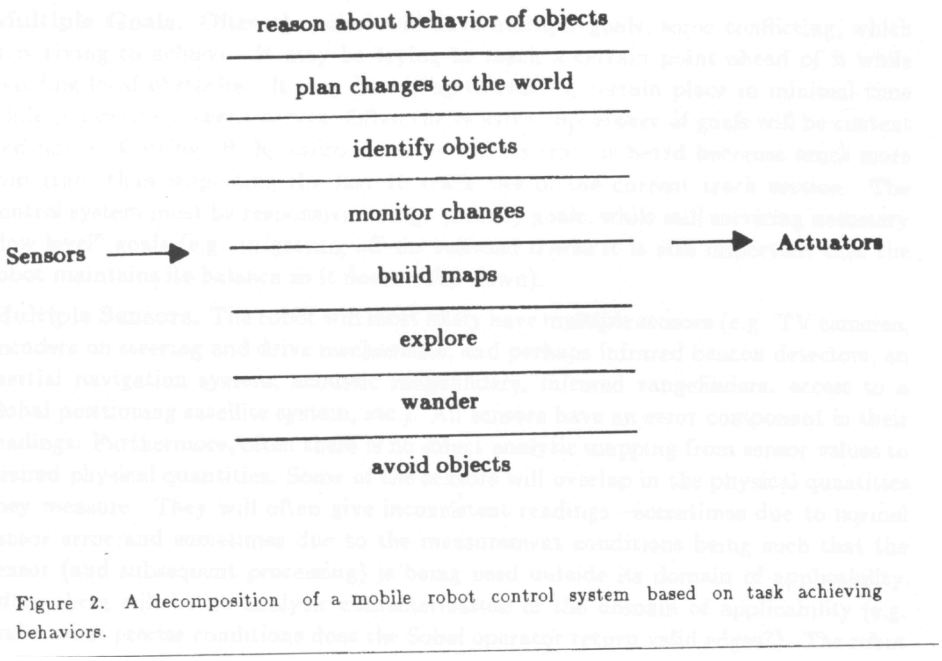


Figure 2. A decomposition of a mobile robot control system based on task achieving behaviors.

- How to design
  - Decomposing the problem vertically
  - Rather than the slice based on internal workings of the solution, the system slices the problem on the basis of desired external manifestations of the system.
  - A designer is able to design each module independently of others.

3.3 Requirements of a robot in Real World

- Getting sensor input and generating behaviors
- Multiple Goals
- Multiple Sensors
- Robustness
- Additivity

3.3.1 Multiple Goals

- There are several goals simultaneously
  - For example,
    - A robot tries to reach a certain place while avoiding local obstacles.
    - It tries to reach a certain place in minimal time while conserving power reserves.
  - The relative importance of goals will be context dependent.
- Behaviors of the horizontal system are selected in response to the structure of

environments.

- It need not have a model of the environments.

3.3.2 Multiple Sensors

- Many sensors
  - CCD cameras, encoders on steering and drive mechanism, infrared beacon detectors, an internal navigation system, laser rangefinders, sonic sensors, GPS, and etc.
- The robot must make a decision under the conditions.
  - No direct analytic mapping from sensor values to desired physical quantities
  - They overlap in the physical quantities they measure.
  - Inconsistent readings may occur.
    - Normal errors
    - Errors based on use outside its domain of applicability
    - No analytic characterization of the domain of applicability

3.3.3 Robustness

A robot must be robust to deal with

- Sensor fails: when some sensors fail, the robot should be able to adapt and cope by relying on those still functional.
- Dynamic environment: when the environment changes drastically, it should be able to still achieve some sensible behaviors.
- Hardware fails: when there are faults in parts of its processors...

3.3.4 Additivity

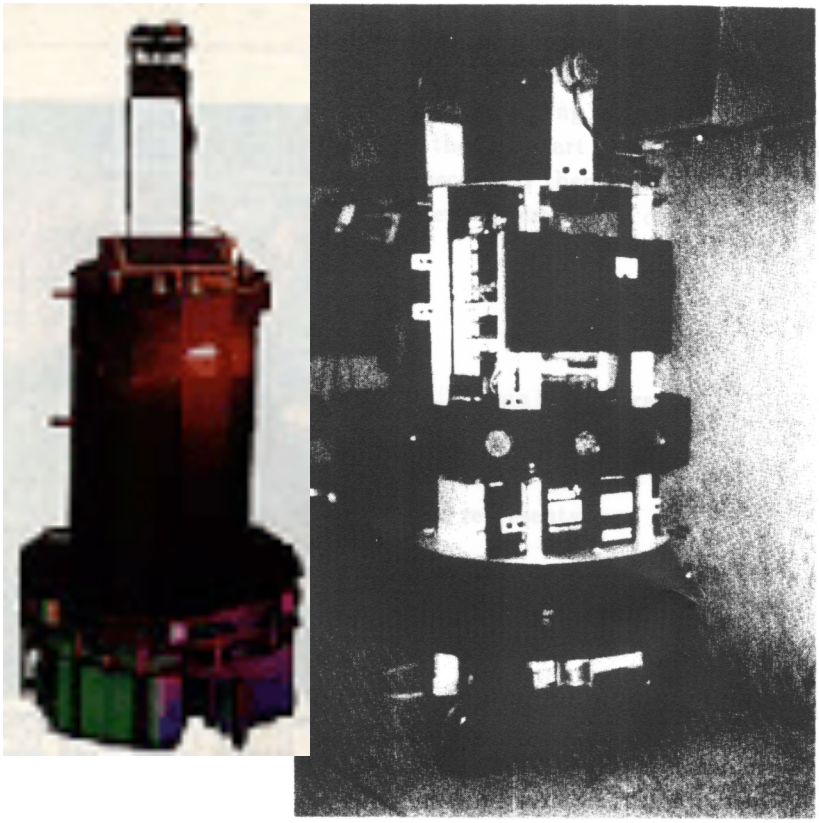
- This comes from a hardware requirement.
  - When more sensors and capabilities are added, it needs more processing power.

3.4 Subsumption Architecture

3.4.1 Robot

R.Brooks used a robot which has the following specifications.

- Mobile robot
- Sensors
  - A ring of twelve sonar
  - Pan tilt camera



3.4.2 SSA design

SSA is an actual example of the horizontal system.

- Decomposing the problem vertically
- Rather than the slice based on internal workings of the solution, SSA slices the problem on the basis of desired external manifestations of the robot control system.

3.4.3 Levels of competence

The following is an example of the vertical decomposition of a mobile robot.

- *Level 0:* Avoid contact with objects
- *Level1:* Wander aimlessly around without hitting things
- *Level 2:* Explore the world by seeing places in the distance which look reachable and heading for them.
- *Level 3:* Build a map of the environment and plan routes from one place to another.
- *Level 4:* Notice changes in the static environment.
- *Level 5:* Reason about the world in terms of identifiable objects and perform tasks related to certain objects.
- *Level 6:* Formulate and execute plans which involve changing the state of the world in some desirable way.

- *Level 7:* Reason about the behavior of objects in the world and modify plans accordingly.

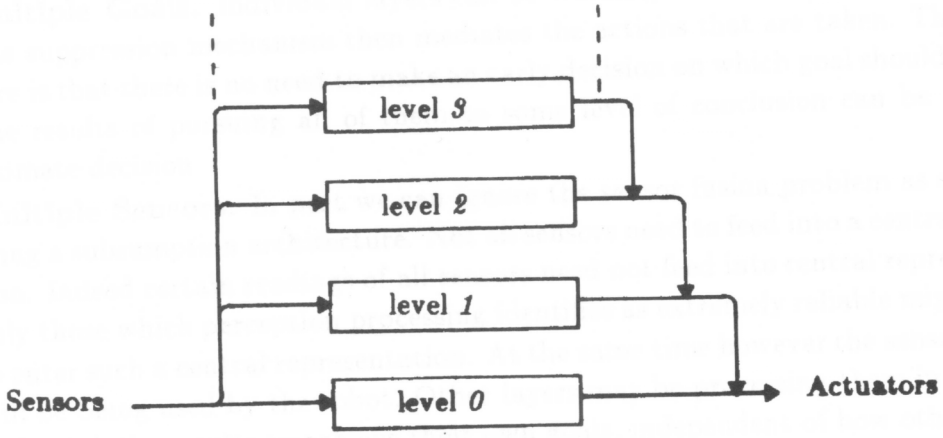


Figure 3. Control is layered with higher level layers subsuming the roles of lower level layers when they wish to take control. The system can be partioned at any level, and the layers below form a complete operational control system.

3.4.4 Tendency of generated behaviors

Behaviors generated by SSA change in response to the structure of an environment. Let us consider SSA which has modules: "avoiding behavior" for level 0, "random movement" for level 1. "going to W" for level 2, and "going to A" for level 3. The avoiding behavior prevents a robot from touching an obstacle. The random movement determines the direction the robot proceeds to randomly. The going to W makes the robot proceed to W when the sensor attached to the robot finds W. The going to A also makes the robot proceed to A when the robot finds A. However, the robot has another sensor to find A. Moreover, let us assume that the sensors have a different range of finding each own target. The range of the sensor for W is much wider than that of the sensor for A.

The SSA now we defined generates different behavior depending on the environment. If A and W are in the same distance from the robot, it proceeds to W even if going to A has higher priority than going to W. The reason why going to W is selected is because the location of A is out of range of the sensor for A. However, if W disappears while the robot is proceeding to W, it moves randomly according to the activation of random movement. If A comes into the range of the sensor for A while it moving randomly, it starts to proceed to A according to the activation of going to A.

Next, let us image an odd environment. W is in the verge of a distance from A. The distance is the same as the range of the senor for A. In the environment, the following

behavior is generated. If the robot is in a location where it can find W but cannot A, it starts to move to W. However, when it approaches the verge of the detectable distance of A, it turns its direction to A to go to A.

These differences of the behaviors come from the difference in the environments. In other word, the generated behaviors reflect the complexity and structure of the environments.

3.4.5 Design process of Layers of control

Designing SSA is achieved based on the following steps.

- Building layers of a control system corresponding to each level of competence
- Adding the developed competence on the lower layer.
  - Easy to add a new layer to an existing set to achieve a new behavior
- Starting the development from level0
- Add a higher layer on the lower layer.

Development of *level 0*

- Starting by building a complete robot system which achieving *level 0* competence.
  - Never alter *level 0* after it is debugged thoroughly because the *level 0* competence guarantees a most basic action for the robot.

Development of *level 1*

- Next building *level 1*
  - Examining data from *level 0*
  - Permitted to inject data into the internal interface of *level 0* suppressing the normal data flow in *level 0*.
- The achievement of *level 1* competence with the aid of *level 0*
- *Level 0* continues to run unaware of the layers above level 0 which sometimes interfere with its data paths.

Development of higher levels

- The work you must do in designing the rest of the layers is to repeating the same process for the higher levels.
  - Developing a competence and adding it on the lower layer.

The feature of the incremental development

- In the scheme of the development, the robot is able to run as soon as you have built the first layer.
- Additional layers can be added later
- The developed lower layers need never be changed.

3.4.6 Components for SSA

Each competence of SSA consists of some software modules. Because of the structure of the competences, designing them is to make the following.

- Internal structure of modules
- Communication structure between modules

The internal structure of modules consists of the following software structures.

- Each module is a finite state machine.
  - The machine is written with Lisp data structure
  - The machine has four state types
    - Output: An output message is sent to line.
    - Side effect: One of the module's variables is set to a new value.
    - Conditional dispatch: A predicate on the module's variables and input buffers is computed. Depending on the outcome, one of two subsequent states is entered. In short, this achieves conditional branch.
    - Event dispatch: A pair of conditions and states to branch to are monitored until one of the event is true. The events are in combinations of arrivals of messages on input lines and expiration of time delays.

The following lisp code is the example of avoid module.

```
(defmodule avoid
    :inputs (force heading)        ← force is calculated with force = distance^5.
    :outputs (command)               The equation expresses the evaluation of an
    :instance-vars (resultforce)     obstacle. The value increases exponentially the
    :states                          distance between it and the robot decreases.
    ((nil (event-dispatch (and force heading) plan))                ← Event dispatch
     (plan (setf resultforce (select-direction force heading)) go)  ← Side effect
     (go (conditional-dispatch (significant-force-p resultforce 1.0)←Conditional dispatch
                                                                start
                                                                nil))
      (start (output command (follow-force resultforce))           ← Output
             nil)))
```

There are two types of communication structure between modules

- Output may be inhibited.
- Input may be suppressed.

The connection is also expressed with lisp code. The followings are the example of the connection.

(defwire (feelforce force) (runaway force) (avoid force))

The output from the feel force module is connected to the inputs of the runaway module and the avoid module.

(defwire (avoid command) ((suppress (motor command) 1.5)))

The output from the avoiding module suppresses the input of the motor module.

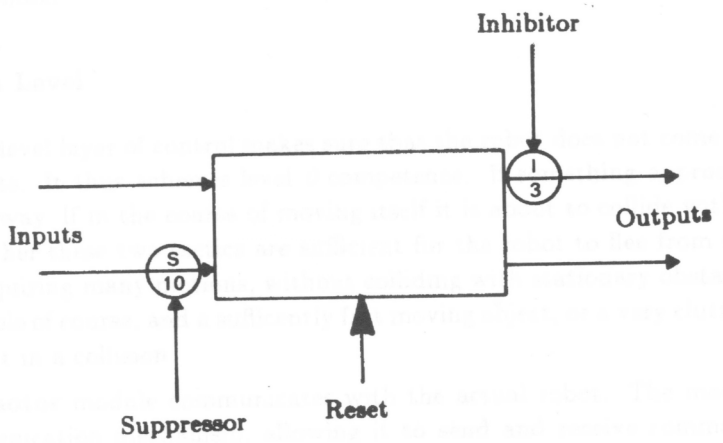


Figure 4. A module has input and output lines. Input signals can be suppressed and replaced with the suppressing signal. Output signals can be inhibited. A module can also be reset to state NIL.

3.4.7 Robot control system instance

Let us consider three layered SSA for an autonomous mobile robot as an example. The following is the example of each layer.

- *level 0*: prevent a robot from contacting with an object.
- *level 1*: wander avoiding obstacles.
- *level 2*: generate a path to reach a certain place.

Level 0

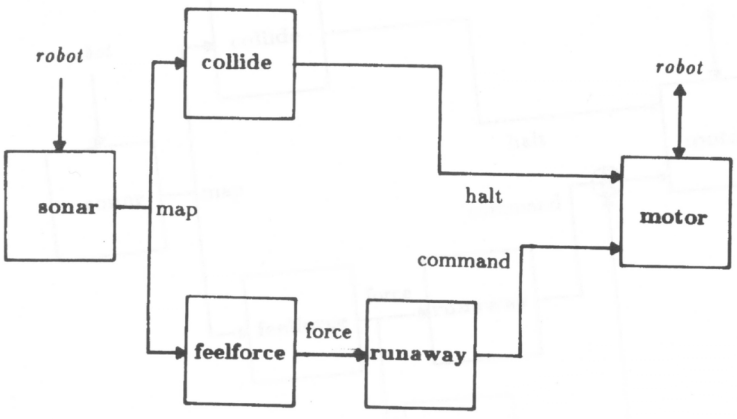


Figure 5. The level *0* control system.

This figure is the instance of *level 0*. It consists of sonar, collide, feelforce, runaway, motor. *Level 0* achieves basic behaviors to prevent the robot from contacting with an obstacle. The actual behavior is stopping the robot or turning its direction opposite the obstacle.

Sonar corresponds to ultrasonic distance sensors attached to the robot. Collide and feelforce are given data from sonar. Collide give motor halt command directly if it detects an object. The direct connection achieves immediate response. Feelforce evaluates the distance between the robot and the obstacle and gives the result named force to runaway. Runaway determines the direction of the robot in response to the force and gives the robot a command.

Level 0 has two paths toward motor. Since each module is calculated in parallel with the others, halt and command reach motor independently. When the robot met an obstacle, halt starts to have effect on motor at first because the path via collide is faster than the one via feelforce and runaway. After stopping the robot, the robot turns its direction according to the result of runaway. This parallel calculation guarantees a robust avoiding behavior. Even if there is no time to turn the robot opposite the obstacle, collide module can stop the robot as soon as possible.

Level 1

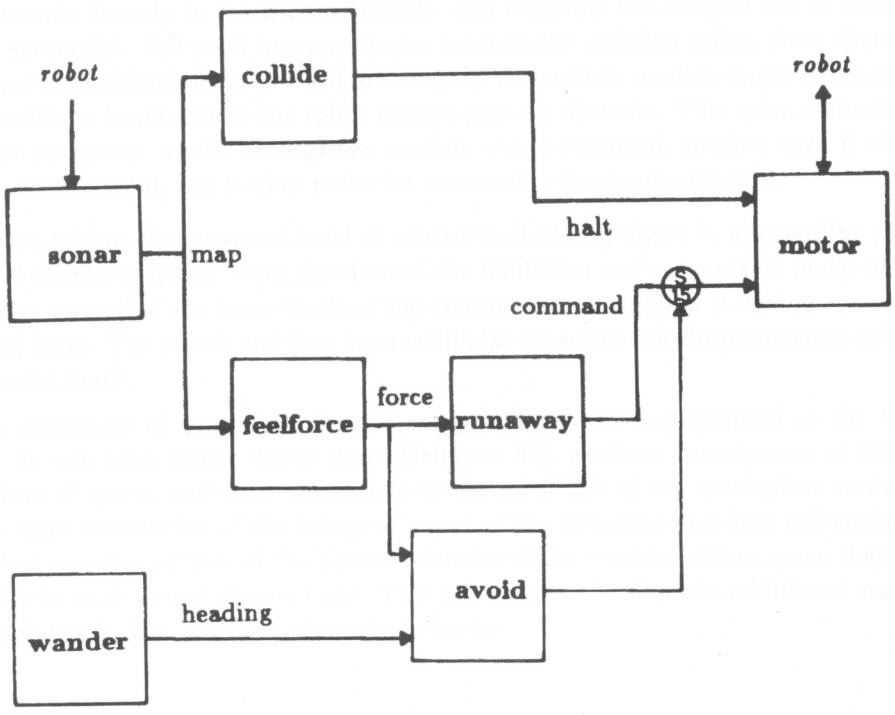


Figure 6. The level *0* control system augmented with the level *1* system.

This figure shows *level 1* and *level 0*. *Level 1* consists of wander and avoid. Sonar, collide, feelforce, runaway, and motor are those which I have already explained at *level 0*. *Level 1* makes the robot move toward random directions. Wander generates the random directions named heading and gives it to avoid. Avoid generates a motor command to make the robot proceed toward the given direction avoiding obstacles. Command from avoid suppresses command from runaway to override motor. However, there is a possibility that it takes longer time for avoid to calculate a new direction than *level 0*. Parallel execution of *level 0* and *level 1* also guarantees the timing of command. Even if command from avoid is delayed, *level 0* can control motor.

Level 2

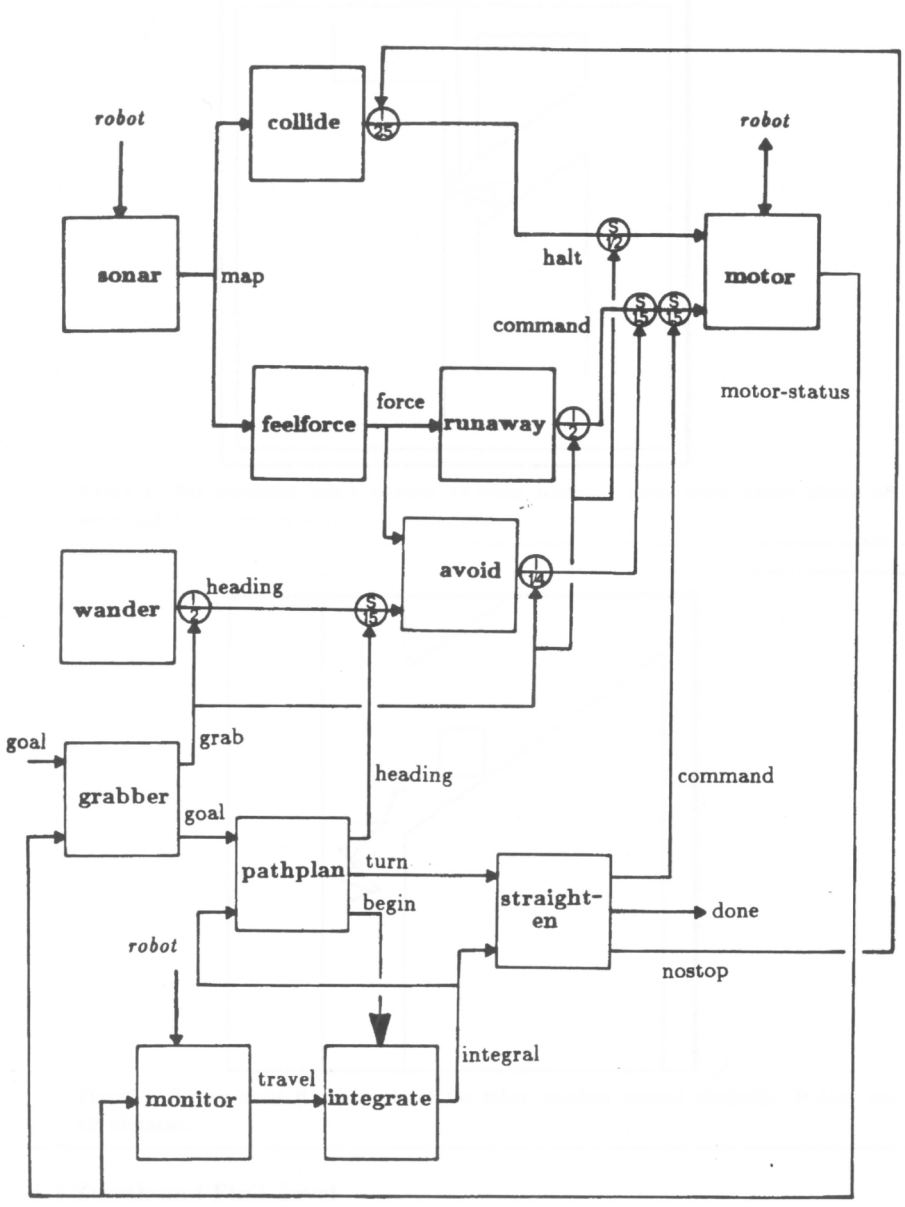


Figure 7. The level *0* and *1* control systems augmented with the level *2* system.

This figure shows *level 2, level 1, and level 0. Level 2* consists of grabber, pathplan, straighten, monitor, and integrate. These modules navigate the robot to a goal location. When a goal is given to *Level 2* by an upper layer or a user, it starts the navigation. At first, the goal is given to grabber. Grabber gives halt command to wander, avoid, runaway, and collide to override motor for a while. After giving halt, it confirms whether or not motor stops. After the confirmation, it gives pathplan the goal. Pathplan generates a path plan toward the goal location by dividing the path to the goal into the paths to sub-goals. Then, it gives avoid the path to each sub-goal one by one according to the current location of the robot. Giving the path is achieved along the heading line which suppresses the output from wander. Avoid make the robot proceeds to the direction avoiding obstacles.

Monitor detects the degree of the movements by referring to motor and other sensors attached to the robot. Then, it gives integrate the moving information along the travel line. Integrate is initiated by pathplan when it starts to plan a path, and maintains the current location of the robot by using information from monitor. The location is given to pathplan and straighten. When pathplan detects that the robot is near the goal, it stops the navigation and initiate straighten. Straighten make the robot move to the exact location of the goal. It inhibits halt from collide to proceed the goal because there is a possibility that collide prevents the robot from going to the goal if there is an object at the goal.

3.4.8 Simulation of SSA

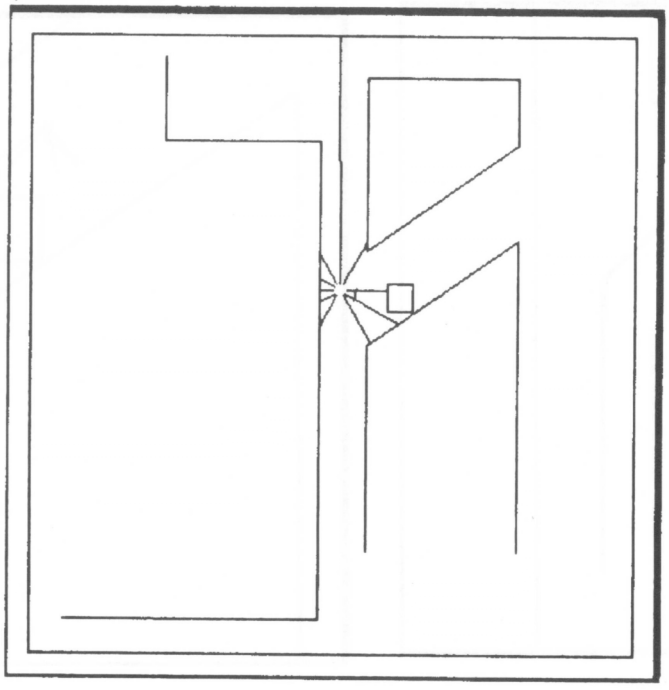


Figure 8. The simulated robot receives 12 sonar readings. Some sonar beams glance off walls and do not return within a certain time.

Let us confirm behaviors generated by the instance of SSA. This figure shows a scene where a robot measures distance from walls or an obstacle by a using sonar ring. The robot is in the crossing. The lines drawn from the robot denote the measurements of the distance.

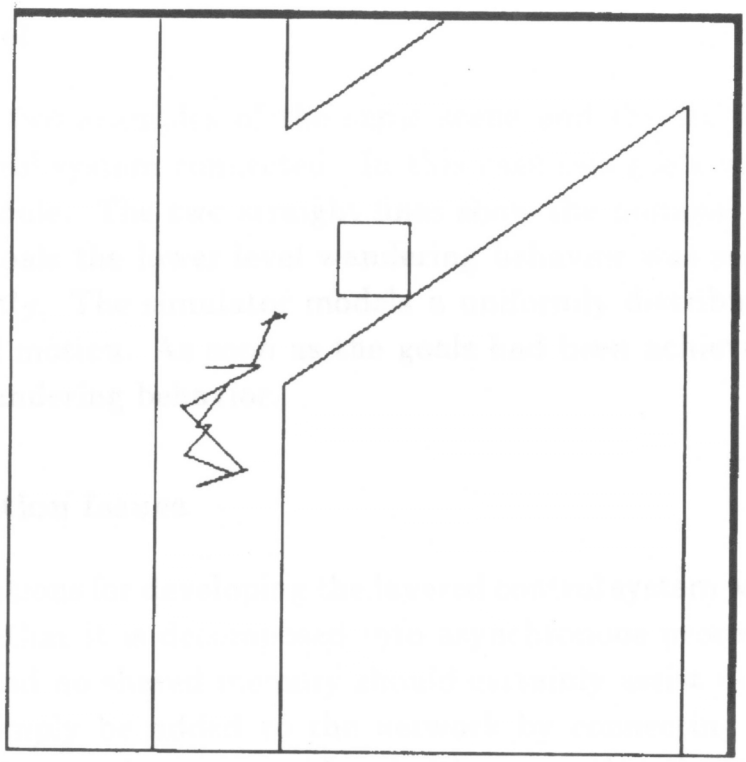


Figure 9. Under levels *0* and *1* control the robot wanders around aimlessly. It does not hit obstacles.

Next figure shows the example of behaviors when *levels 0* and *1* control the robot. If no goal is given, the robot behaves like this. Since *level 1* generates directions for the robot to proceed to randomly, it behaves aimlessly. However, it never hit the walls and the obstacle.

The figures in the next page are examples of behaviors generated by *levels 2*, *1*, and *0*. In the examples, the robot is in the lower part of the vertical corridor. Also, the goals are the upper right part of the corridor. Pathplan generates almost the same paths in the figures. The straight lines denote the paths. Also, there is an object on the way to the goal from the robot location when it proceeds along the paths.

In the figures, *level 2* make the robots proceed along the generated paths before finding the obstacles. However, avoid of *level 1* generates motor commands to avoid the obstacles. Moreover, since avoid gets information about both of the locations of objects and the location of a goal, the robots can proceeds to the goal even though it does not go along the panned path. After reaching the goals, *levels 1* and *0* starts to control the

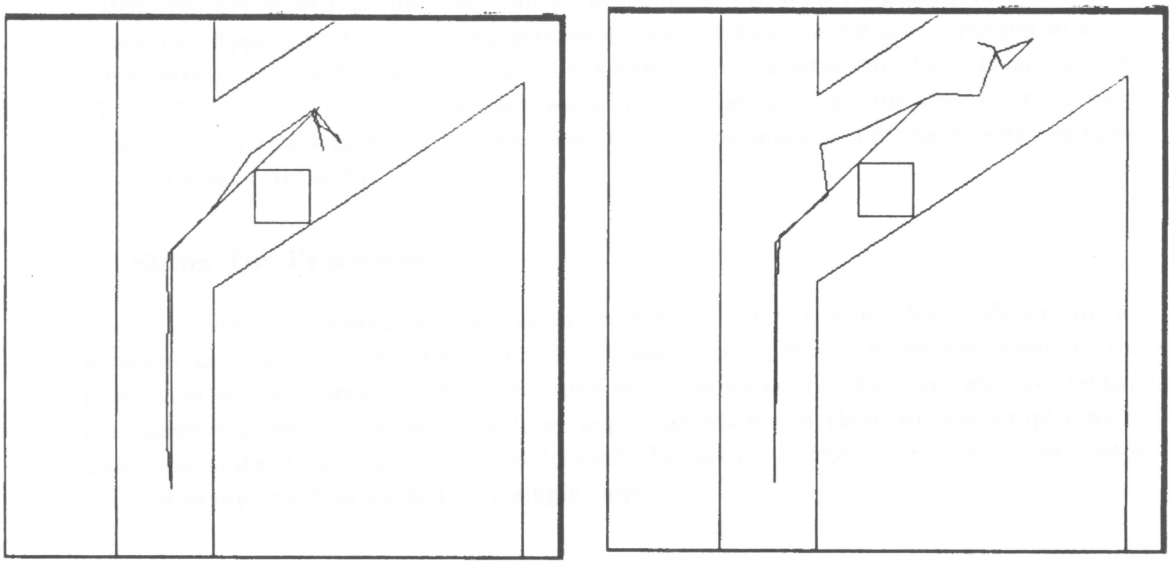


Figure 10. With level 2 control the robot tries to achieve commanded goals. The nominal goals are the two straight lines. After reaching the second goal, since there are no new goals forthcoming, the robot reverts to aimless level 1 behavior.

robot. The several lines around the goals are the result of the aimless behaviors.

In addition, the trails on the ways to the goals are different in the figures. The difference comes from the difference of timing of reading sonar data. The robot of the left figure finds the obstacle earlier than the one of the right. As a result, the robot of the left avoids the obstacle gradually. On the other hand, the right one avoids it suddenly.

3.4.9 Problems tackled by SSA

SSA can deal with the requirements for a robot in the real world. I already explain each requirement in 3.3. The following list explains how SSA achieves each requirement.

- Multiple Goals
  - Since individual layers works on individual goals concurrently, the robot designed with SSA can pursue multiple goals.
  - The suppression mechanism between layers mediates the action selection.
  - The advantage is that there is no need to make an early decision which goal should be pursued.
- Multiple Sensors
  - SSA can ignore the sensor fusion problem which comes from integrating sensor data into a central representation.
  - Not all sensors need to feed into a central representation.
  - Only those which perception processing identifies as extremely reliable might

be eligible to enter it because unreliable sensor data cannot activate a module.

- At the same time, other layers may be processing them to achieve their own goals independent of how other layers use them.

- Robustness
  - The use of multiple sensors guarantees that the system goes well regardless of some faults in some sensors.
  - Lower levels have been well debugged and continue to run independent of higher layers.
    - The lower level continues to produce outputs if the higher layers cannot produce results in a timely fashion.
- Additivity
  - Easy to spread the layers over many loosely coupled processors.
  - Each layer can be implemented on its own processor.
  - Communications between them require fairly low bandwidth.

3.4.10 Advantage of Layers design

The most prominent features of SSA are the responsive behaviors like following.

- Behavior selection reflects the structure of environments themselves.
  - No need to knowledge to select them.

R.Brooks calls the responsive feature intelligent without reasoning or intelligent without representations. These names implie the big difference between SSA and conventional intelligent system which has a central representation and carries out symbolic inference based on the representation.

The following list shows other advantages in using SSA.

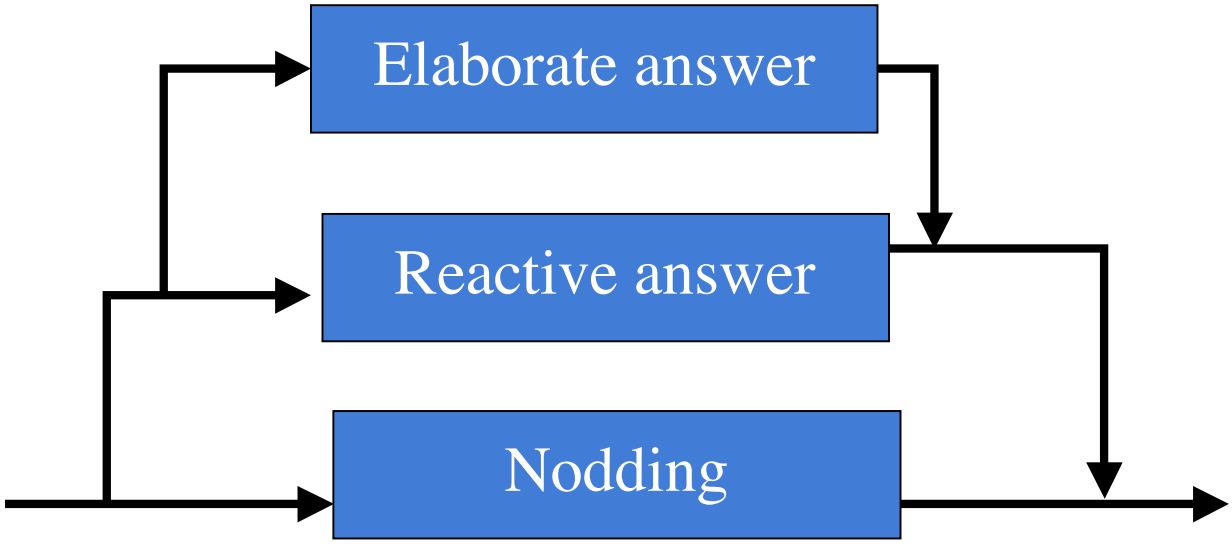
- The designers are free to use different decomposition for different sensor-set task-set pairs. This advantage make SSA be able to be used in designing many types of autonomous system; an intelligent robot, an intelligent room, an interactive system.
- The processors send messages over connecting wires.
  - There is no handshaking or acknowledgement of messages.
  - There is no other communication path. Ex. Shared memory.
  - The processors run completely asynchronously monitoring input wire and sending messages on output wires.
- There is no central control.
  - All processors are created equal.
- Higher level layers subsume the role of lower levels. The connection make it easy

for a designer to add a new module to SSA in existence.

- Inputs to modules can be suppressed and outputs can be inhibited by wires terminating form other modules.

3.4.11 Example of implemented systems

This picture is an example robot designed using SSA. The robot is an insect type robot which has six legs. Each leg is controlled by each individual module. In spite of the distributed control of the legs, it can move naturally in response to the figure of the surface of the land.



We can also design other intelligent system using SSA. The above figure is an example of a dialogue system. It consists of three layers. *Level 0* generates nodding

behaviors in response to a use's input. The response is quickest in the layers. *Level 1* generates a reactive utterance. For example, the system finds an utterance by searching database by keyword matching. The speed of the *level 1* response is slower than that of *level 0*. The response of *level 2* is much slower than the others. However, it can generate an appropriate answer by using inference and referring a communication context. The dialogue system achieves flexible communication. For example, it can response to a user's utterance by nodding, after nodding, it generates reactive answer. Finally, it gives a user an appropriate answer. The gradual response gives detail of information incrementally. On other hand, conventional dialogue system cannot response to user's utterance immediately because it generates all responses by using elaborate answer. The slow response will make communicative relationship between the user and the system collapse.

Moreover, there are more challenging attempts to develop communicative systems using SSA. For example, the following two systems try to establish cognitive architectures for human conversations.

- Stammering Computer (Incremental utterance generation)
  - 口ごもるコンピュータ
  - When a human talks with others casually, he/she does not always generate a complete sentence. He/she selects his/her words incrementally along progress of generating the sentence. The incremental generation can be achieved by SSA. The modules in lower layers select words in response to the other's utterance, and output the words as a part of the sentence. The modules in upper layers generate following sentence complementing or correcting the already generated words.
- Computer keeping its ears cocked
  - 聞き耳をたてるコンピュータ.
  - Even though a human does not participate in a conversation between others, he/she may sometimes keep his/her ears cocked. In the situation, he/she may hear just the sounds of their utterances, interpret the meaning of their each utterance, or infer their communicative intention from their utterances. The three cognitive processes differ in the degree of understanding the utterances. Also, the computational costs of these processes differ. We can accomplish a robot which keeps its ears coked by employing the three processes as each layer of SSA. The robot always hears human's utterances by *level 0*. If there is an interesting keyword, it generates a brief response to the human. Then, *level 1* generates a sophisticated utterance by referring to the meaning of

his/her utterance. Moreover, *level 2* starts to infer his/her communicative intention to make a plan to interact with him/her.

3.4.12 Disadvantage of SSA

In spite of the advantages of SSA, there are several disadvantages.

- Consistency between goal-oriented behavior selection and environment-oriented behavior selection
  - The behaviors SSA generates are based on the dynamics and the structure of environments. The generation sometimes causes a contradiction when it is given a goal.
- Difficulty of developing actual implementation
  - As you can see the implementation of *level 2*, the circuit is not easy to be understood at a glance. There may be some difficulties when you design the circuits of much higher layers.
- Self-recognition
  - Some of behaviors of a robot consist of the combination of actions which some of SSA modules generate separately. For example, the insect robot with six legs proceeds forward as a result of behaviors of each leg. However, SSA does not have a mechanism for self-recognition. Almost SSA systems are not conscious of what they do. The absence of the self-recognition makes it difficult for a system to achieve a given goal.

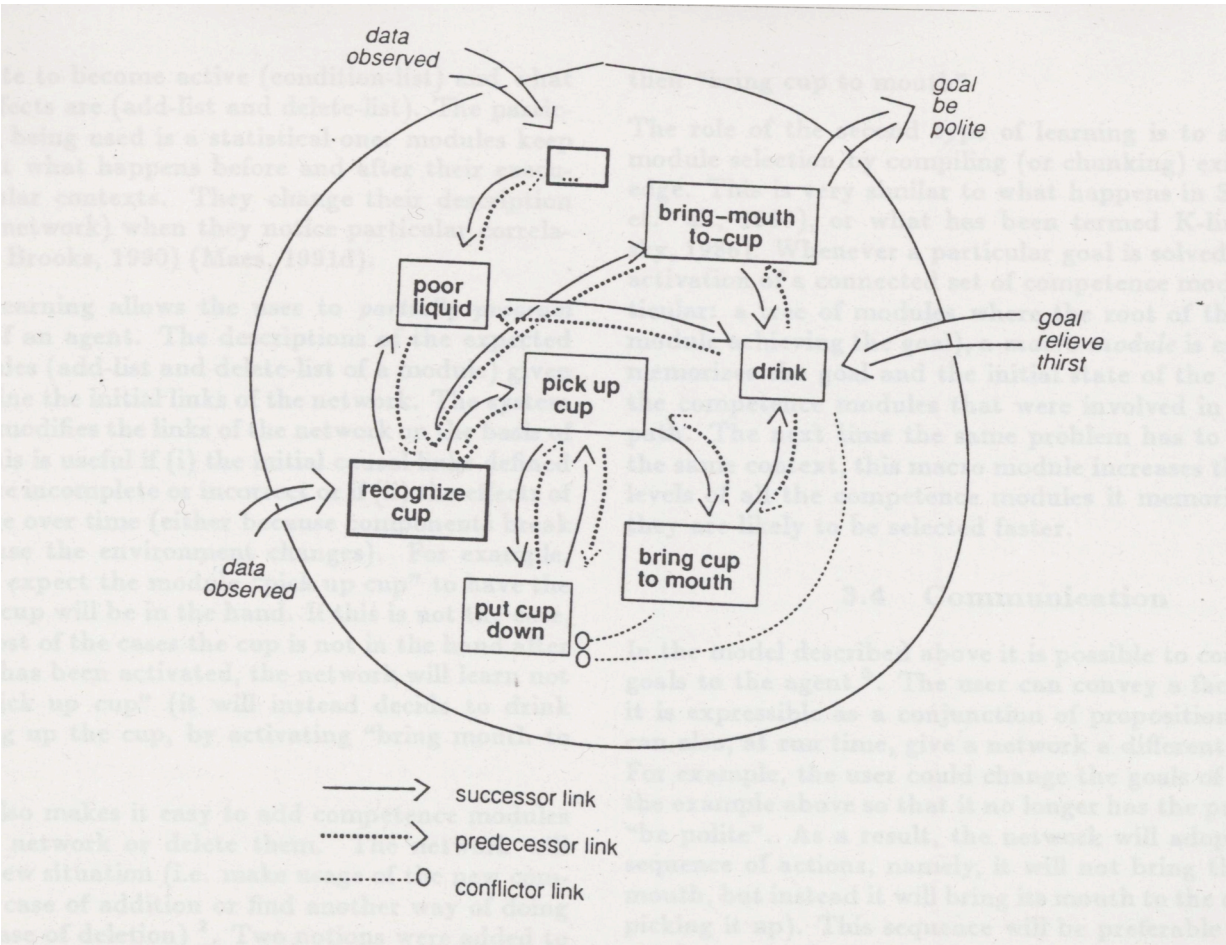
3.5 ANA (Agent Network Architecture)

- Agent Network Architecture (ANA) generates behaviors based on goal-oriented method and an environment oriented method.
- ANA has both features of planner (an example of goal oriented method) and SSA (an example of environment oriented method).
- The features vary depending on parameters.

3.5.1 Example of ANA

The figure in the next page is an example of ANA which can achieve drinking behaviors. There are seven agents expressed as rectangles in the figure. Also, one rectangle except for them of the seven agents is not given a name. It denotes the set of agents to find a content of the drink (ex. a bottle). Some of agents are connected to the environment (data observed in the figure). Also, some of them are connected to goals (goal relieve thirst and goal be polite in the figure). Moreover, each agent is connected

to others via three types of directed links: successor links, predecessor links, and conflict links. The successor link denotes anteroposterior relationship between agents.



An agent in the posterior part of the link can be carries out when the agent in the anterior part achieved its task. The predecessor link also is an anteroposterior relationship. This link activates an agent in the posterior part of the link when an agent in the anterior part must be carries out because the anterior agent requires the result of the execution of the posterior agent. The conflict link denotes a contradictory relationship. If two agents connected each other with the link, they generate a contradict effect on the world. The link inhibits the other agent to avoid the contradiction.

ANA employs energies distributed on the agents. The agent which has highest energy is carried out when an environment satisfies a condition which the agent requires for the execution. The links are used to distribute the energy. If an agent is given a value of energy, an agent which prepares a condition for the agent is given a value of energy along a predecessor link. Also, an agent which can be executed after the agent is given a value of energy along the successor link. Moreover, an agent which generates a contradictory result is inhibited along the conflict link.

In addition, some agents are given a value of energy from goals to achieve it. On the

other hand, there are agents who are given a value of energy from events in environments, which means that the environments satisfy the requirements of the agents.

Calculation process of the energy distribution and agents activation is done as follows.

1. Energy from environmental information
2. Energy from goals
3. Energy from other agents
4. Agents which has highest energy and executable conditions and whose energy is beyond a threshold θ are executed.
5. Update world model (add-list and delet-list)
6. Try 1-5 until given goals are satisfied.

3.5.2 Actual code of agents

Each agent is defined as four components: condition-list, add-list, delete-list, and activation-level. The next codes are examples of agents RECOGNIZE-CUP and PICK-UP-CUP.

```
def module RECOGNIZE-CUP
   condition-list: OBJECT-OBSERVED
   add-list: CUP-OBSERVED
   delete-list: OBJECT-OBSERVED
   activation-level: 53

def module PICK-UP-CUP
   condition-list: CUP-OBSERVED, HAND-EMPTY
   add-list: CUP-IN-HAND
   delete-list: HAND-EMPTY
   activation-level: 65
```

There are two links between the agents: a predecessor link from PICK-UP-CUP to RECOGNIZE-CUP and a successor link from RECOGNIZE-CUP to PICK-UP-CUP. The condition-list denotes a condition for executing the agent. The condition-list of PICK-UP-CUP expresses the requirement of CUP-OBSERVED and HAND-EMPTY. Since RECOGNIZE-CUP achieve CUP-OBSERVED which expressed in the add-list of RECOGNIZE-CUP, the two agents are connected with the predecessor link and the successor link. Also, RECOGNIZE-CUP is achieved when sensors detect an object and adds OBJECT-OBSERVED to a database. After confirming whether the object is a cup,

it adds CUP-OBSERBED to the database and deletes OBJECT-OBSERVED from it. The number 53 in the activation-level denotes a threshold. If a value of energy which RECOGNIZE-CUP has is bigger than 53, the agent has a possibility to be executed. Each link of PICK-UP-CUP also achieves the same function as RECOGNIZE-CUP.

The following example corresponds to a conflict link.

```
def module PUT-CUP-DOWN
  condition-list: CUP-IN-HAND
  add-list: HAND-EMPTY
  delete-list: CPU-IN-HAND, CUP-AT-MOUTH
  activation-level: 32

def module DRINK
  condition-list: CUP-AT-MOUTH
  add-list: RELIEVE-THIRST
  delete-list:
  activation-level: 87
```

Since the delete-list of PUT-CUP-DOWN has CUP-AT-MOUTH which the condition-list of PUT-CUP-DOWN also has, the execution of DRINK disrupts the condition of DRINK. To prevent the disruption, there is a conflict link between the agents. Also, the add-list of DRINK has RELIVE-THIRST. If the system is given a goal "relieve thirst," the energy of DRINK increases because energy comes from the goal.

3.5.3 Behaviors of ANA

The next figure shows the execution of ten agents to relieve thirst: RECOGNIZE-CUP, RECOGNIZE-BOTTLE, PICK-UP-CUP, PUT-CUP-DOWN, PICKU-UP-BOTTLE, POOR-LIQUID, PUT-BOTTLE-DOWN, BRING-CUP-TO-MOUTH, BRING-MOUTH-TO-CUP, and DRINK. Each graph shows a temporal change in the energy of each agent.

In the figure, DRINK has highest energy at first by giving a goal "relieve thirsty." However, the condition of DRINK which means that there is a cup of drink near mouth is not satisfied. This is why DRINK is not carried out at that time.

DRINK distributes energy to other agents along predecessor links. There are three agents BRING-CPU-TO-MOUTH, BRING-MOUTH-TO-CUP, and POOR-LIQUID in the posterior parts of the links. The energy of BRING-CPUT-TO-MOUTH and POOR-LIQUID become higher in response to the energy distribution. On the other

hand, the energy of BRING-MOUTH-TO-CUP does not increase as much as the others because it is inhibited by a goal “BE-POLITE.”

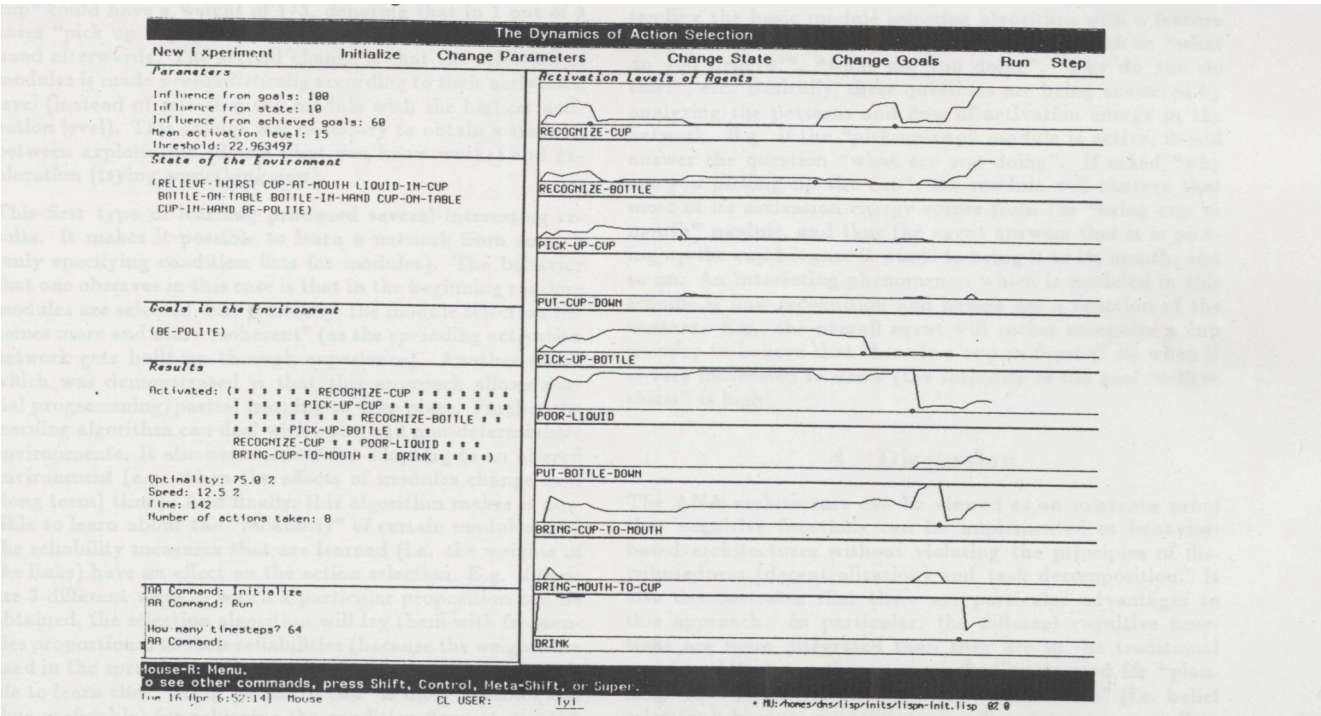


Figure 2: An example run of the module selection algorithm. The panes on the right side of the interface show the activation levels of the modules evolving over time. The small circles on the time axis denote when a module has been activated. The panes on the left show (from top to bottom): the global parameter values, the current state of the environment, the current goals of the agent, the list of modules that have been activated (chronologically, where “*” stands for a timestep in which no agent was strong enough) and a lisp interactor window.

Next, the energy of PICK-UP-CUP, RECOGNIZE-BOTTLE, and RECOGNIZE-CUP increase because these are predecessor of BRING-CUP-TO-MOUTH and POOR LIQUID. Moreover, RECOGNIZE-CUP has much higher energy than them because it gets energy from PICKU-UP-CUP and POOR-LIQUID. Then, it finds a cup. See a small dot on the time line of RECGONIZE-CUP. This figure employs dots when ANA carries out the agent. After recognizing the cup, the value of the energy of RECOGNIZE-Cup decreases.

The energy of PICK-UP-CUP increases because its condition is satisfied by RECOGNIZE-CUP. Then, ANA executed PICK-UP-CUP. Even though the energy of PICKU-UP-BOLLE is a high value, ANA does not carry out it because its condition is not satisfied. However, after RECOGNIZE-BOTTLE finds a bottle, ANA immediately carries out it.

Next, ANA confirms the location of the cup with RECOGNIZE-CUP and it carries out POOR-LIQUID because all conditions of it are satisfied. Then ANA carries out BRING-CUP-TO-MOUTH and DRINK. Since DRINK satisfies the goal “relieve thirst,” the energy of DRINK decreases. Also, energy of all agents decrease in response to the DRINK energy.

3.5.4 Effect of parameter setting

$$\theta \propto \frac{1}{speed} \qquad \theta \propto optimality$$

$$\beta \propto goal-oriented$$

There are two parameters in ANA: β and θ. β is a coefficient of a predecessor link. The range of the value is between 0 and 1. If the value of β increases, the agents connecting to the agent which has high energy get energy along the predecessor links. The energy distribution means that larger β becomes, more the behaviors of ANA depends on goals.

θ is a threshold of the activation of each agent. Larger θ becomes, more difficult the activation becomes. Since only most significant agents are carried out, the optimality of the system increases. However, the speed to achieve a goal decreases because the agents are seldom executed.

3.5.5 Effect of environments and goals

$$\#\,agents-realizing-goals \propto time-necessary-for-goal$$

$$\#\,goals \propto \frac{1}{goal-orientedness}$$

$$\#\,propositions-in-state \propto \frac{1}{data-orientedness}$$

The conditions of environments and goals also have effects on behaviors of ANA. If the number of goals which agent must realize increases, it takes long time to achieve them. If the number of goals increases, ANA does not pursue them because the effect of each goal becomes weaker. If the number of propositions which correspond to conditions of each agent increases, the effect of environments becomes weaker.

3.5.6 Applications of ANA

- Applications for dialogue system
  - ANA as dialogue planner
  - Easy to adapt changes in subjects

3.5.7 Disadvantages of ANA

- It is the same as the problem of planner.
  - Symbol grounding problem

4 Communication System in Real World and Focus of Attention

This section explains focus of attention which is used to interpret or generate utterances in the real world communications. I employ dialogue systems as examples for the explanation. The topic of this section is as follows.

- Dialogue system with a robot
  - Disambiguation with sensor information
- Logical frame work between verbal expressions and nonverbal information (situations)
- Focus of attention
- Recognizing situations with focus of attention for a dialogue

4.1 Dialogue system using sensor information

4.1.1 Linta-I

I introduce a dialogue system named Linta-I which employs sensor information when interprets utterances of a user.

4.1.2 Dialogue in the real world

We frequently use situated utterances in a conversation with someone. Since our utterances depend on a current situation or context, we must refer to the situation when generating or interpreting the utterances.

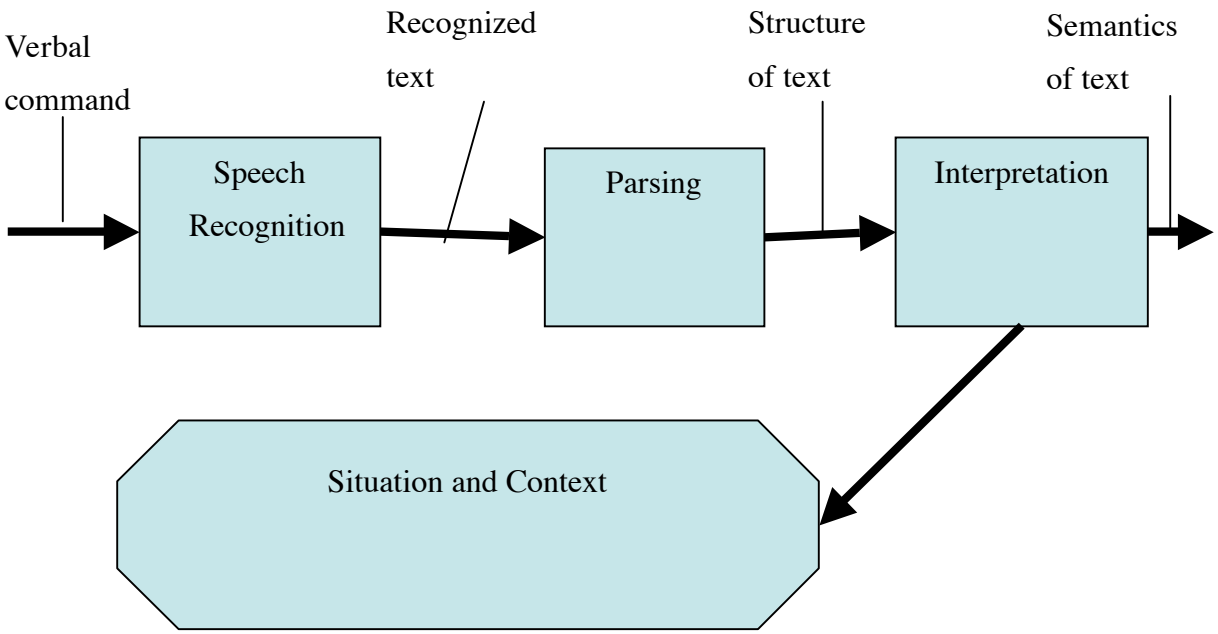
- They omit obvious information from their utterances. For example,
  - Close!
  - A robot must use real world information to interpret it.

4.1.3 Ambiguity in Dialogue System

Linta-I tries to interpret user's situated command with sensor information because the command has incomplete expression.

The figure in the next page shows a sequence when a dialogue system interprets a verbal command from a user. The command is recognized at a speech recognition module, and it outputs a recognized text string. The parsing module analyses the text string to obtain the grammatical structure. The interpretation module obtains the semantics of the text string by referring to situations and contexts.

The diagram in the figure contains three ambiguities. The ambiguities mean that the modules obtain some outcomes in their processes. They must select one of the outcomes



appropriate to an ongoing conversation.

- Ambiguity at Speech Recognition
- Ambiguity at Parsing
- Ambiguity at Interpretation

The ambiguity of speech recognition comes from matching voice to words. When the speech recognition module recognizes the voice command, there are several words in the candidate for the voice.

The ambiguity at parsing is that the parsing module can obtain a variety of parsing trees from the recognized text string.

The ambiguity at interpretation indicates that the parsing module can obtain a variety of semantic structures from the parse tree.

Each module must select appropriate one from each own outcomes. What information each module refers is crucial in selecting the outcomes. Linta-I focuses on ambiguity at interpretation. The context and situation are significant information for selecting the semantic structures. This subsection explains how to deal with the ambiguity by using sensor information.

4.1.4 Example of Interaction

The following example shows interaction between a user (U) and a robot (R). In the situation, he gives it a definite command “go forward.” And the robot proceeds forward. To achieve the command, Linta-I must select a command in response to his verbal expression to move the robot.

Ex.1

- U: Go forwad!
- R: Should I go forward?
- U: Yes.
- R: I see. I proceed.
- U: Stop!
- R: Yes. I will stop.

The next example shows that he gives the robot an incomplete command “go.” However, the robot cannot understand where it should go. Then, it asks him the direction to go. To achieve the interaction, Linta-I must find what information the command lacks, and generates a question to obtain the omitted information from him.

Ex.2

- U: Go!
- R: Which direction should I go?
- U: Right!
- R: I cannot proceed to right.

In Example 3, he also gives the robot an incomplete command “wrong” while it moves forward. Since the command does not suggest what is wrong, Linta-I cannot select the command of the robot in response to the verbal expression. So, the robot asks him what is wrong.

Ex.3

- U: Go forward!
- R: I see. I proceed.
- U: Wrong!
- R: What is wrong?
- U:Proceeding.
- R:Yes. I will stop.

In Example 4, he also gives the robot the same command as Ex.3. However, situation around it is different from Ex.3. In this situation, there is an obstacle at some distance away from the front of the robot. In short, the robot is moving toward the obstacle. At the situation, it can infer what is wrong if Linta-I has an inference mechanism. Since Linta-I infers that moving forward endangers the robot, it can conclude that what is wrong is moving forward. In Ex. 4, the robot generates an utterance “do you mention that proceeding is wrong?” to confirm the result of the inference.

Ex.4

- U: Go forward!
- R: I see. I proceed.

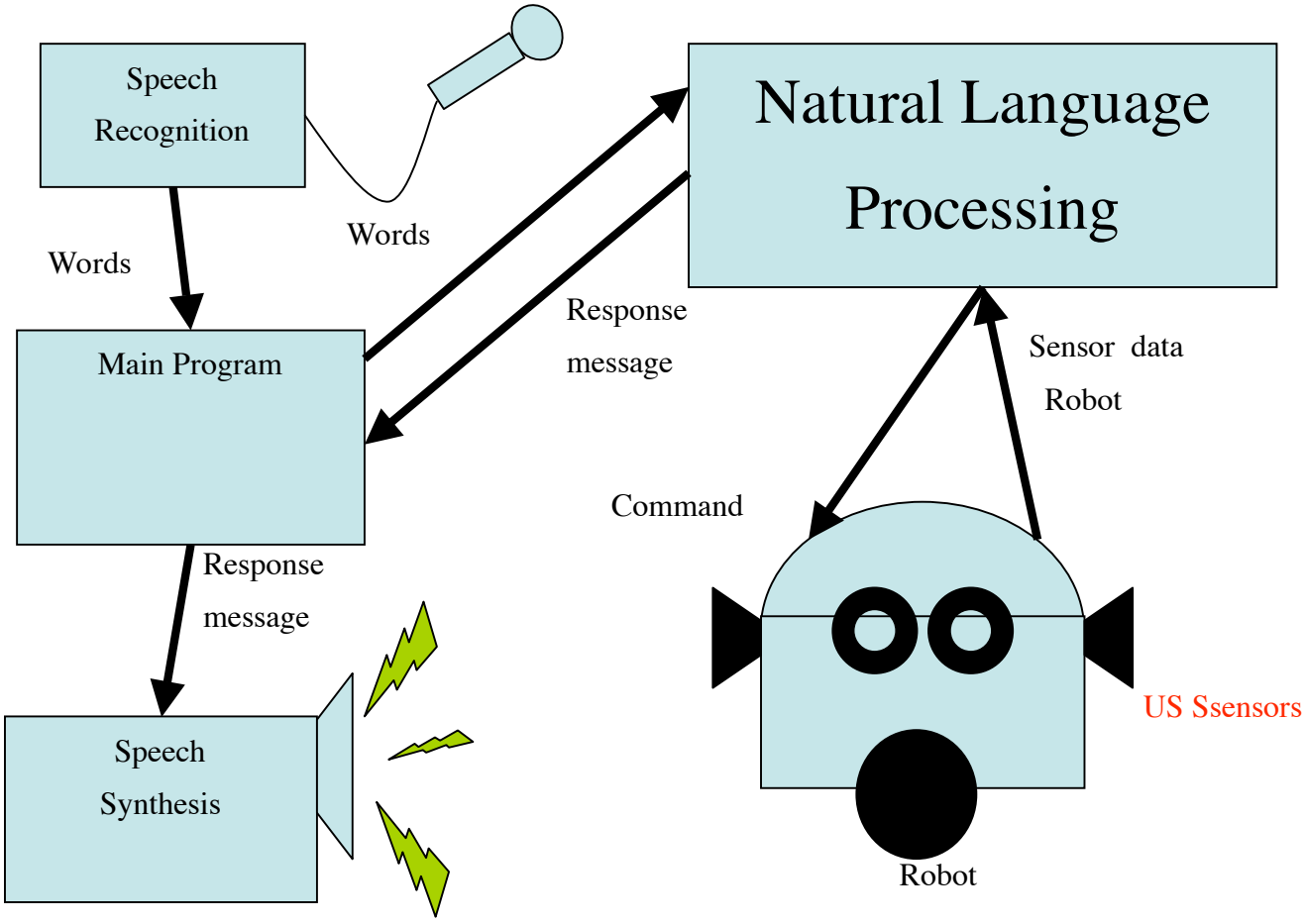
- U: Wrong!
- R: Do you mention that proceeding is wrong?
- U: Yes.
- R. I see. I will stop.

In Example 5, the distance between the robot and the obstacle is much closer than Ex.4. In the situation, the robot almost hits it. Since it is not sufficient time for Linta-I to confirm the inference, it decide that what is wrong is moving forward and stops the robot immediately.

Ex. 5

- U: Go forward!
- R: I see. I proceed.
- U: Wrong!
- R. I see. I will stop.

4.1.5 Structure of Linta-I



This figure shows the system structure of Linta-I. It consists of a speech recognition device, a speech synthesis device, a main program, a natural language processing module, and an autonomous mobile robot. The robot has four wheels (two drive wheels

and two caster). It also has four ultrasonic distance sensors. It can find obstacles in the direction of its front, left, right, and rear.

Voice sound is recognized by the speech recognition device. The main program sends the recognized words to the natural language processing module and obtains a response message from it. The speech synthesis device generates voice sound from the response message.

The natural language processing module analyses the recognized words by referring to sensor information and generates the response message and a command to the robot.

4.1.6 Natural Language Processing

The natural language processing module employs several techniques to analyses a command which is expressed with natural language.

- DCG (Definite Clause Grammars)

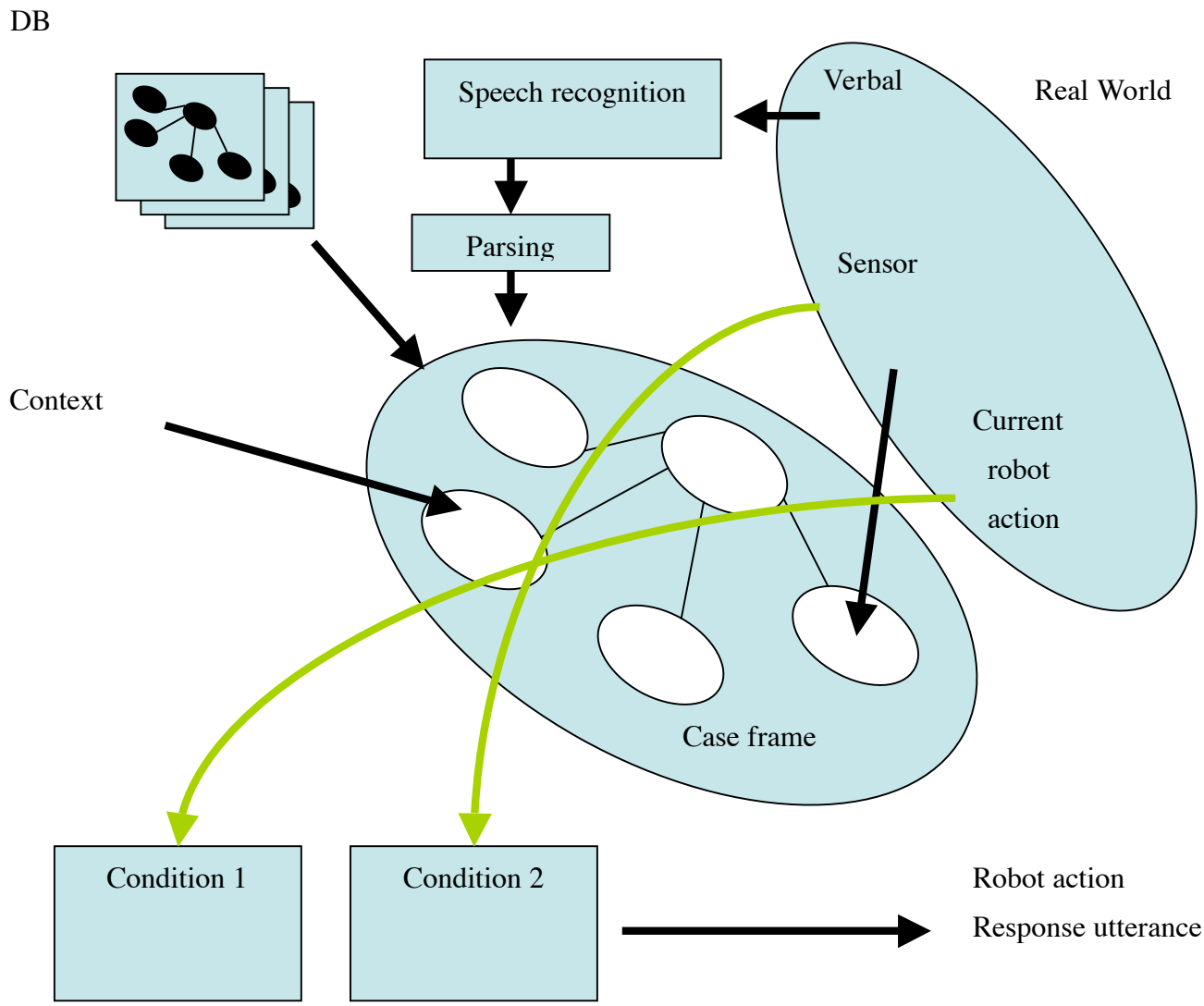
DCG is a way to define a grammar which is used to get a parse tree. The notation of DCG is based on context free grammars which is the same as the one used in the grammar module of SHLDRU.

- Case Frame
- Ex. Go forward!
  - [proceed, [agent [robot]], [goal, [forward]], [source, X], [adv, Y]]

Case frame is a semantic expression. The main part of case frame is a verb word. Also, it has several slots which form the semantic expression with the verb. The above example is the semantic expression of a command “go forward.” The verb is “proceed.” It has four slots: agent, goal, source, and adv. The agent slot shows a subject who achieve a verb. Since the example is a command, the agent slot has the value of robot. In other words, it means that the verb “proceed” is achieved by the robot. The goal slot expresses a location or a direction of “proceed.” The example indicates that the goal is the direction of the front of the robot. The source slot suggests the initial location of “proceed.” The adv slot suggests how degree the robot carries out “proceed.” Since the example “go forward” does not suggest source and adv, these slots do not include values. Slots in a case frame are prepared based on each verb.

The figure in the next page shows a conceptual sequence in interpreting a user’s command. Each slot of a case frame which is the network in the middle of the figure is filled by each word of a verbal command and information stored in the context.

The case frame is selected from DB in response to the verb of the command. Linta-I selects an action rule in response to the case frame. If there are slots whose values are not filled, the slots causes an ambiguity in selecting a robot’s action and response.



4.1.7 Knowledge of Linta-I

Linta-I has the following knowledge about its actions and objects.

- Robot action
  - Go forward
  - Turn right
  - Turn left
  - Go Backward
  - Stop
- Object

Also, it has a knowledge frame to express the above knowledge.

$$\begin{bmatrix} name\_of\_knowledge \\ attribute[value] \end{bmatrix}$$

The knowledge frame consists of the set of attributes and values. Linta-I has three

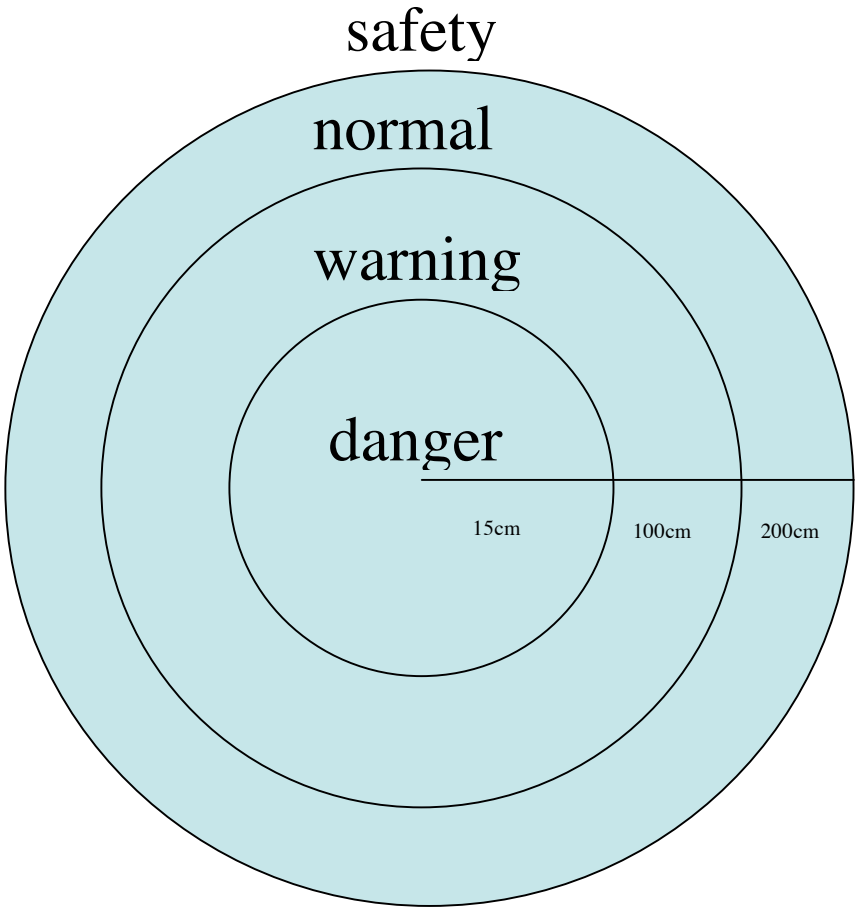
types of attributes in the knowledge frame.

- attribute
  - safety
  - distance
  - robot_action

In addition, the value of each attribute is determined according to the following table.

attribute	value
safety	danger warning normal safety
distance	0 - 250 cm
robot_action	running no_running

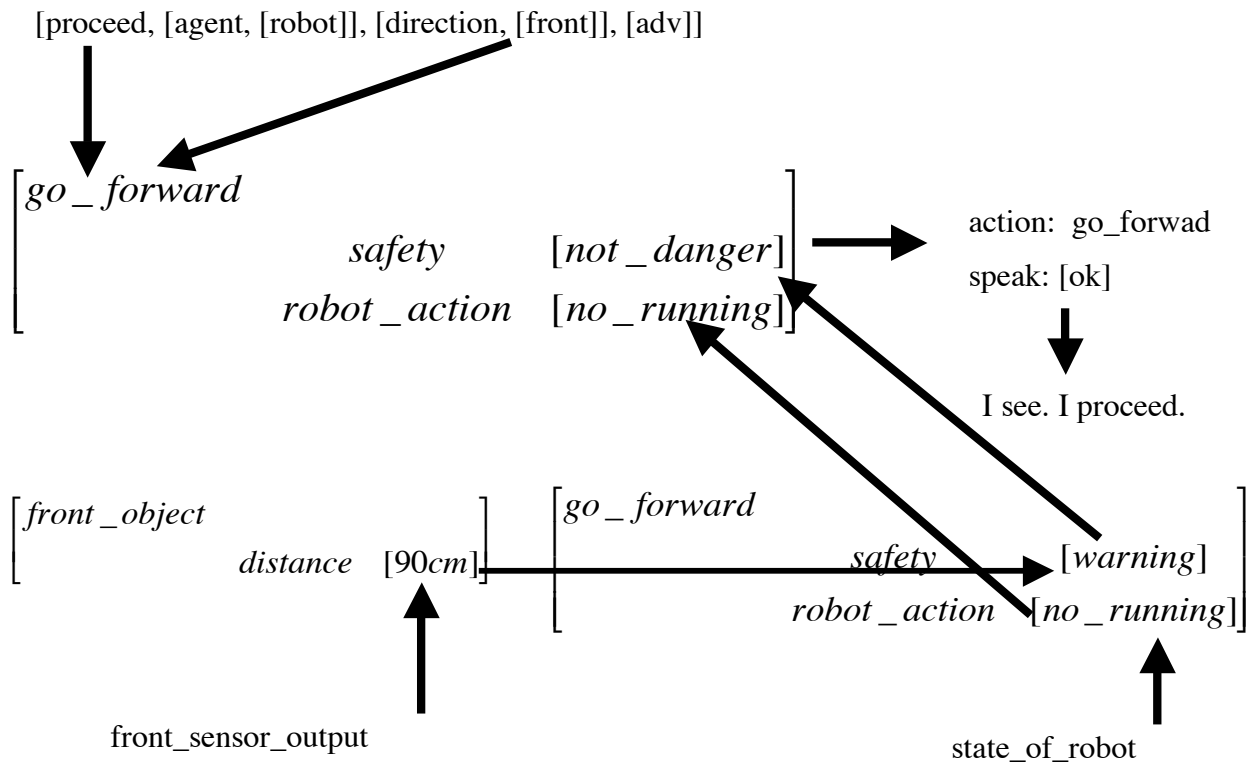
The value of the safety attribute is determined based on the constraints expressed in the following figure.



The center of the figure expresses the position of the robot. Linta-I determines the value according to the distance between the robot and an obstacle. For example, if an obstacle is in 90 cm from the robot, the value of the safety attribute becomes warning.

4.1.8 Example of interpretation

Let's consider an interpretation process. The case frame [proceed, ….] is generated from a user's command "go forward." Linta-I selects a knowledge frame "*go_forward*" according to proceed and [direction, [front]] in the case frame.



The attributes *safety* and *robot_action* in the knowledge frame are not filled at the beginning. Linta-I determines the values by referring to the knowledge frame of a sensor and a robot action. Since the value of sensor is 90 cm, the knowledge frame of *front_object* becomes the one in the figure. Then, the value of the safety in *go_forward* becomes *warning*. Also, the value of *robot_action* becomes *no_running* by referring to the ongoing robot action.

According to the filled knowledge frame, Linta-I selects an action go_forward and a response massage [ok].

4.1.9 Constraints on decision

Linta-I has constraints to determine whether it carries the robot action. The constraints are used when Linta-I interprets a user's command. For example, the user

command “wrong” is expressed as [reject, [object A]]. When Linta-I determines [object A], it refers [object A] by using the following constraint.

$$\begin{bmatrix} go\_forward & & \\ & safety & [danger\_or\_warning] \\ & robot\_action & [running] \end{bmatrix}$$

The constraint is based on the knowledge frame of an action “*go_forward.*” It indicates that the robot carries out an action and has a value of safety which is *danger* or *warning*. They are indicated by “*robot _action* [*running*]” and “*safety* [*danger _or _warning*].” If a current knowledge frame satisfies the constraint, Linta-I determines that [object A] is the action that the robot is executing now.

Moreover, the response of the robot varies depending on the value of the safety. If the value is *danger*, Linta-I immediately stops the robot. On the other hand, If the value is *warning*, it generates an utterance for confirmation.

The other example is a situation where a user gives the robot a command “go forward.” In response to the command, Linta-I has a case frame [proceed, [agent, [robot]], [direction, [front]], [adv]]. Although the case frame indicates a robot action “*go _forward*” directly, Linta-I does not execute it soon. It confirms a current robot action by referring to the knowledge frame of the current robot action and a constraint for the command. The following constraint indicates that a robot does not moving forward. If a current knowledge frame does not satisfy the constraint, Linta-I generates an utterance “I am already moving forward.”

$$\begin{bmatrix} go\_forward & & \\ & safety & [S] \\ & robot\_action & [no\_running] \end{bmatrix}$$

4.1.10 Response utterance

Linta-I has a set of utterances for the response to user’s commands. The following list is the examples of the response utterances. Linta-I generates a case frame to select a response utterance. Also, since the case frame has a variable, Linta-I generates an appropriate utterance by set a value on the variable.

- [where]
  - Where should I go?
- [yes_no, [proceed, X]]
  - Do I go toward X?
- [yes_no, [subject, X]]

- Do you mention X?

・ [what_object, X]

- What is X?

4.2 Logical framework between verbal expressions and nonverbal information

4.2.1 Situation and Attitude

・ Theory of meaning of situated expression

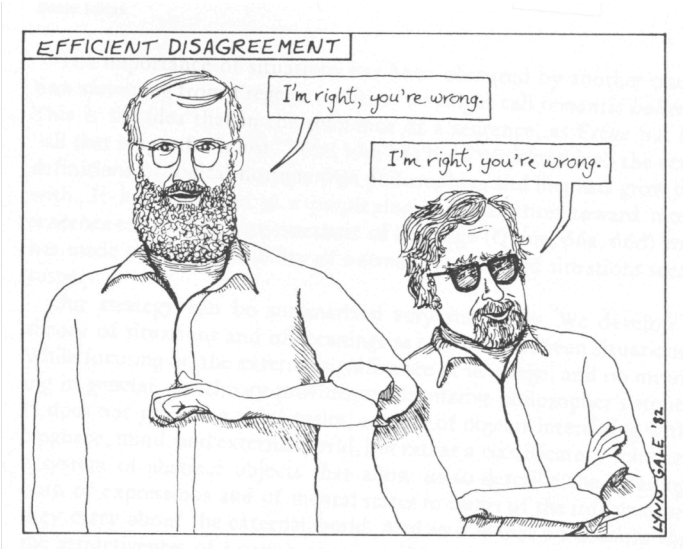
- Jon Barwise and John Perry, 1983 proposed the theory.
- The meaning changes according to a current situation.

・ Relation between verbal expression and situations.

4.2.2 Situated Expression

The meaning of situated language expressions depends on a situation. The following figure shows the example of situated utterances.

・ I'm right, you're wrong.

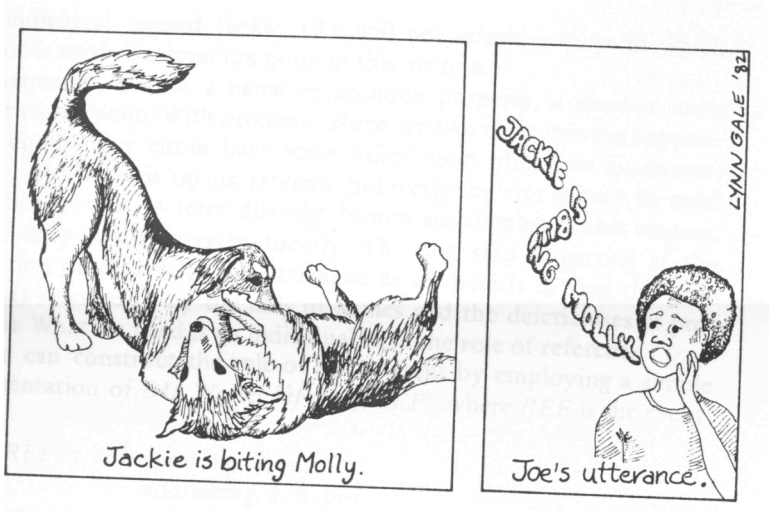


In the figure, two persons say the same utterance. However, the meanings of the utterances are different depending on the situation. Let's consider the difference by yourself.

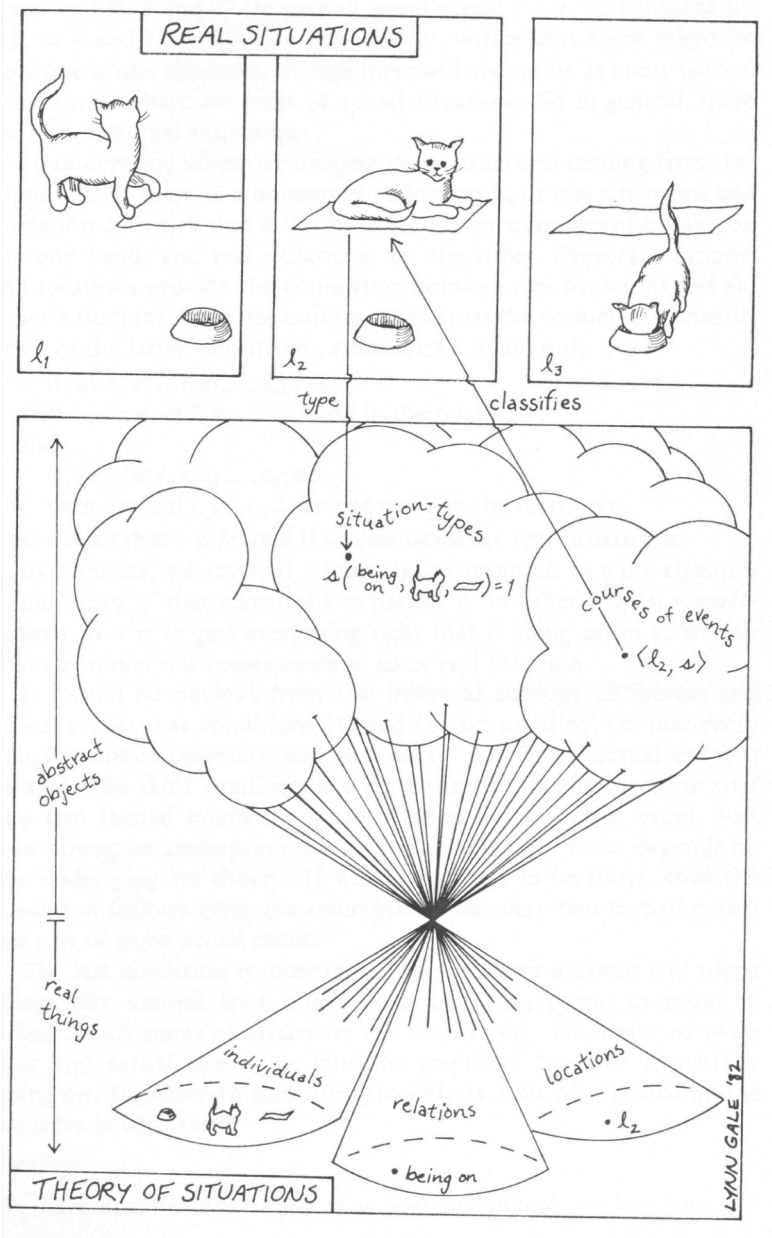
The next figure also shows the example of a situated utterance. The utterance is as follows.

・ Jackie is biting Molly.

In the example, a speaker Joe says an utterance which describes a situation at a distance from Joe. The meaning of the utterance also depends on the situation.



4.2.3 Relation between Physical event and Logical from



The figure shows the relation between real situations and logical expressions. The

concept is important to understand the meaning of the situated utterance. In the real world, there are many entities and events. However, humans cannot notice all of them. They focus on a part of them in their everyday life.

The figure indicates that humans have constraints (abstract objects in the figure) which extracts relevant real things from the real world: individuals, relations, locations, and so on. Humans always have the constraints in their activities. Moreover, utterances of humans introduce the constraints to others. In short, the meaning structure of an utterance is a relationship between the real things recognized through the introduced constraints.

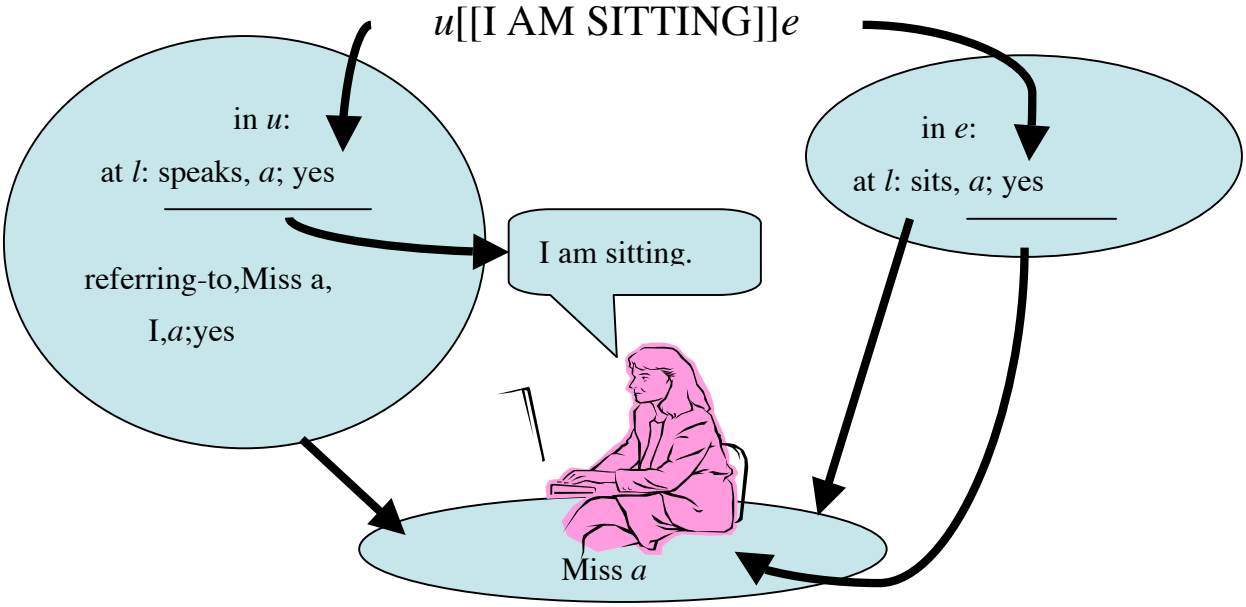
4.2.4 Logical From

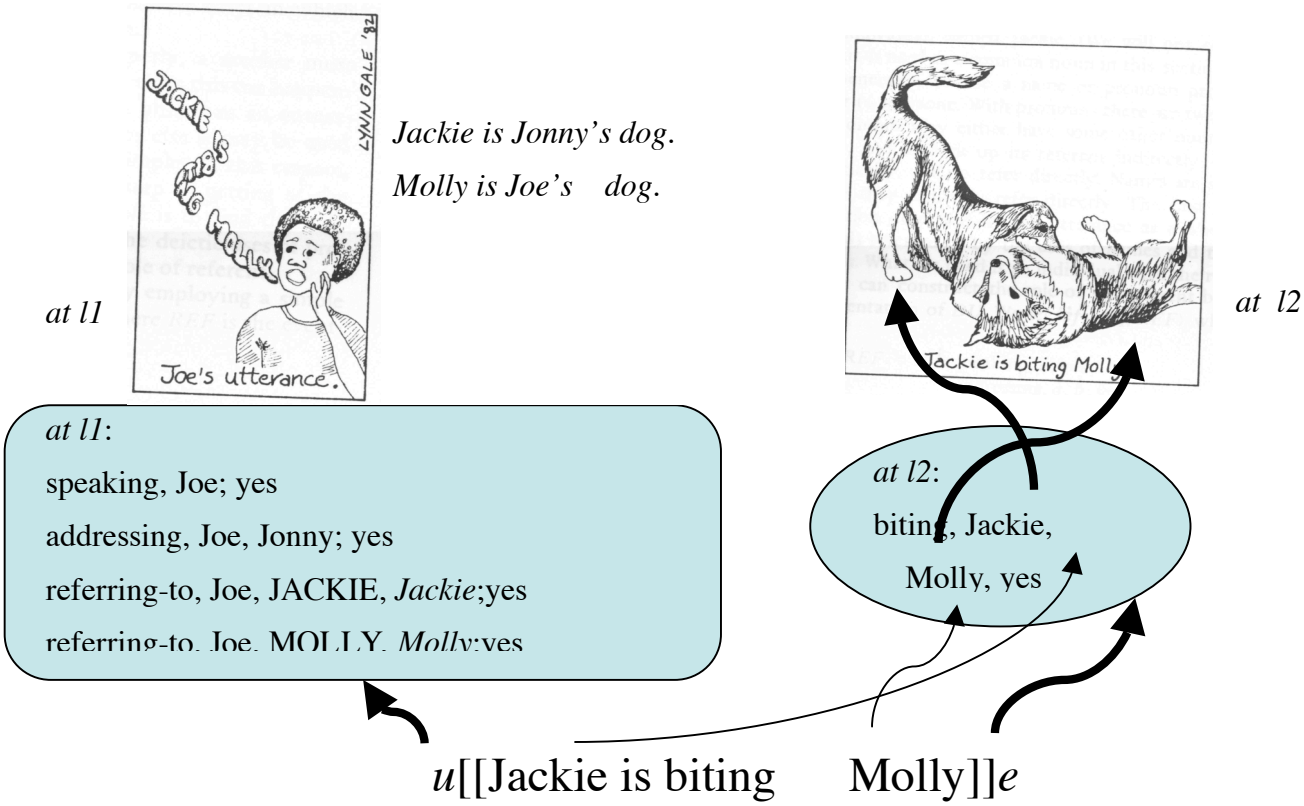
The theory of situation deals with the meaning of an utterance using the following logical form.

- *u*[[I AM SITTING]]*e*
  - *u* denotes a discourse situation; who is speaking, when and where, what words are being uttered, and to whom.
  - *e* denotes a recognized/described situation; what events occur and where.

  - in *u*: at *l*: speaks, *a*; yes
  - in *e*: at *l*: sits, *a*; yes

The following figure shows an example situation when a person says "I am sitting." The discourse situation indicates that Miss a refers herself by using a word I. Since I = a in the discourse situation, we can obtain logical expression "at l: sit, a; yes" from the utterance "I am sitting." The logical expression forms a described situation.





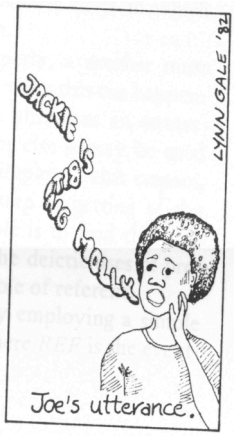
This figure shows the interpretation of "Jakie is biting Molly." In the figure, the discourse situation is different from the described situation. The discourse situation is at *l1* where Joe tells Jonny about the situation at *l2*. Also, the discourse situation assumes that both Joe and Jonny knows that Jackie is Jonny's dog and Molly is Joe's dog. Because of the mutual knowledge, Joe and Jonny can connect the words "JACKIE" and "MOLLY" with actual dogs "Jackie" and "Molly" which exist in the real world. As the example suggests, the discourse situation introduces a basis for bindings between a word and a real thing.

According to the bindings, Jonny can obtain a described situation where Jackie is bigint Molly at *l2*. *L2* does not suggest a exact location but indicates that the location is different from *l1*.

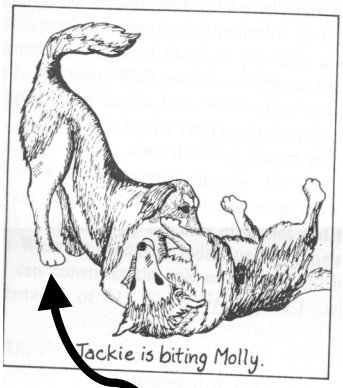
Next figure shows the other discourse situation. In this situation, although Joe knows Jackie is Jonny's dog and that Molly is someone's dog, Jonny is unfamiliar to Molly.

The discourse situation indicates that Jonny cannot connect the word "MOLLY" with an actual dog "Molly." Instead of "Molly", the discourse situation uses a variable *b* to express the unfamiliar dog Molly. Moreover, there is possibility that Jonny imagines that Molly is a human.

Since the discourse situation is different from the former one, Jonny obtains a described situation where Jackie is biting *b*. If Jonny imagines that *b* is a human, the



Jackie is Jonny's dog. Jonny is unfamiliar to Molly.



at l1

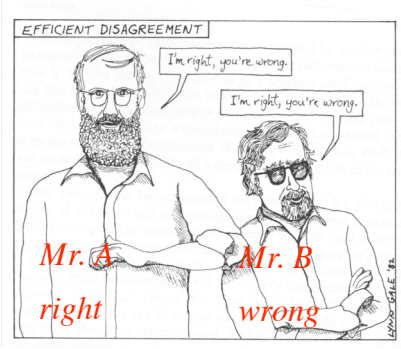
at l2

at l1:
speaking, Joe; yes
addressing, Joe, Jonny; yes
referring-to, Joe, JACKIE, *Jackie*;yes
referring-to, Joe, MOLLY, *b*;yes

at l2:
biting, Jackie,
b, yes

u[[Jackie is biting Molly]]*e*

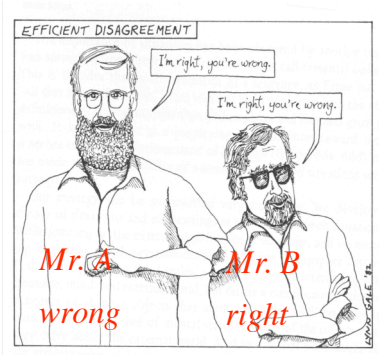
described situation becomes different from the above figure where a dog is biting the other.

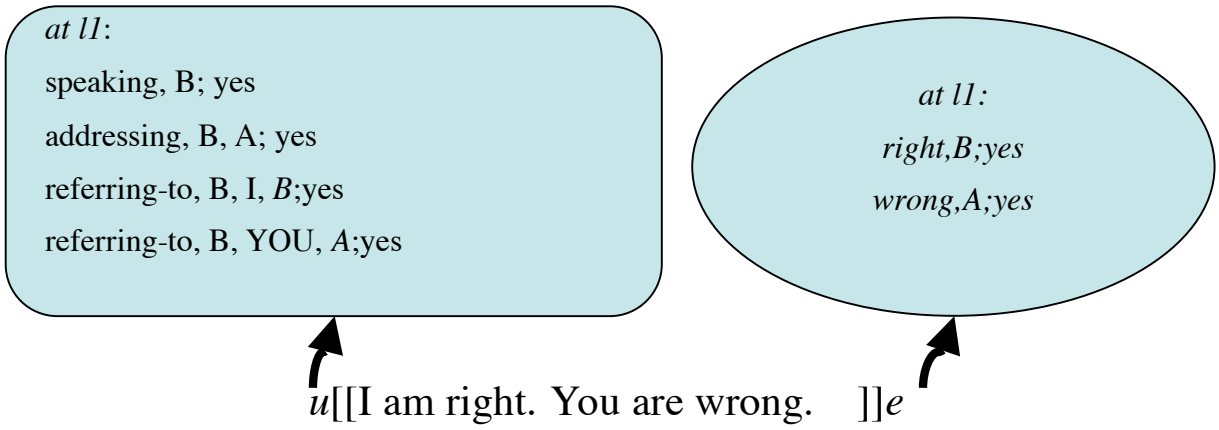


at l1:
speaking, A; yes
addressing, A, B; yes
referring-to, A, I, *A*;yes
referring-to, A, YOU, *B*;yes

at l1:
right,A;yes
wrong,B;yes

u[[I am right. You are wrong.]]*e*





The above two figures shows the interpretation of the same utterance "I'm right. You are wrong," which I already used as an example of a situated utterance. The difference of the figures is in their discourse situations. In the discourse situation of the first figure, the speaker is A and the hearer is B. On the other hand, the speaker is B and the hearer is A in that of the second figure.

The difference of the discourse situation causes the difference of the described situation because the words "I" and "YOU" are connected to the different real things: I = A and YOU = B in the first figure, and I=B and YOU=A in the second figure. As a result, the utterance in the first figure describes that A is right and B is wrong. Also, that in the second figure describes that B is right and A is wrong.

4.2.5 Applications for Actual System

The theory of situation is significant to design a communication system for the real world. However, the following questions arise to employs the theory on our systems.

- How to recognize a situation
- How to select relevant information
  - ➢ The selected information forms a discourse situation.
- Robot must recognize relevant situation from sensor data.

4.3 Focus of Attention

4.3.1 Requirement of Focus of Attention

As the theory of situation suggest, the meaning of a language expression heavily depends on a situation. Focus of attention is important idea to deal with the expressions under a recognized situation.

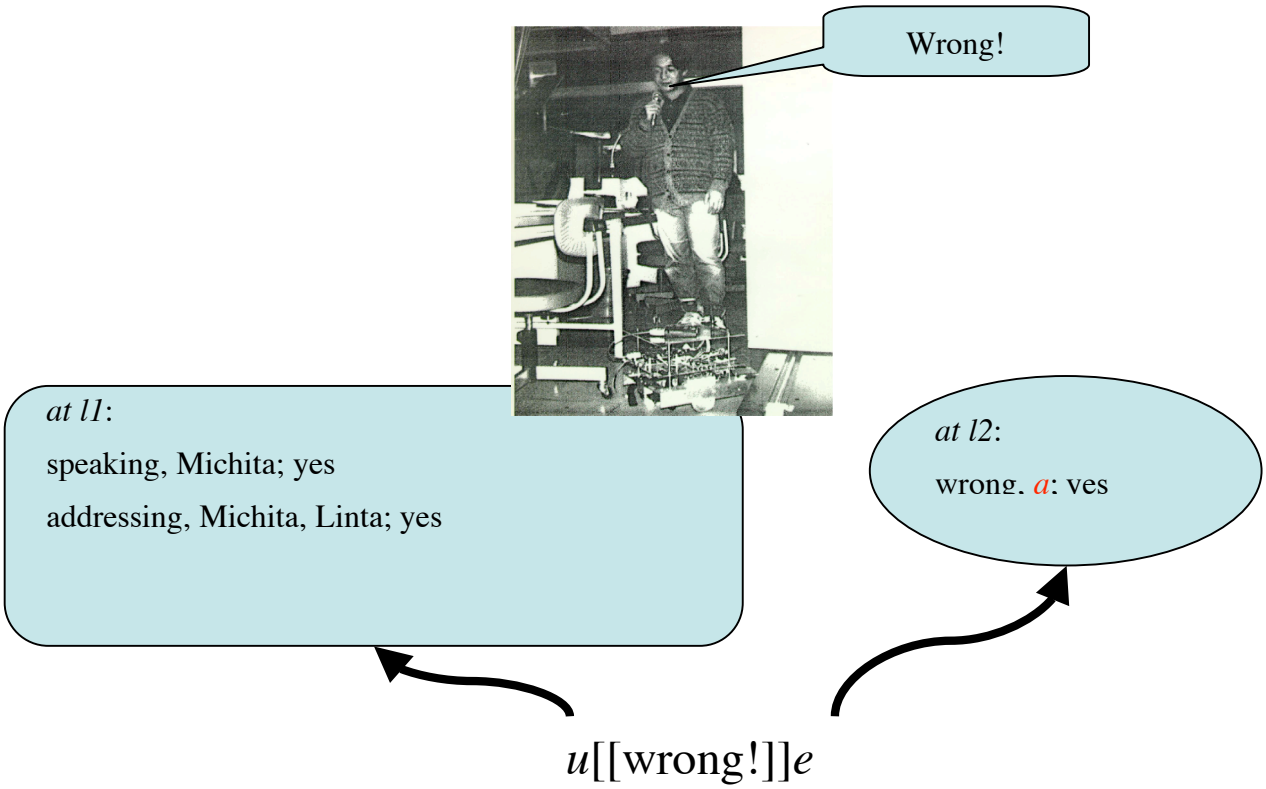
- Natural Language and Situated Expression
  - Situated logical expression introduced by Situation and Attitude
- Focus of Attention
  - Select relevant information from sensor information
  - Construct relationship between situation and language

Linta-I is a good example to consider the relationship between languages, situations, and focus of attention. That is

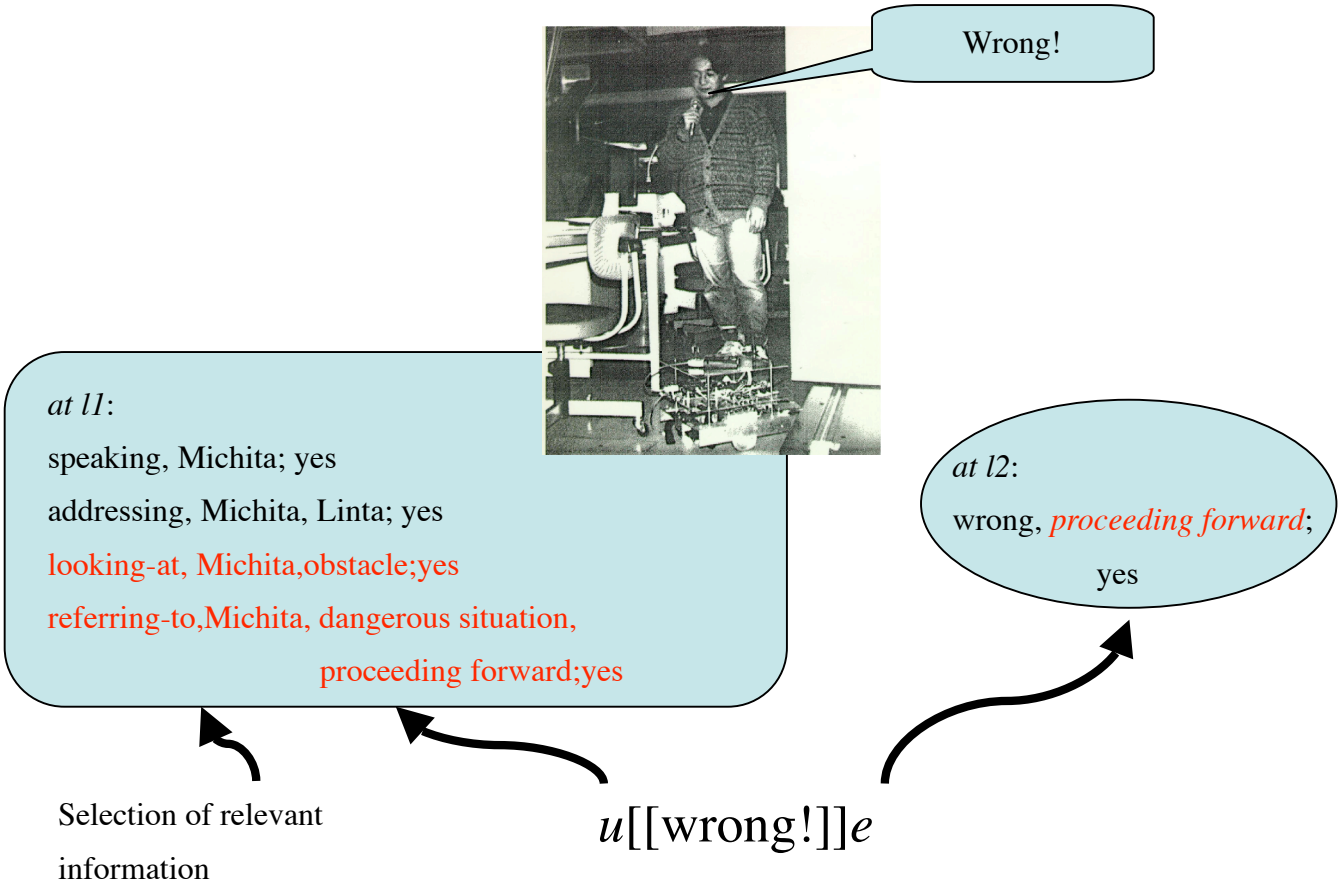
- How to recognize a situation
- How to select relevant information for the expression.

The focus of attention takes a role to form a discourse situation. The robot extracts described situations from given utterances and the discourse situation. Here What the robot must done to achieve the utterance interpretation is to recognize relevant situation from sensor data and make the discourse situation.

4.3.2 Application for Actual System



The figure shows the interpretation based on the theory of situation when Linta-I interprets an utterance “wrong.” Since the discourse situation include only who is a speaker and who is a hearer, Linta-I cannot identify what is wrong. The discourse situation results in the described situation “wrong, *a*.” That is, the variable *a* denotes that Linta-I cannot find the wrong point.



On the other hand, Linta-I infers two facts in this discourse situation. Those are the facts that Michita is looking at an obstacle and that Michita is referring to the robot’s action “proceeding” as a dangerous situation. Since the utterance “wrong” may indicate a dangerous situation, Linta-I can determine that what is wrong is proceeding forward.

The facts added by inference are an example of focus of attention. If the robot pays attention to the front direction when proceeding forward, it obtains the same discourse situation as the example. As a result, the focus of attention makes the robot understand the situated utterance.

4.3.3 Attention Mechanism

Attention Mechanism selects relevant sensor data to recognize current situation. The difficult to select sensor information is what information a system uses for the selection. We can use many types of information for the selection. For example, the context of a

conversation or the action of a robot can be employed when we develop a dialogue system for the robot. This section shows an attention mechanism which employs robot's actions for selecting relevant information.

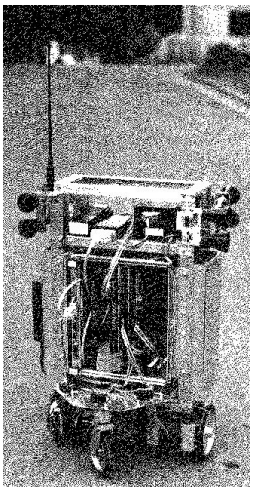
The remaining part of this section introduces a dialogue system named Linta-II which employs an attention mechanism to interpret user's utterances. The attention mechanism selects sensor information relevant to current human robot interaction. It has two types of attention.

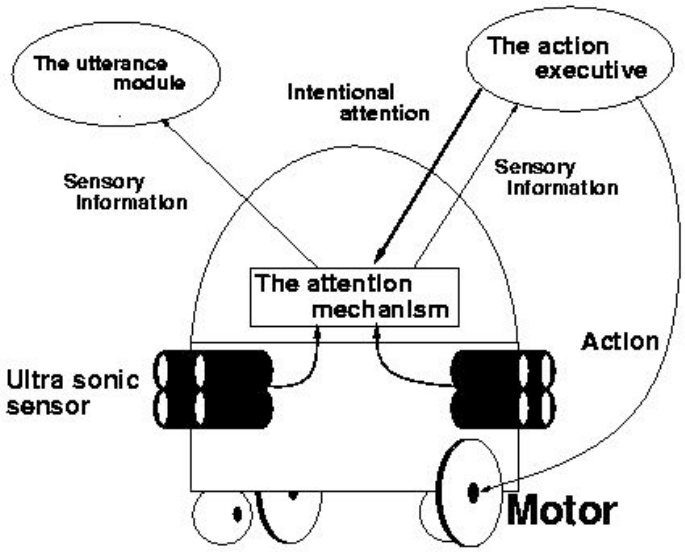
- Top-down attention which selects sensor data related to a current robot action or plan.
- Bottom-up attention which selects sensor data in response to an event around a robot.

The attention mechanism provides the following two benefits.

- Decrease of the use of constraints: When a dialogue system interprets an utterance, it uses constraints which are rules to select relevant sensor information and establish a discourse situation. However, the action mechanism selects relevant information prior to using the constraints.
- Easy to design the system: Many control modules of a robot have effects on each other through the information selected by the attention mechanism. Since they do not communicate directly, we can design each module separately.

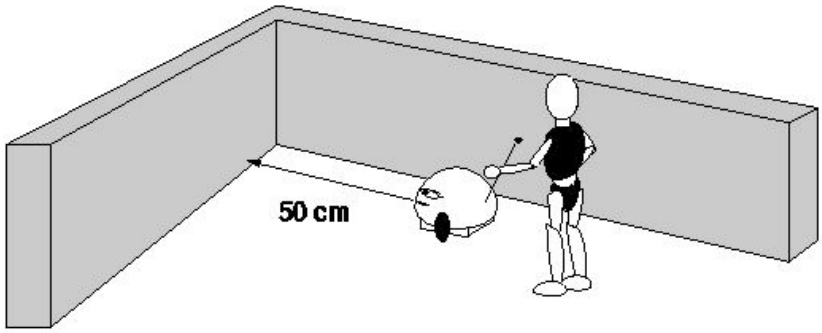
4.3.4 Linta-II





The figure and the picture show a dialogue system named Linta-II. Linta-II is implemented on the autonomous robot ASPIRE which is shown in the picture. Linta-II consists of the attention mechanism, the action executive, and the utterance module. The attention mechanism extracts information of an environment from sensors and

gives it to the utterance module and the action executive. The action executive selects relevant sensor information by giving the attention mechanism top-down attention.



Let's consider utterances generated at the situation of the above figure. The utterances are as the following.

- Ex. 1.a User: "What's next?"
  ASPIRE: "We cannot proceed toward front direction soon."
- Ex. 1.b User: "What's next?"
  ASPIRE: "We cannot turn right."

The example 1.a and 1.b shows the effect of top-down attention. There is a difference between the utterances of ASPIRE even though the user says the same utterance. The difference between the utterances of ASPIRE comes from the difference of attention which ASPIRE has.

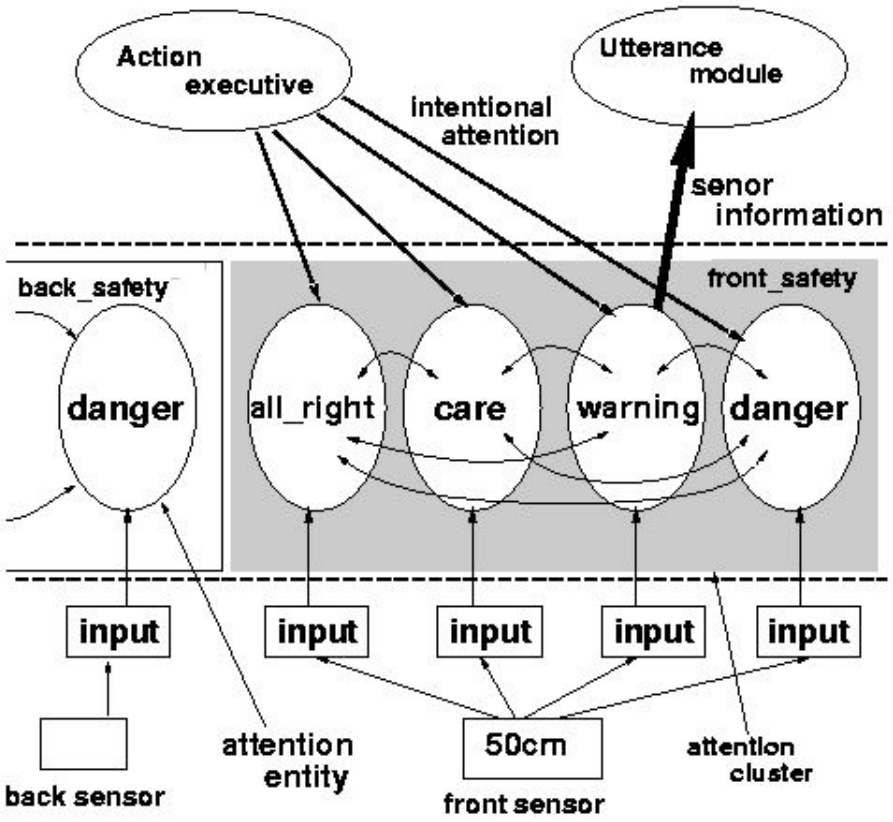
In the example 1.a, ASPIRE and the user moves forward. Since the action executive gives top-down attention to the attention mechanism, it gives a front sensor priority over other sensors. The selection of the front sensor is the reason why ASPIRE generates the utterance related to the front information.

The utterance of the example 1.b is also generated by ASPIRE at the same location as the example 1.a. However, ASPIRE does not move in this example and is searching a way for moving. When the user gives the utterance, the action executive just checks the right sensor by giving top-down attention to the attention mechanism. Since the utterance module has only information from the right sensor at that time, ASPIRE generates the utterance related to the right direction.

- Ex. ASPIRE: "I'm surprised!"

The utterance is an example of bottom-up attention. It is generated when an object falls in the left side of ASPIRE. The sudden appearance of the object causes bottom-up attention. The attention mechanism informs the utterance module and the action executive of the appearance of the object. In the example, ASPIRE generates an utterance related to the object.

4.3.5 Attention Mechanism of Linta-II

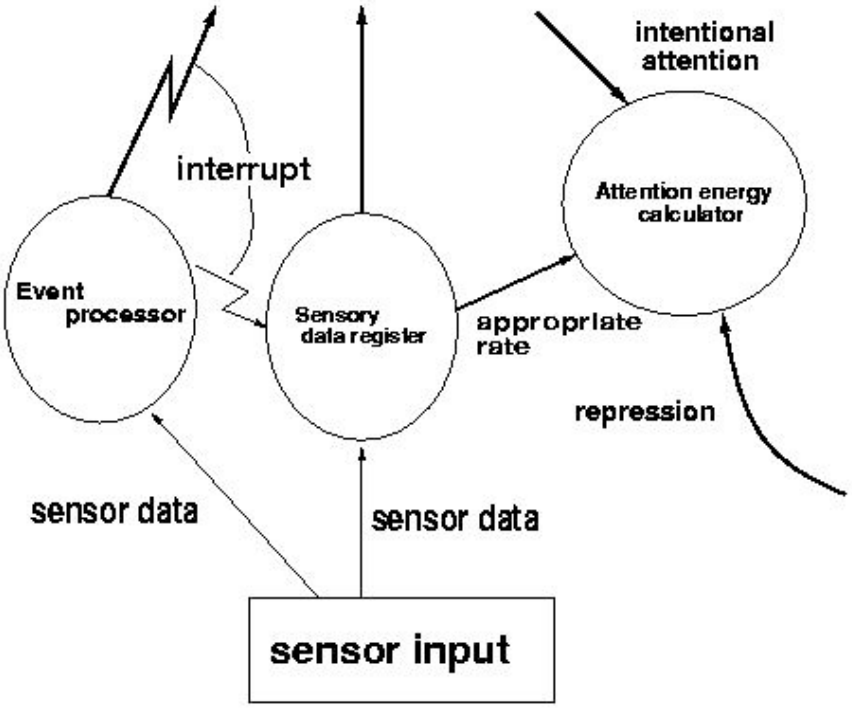


The figure shows the structure of the attention mechanism used by Linta-II. The attention mechanism consists of the sets of attention entities. Each set denotes the feature of sensor information and each entity expresses the value of the feature.

The figure shows the feature of the front_safety which has four values (entities): all_right, care, warning, and danger. Each entity has attention energy. The entity which has highest energy gives formation to the utterance module and the action executive.

In the figure, the action executive gives top-down attention to the feature front_safty. According to the top-down attention, the entities belonging to front_safety have higher energy than the other entities belonging to other features. Also, the entities of front_safety are given sensor data 50cm from the front sensor. While each entity calculates its own energy from the sensor data and top-down attention, they inhibit each other. An entity which has highest energy wins the conflict and gives its value to the action executive and the utterance module. In the figure, the entity front_safeity(warning) is a winner of the conflict.

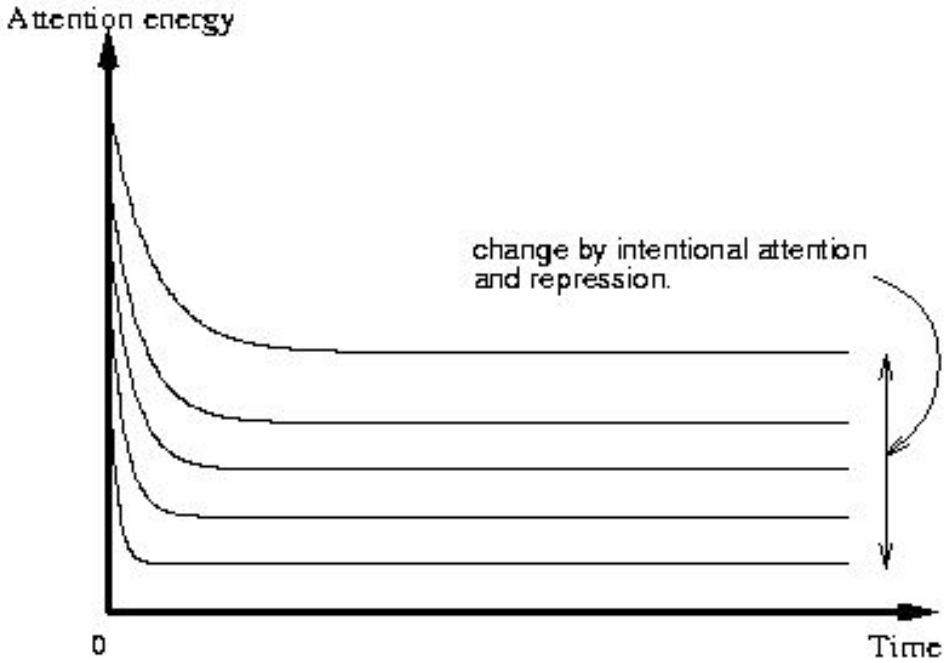
Each attention entity has a structure like the next figure: an attention energy calculator, a sensory data register, and an event processor. The sensor data register has its own value which is given to the utterance module and the action executive when



the entity has highest energy. The value of the energy is calculated at the attention energy calculator. The event processor detects a sudden change in the sensor data and interrupts the utterance module, the action executive, and the sensor data register. The activity of the event processor achieves bottom-up attention.

The attention energy calculator calculates the energy based on the following equation.

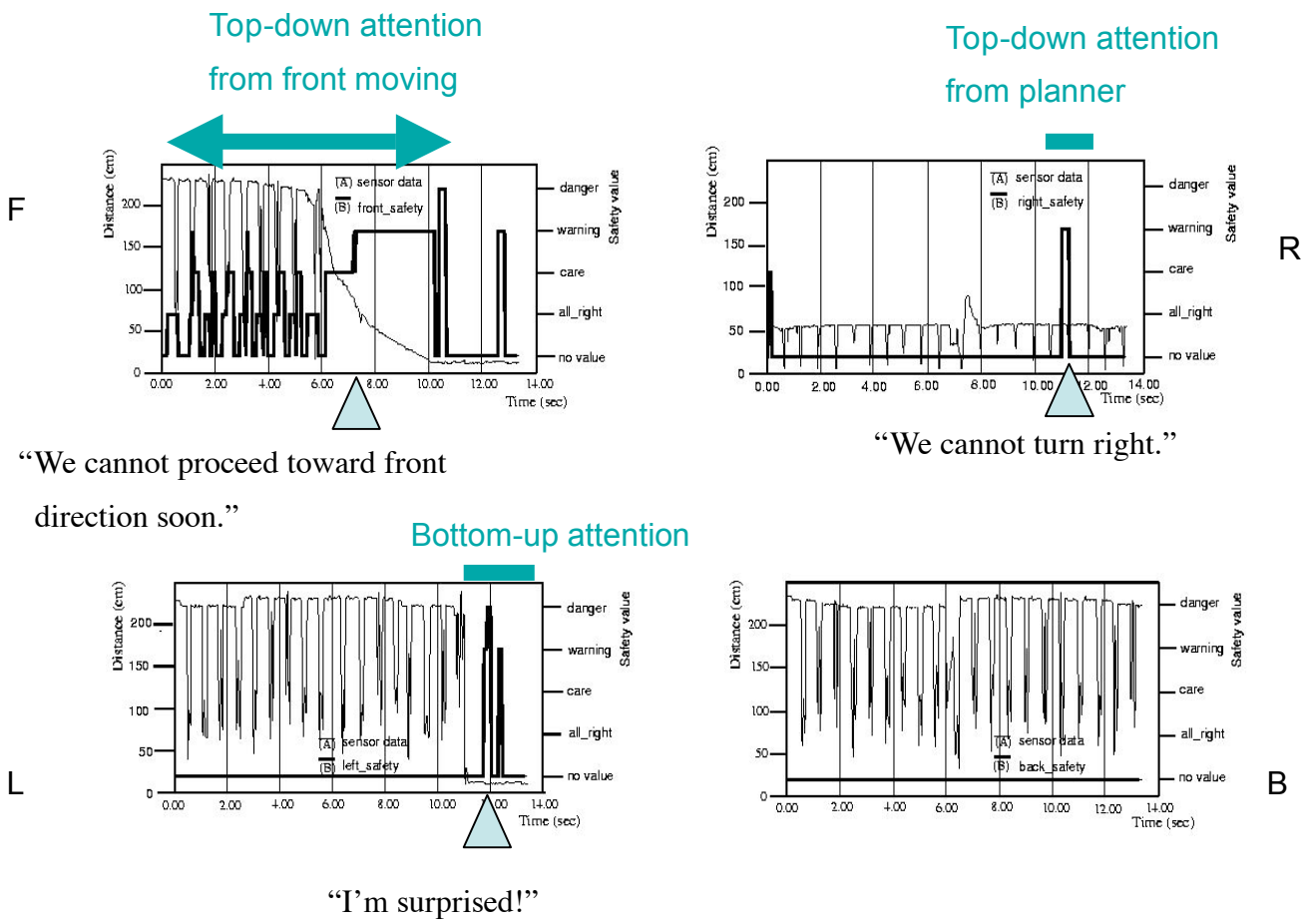
$$f(t) = Ae^{-t/(IE-\sum_j RR_j)} + (IE - \sum_j RR_j)$$



IE denote the value given by the action executive as top-down attention. *RRj* denotes inhibition from the other entities. The value of the inhibition is each energy value of the other entity. The value of *A* comes from the evaluation of sensor data. Since each entity has its own evaluation function, the value is determined based on given sensor data. As a result, the energy of an entity becomes higher when it is given top-down attention and gets appropriate sensor data. Also, since the entity which has highest energy suppresses the other entities through *RRj,* it has highest energy in the entities.

Moreover, the value of energy decreases according to time. The duration of the high energy also depends on $IE - \sum_j RR_j$. That is, it depends on top-down attention and suppressions from other entities.

4.3.6 Example



These graphs show the sensor data of front (F), left (L), right (R), and back (B) ultrasonic distance sensors. The thin line of each graph shows the distance from ASPIRE to an obstacle or a wall. The thick line shows the value of attention: no_value, all_right, care, warning, and danger. I captured the data while ASPIRE moves in a

corridor as shown in 4.3.4.

At first, the robot moves forward. Since the action executive gives top-down attention to front_safety of the attention mechanism, the graph F shows the attention value all_right or care. On the other hand, the attention values of the other graphs are no_value because the attention mechanism does not pay attention to the other direction.

However, when the robot approaches the wall, the value of front_safety becomes danger and the robot stops. If a user asks the robot "what's next" at this situation, it answers "we cannot proceed forward" by referring to the front sensor.

After stopping, the robot searches a way for proceeding. The graphs R shows that the action executive gives top-down attention to right_safety to check a right direction. However, the value is waring. If the user asks the robot the same question at this moment, it answers "we cannot turn right."

The graph L shows that an object suddenly appears in the left side of the robot. In response to the event, the attention mechanism generates bottom-up attention and generates an utterance "I'm surprised."

The graph B shows that the attention mechanism never pay attention to the back sensor. The no_value of graph B indicates the lack of attention to the back.

4.3.7 Conclusion of Linta-II

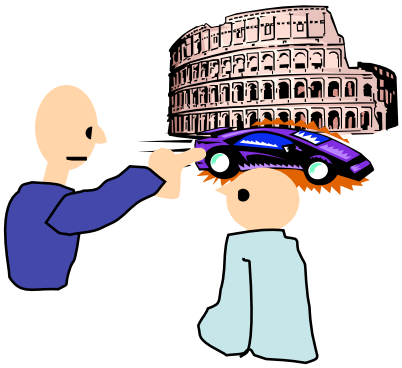
- Linta-II employs attention mechanism to obtain a situation for utterance generation.
- Attention mechanism: Top-down attention and Bottom-up attention
  - Extract situated meaning under the attention
- A system designer can develop a dialogue system and an action system independently. These are connected through the attention mechanism.
- Problem
  - Robot obtains the situation selfishly.

5 Joint Attention and Theory of Mind Model

The previous section explained an attention mechanism which is used for dealing with situated utterances in human-robot communications. However, the attention mechanism is not sufficient in the actual interaction because it selects relevant information selfishly. If a system wants to get the same interpretation as the user, it must pay attention to the same situation as the user. The phenomenon where two or more persons pay attention to the same target is called joint attention in the social psychology. The topic of this section is as follows.

- Joint Attention
  - Share information about physical world
  - Sonja: joint attention with a computer
- Development of Infants' Joint Attention Skill
  - Gaze-awareness and Joint Attention
- Theory of Mind Model and Infants' Joint Attention Skill
- Human-robot interaction and theory of mind model.
- Joint Attention Mechanism
  - Joint attention with gaze direction
  - Joint attention with gaze and pointing gestures

5.1 Joint Attention



Joint attention is a cognitive phenomenon when tow persons pay attention to the same thing. The figure shows a situation where a person talks about a car passing in front of him. A hearer beside him must also direct his attention to the car to interpret his utterance. In the situation, joint attention is established between them.

Establishment of joint attention between a human and a system requires

- We find out an appropriate setting or
- Give it a mechanism to generate the joint attention.

This subsection considers a method to establish joint attention for a computer system.

5.1.1 Joint Attention with Computer

This subsection employs an interactive system named Sonja which is based on a role playing game called Amazon. Sonja establishes joint attention by utilizing a game scenario of Amazon. Since the scenario guides a user's attention to a relevant part of the scene, Sonja easily assumes that the user pays attention to the relevant part. Because of the assumption, Sonja can establish joint attention by just directing its attention to the same part of the game scene.

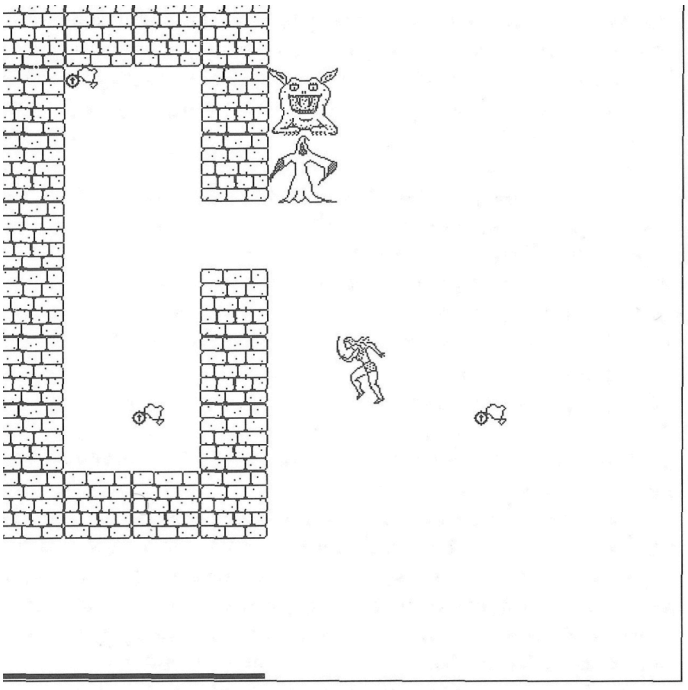
5.1.2 Sonja

Sonja interprets user's utterances based on the scene of a video game. Since Sonja takes visual-guided activities, it assumes that the user gives utterances to it based on the scene.

The significant points which Sonja employs establishing joint attention are as follows.

- The scene and scenario of Amazon draw a user's attention to a certain part of the game scene. In other words, this system is an example where a system designer finds appropriate setting to draw human's attention.
- Sonja has attention mechanism to obtain relevant information from the scene of the game. What is important for the attention mechanism to establish joint attention is that it has knowledge where a user pays attention in each scene of the game. Also, the attention mechanism is a simulated but realistic visual system.

5.1.3 Scene on Amazon



Sonja imagines a game scene where a player controls a woman warrior named

Amazon. There are walls, various enemies, and tools in the scene. It also imagines that the other person may give advice to the player as if you sit by him/her.

5.1.4 Role of Sonja

Sonja is a player of Amazon. A person (a user) is an adviser of Sonja. He/she gives advice to Sonja as an adviser sitting by the next of Sonja.

A visual system Sonja has is similar to human's. The similarity is the basis of establishing joint attention.

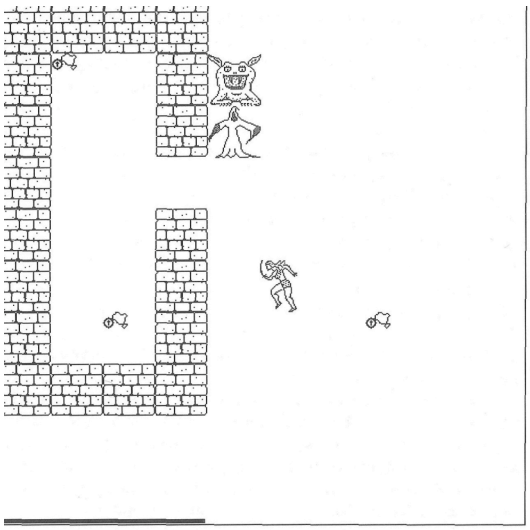
Moreover, Sonja models the person's strategy to select his/her behaviors.

5.1.5 Knowledge of Sonja

To achieve human-system interaction based on joint attention, knowledge Sonja has is as follows.

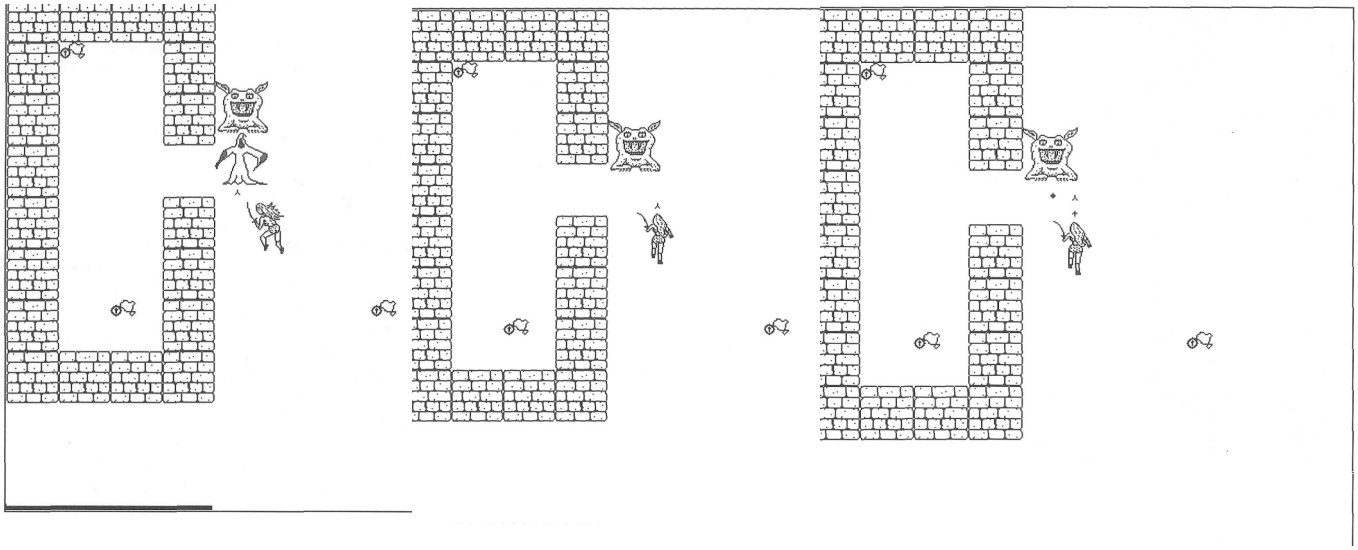
- Sonja knows
  - How to kill off monsters.
  - How to pick up and use tools.
  - How to get about in the dungeon maze.
- Sonja does not know
  - How to choose the right thing to do when there are several plausible actions available.
- Instructions from an advisor (a human) have effect on the selection.

5.1.6 Example of Instruction

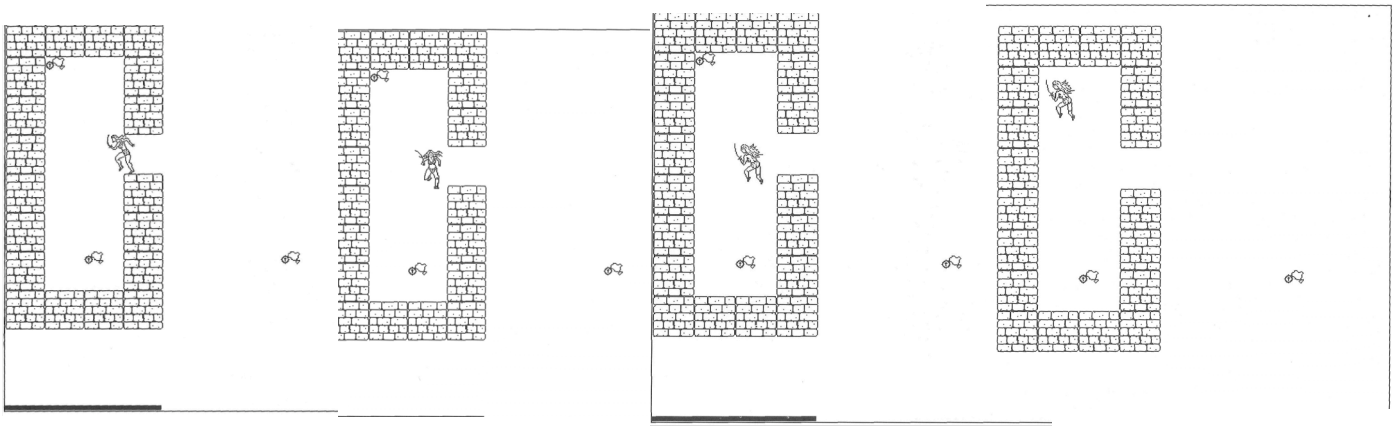


This picture shows an example of playing Amazon where there are three amulets, a ghost, a monster, and walls. At this scene, a user can give it advice "Go in and get the

amulet!" Since Sonja cannot identify what amulet is relevant, it tries to take one of them randomly. However, the user gives the advice to Sonja in this example.



On the way to the room, Amazon meets the ghost and the monster. There are several ways to fight with them. At first, it kills the ghost by throwing a star. After this, the next enemy is the monster. When fighting with it, a user gives Sonja advice "use a knife." Sonja selects a knife in response to the advice and battles with it (the right figure in the above three figures).



After killing enemies, Amazon goes in the room and tries to get the lower amulet (two figures in the left side of the above figures). At the situation, the user gives Sonja advice "No, the other one!" This advice does not include an exact action of Amazon.

The advice is an example of a situated expression. Also, humans frequently use such a situated expression when giving the player of a game. Since the player and the adviser focus on a part of the game scene, they can communicate by using the situated expression.

Sonja also must refer to situation for interpreting the situated advice. What the other one refers to and what action the adviser intends Sonja with the advice are crucial

information for the interpretation.

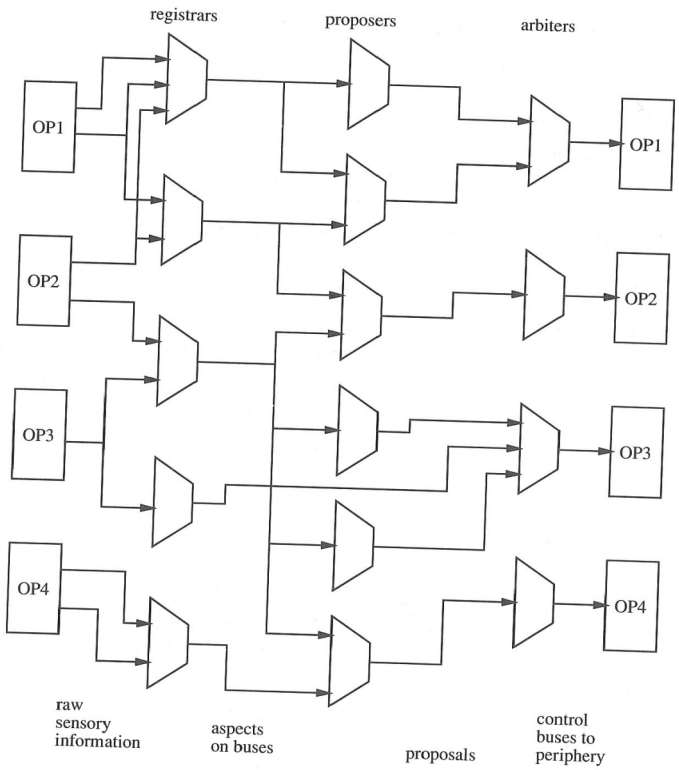
Since Sonja focuses on the lower amulet, it can easily notice that the other suggest the upper amulet. Moreover, since it also realizes that the current action of Amazon is getting amulet, it carries out “getting the upper amulet.”

Throughout the example, we can conclude that the interpretation of the situated advice is to select an appropriate behavior under the focused information of an environment. In short, Sonja must have the following functions to deal with the situated advice.

- Selection of available behaviors
- The same visual skill as a human

5.1.7 Hardware and Architecture

Sonja consists of a central system and a visual architecture.

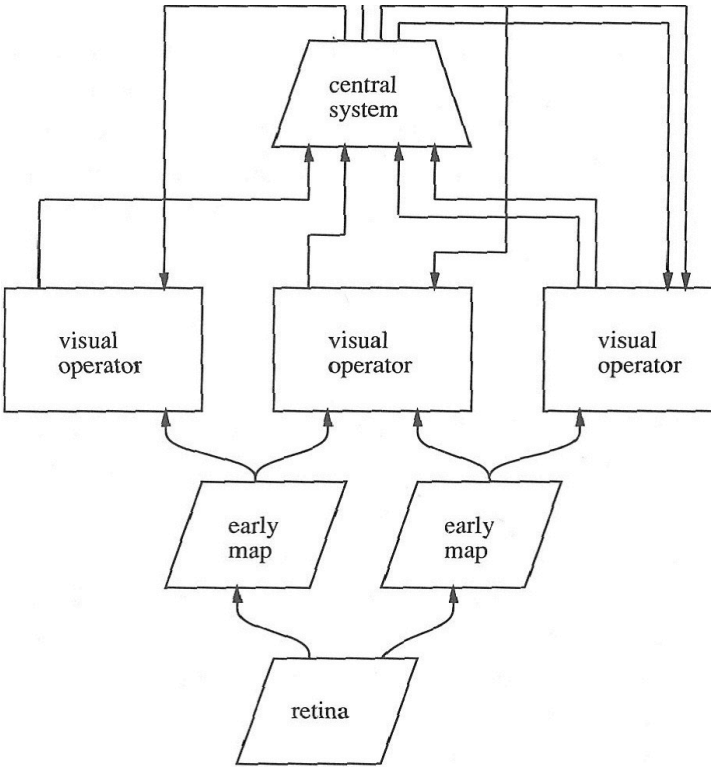


The central system has the structure of the above figure. It obtains sensory information and selects an appropriate behavior of Amazon. The components of the central system are registers, proposers, and arbiters. OPs denote operators which are used to control Amazon and the visual architecture.

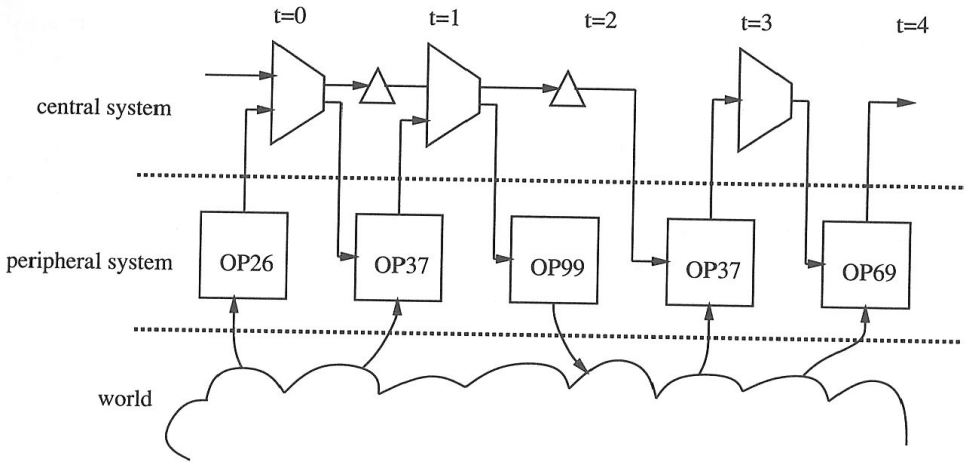
The registers stores sensory information and the proposers generate behaviors in response to the current sensory information. The arbiters select appropriate behaviors when a conflict arises between generated behaviors.

The visual architecture is a task-dependent vision system. And the visual functions

of the architecture are activated by central system directly.



This figure shows the overall structure of the visual architecture. The retina corresponds to the scene of the game. The early maps obtain basic visual features from the retina. Each visual operator is activated by the central system and obtains information relevant to the central system.



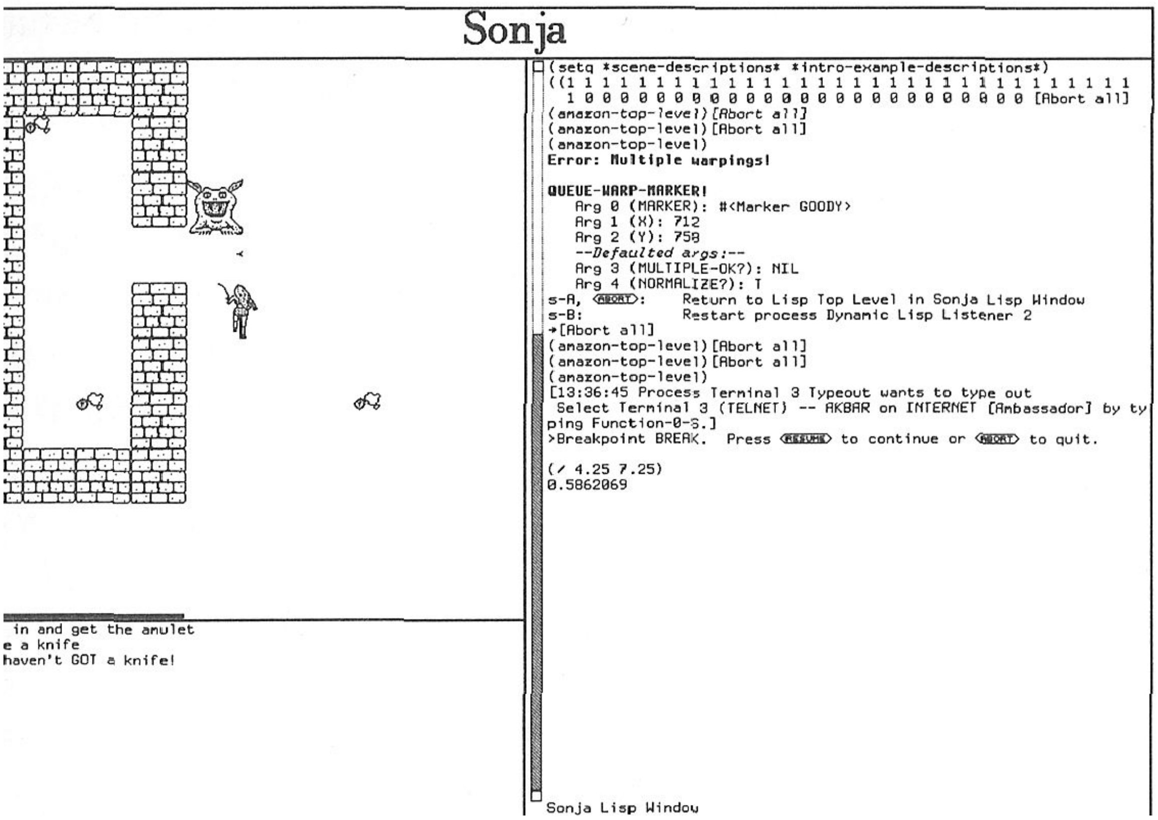
This figure shows the sequence of executions of Sonja. At first, the central system obtains world information with OP26 (a visual operator) and generates OP37 (a visual operator). The triangle between t=0 and t=1 is a temporary storage for storing the result of the central system. The central system generates OP99 (the control of Amazon) in response to data from the temporary storage and information recognized

by OP37. It executes OP99 at t=2. Also, it recognize the effect of OP99 by delaying the execution of OP37 until t=3. After that, it executes OP69 (a visual operator).

5.1.8 Instruction Use

The researchers of Sonja have conducted an experiment to investigate the interaction between people when they played video games. As the previous section has explained, the adviser used frequently situated expressions. Moreover, the expressions were based on the scene of the game. The situated expressions were supported by the same visual scenes when the player and the adviser were sitting next to each other.

The situated expressions had the following features: immediate and less than five length. For example, "pick up that items", "danger", and "other way." They have referred to information of the game scene.



This figure is a captured scene of the execution of Sonja. The window under the graphic is for the input of the advice of the user. The right window shows the execution of lisp code.

The next table shows the list of instructions which the adviser of Sonja can use. The first column indicates the exact texts of the advices. The second one indicates the lisp

Instruction	Instruction buffer(s) set	Field
Get the monster/ghost/demon	`kill-the-monster`	`type`
Don't bother with that guy	`dont-kill-the-monster`	
Head down those stairs	`go-down-the-stairwell`	
Don't go down yet	`dont-go-down-the-stairwell`	
Get the bones	`kill-the-bones`	
Ignore the bones for now	`dont-kill-the-bones`	
Get the *goody*	`pick-up-the-goody`	
	`register-the-`*goody*	
Don't pick up the *goody*	`dont-pick-up-the-goody`	
	`register-the-`*goody*	
Head *direction*	`suggested-go`	`direction`
Don't go *direction*	`suggested-not-go`	`direction`
Go around to the left/right	`go-around`	`direction`
Go around the top/bottom	`go-around`	`direction`
Go on in	`go-in`	
OK, head out now	`go-out`	
Go on in and down the stairs	`in-the-room`	
	`go-down-the-stairwell`	
Go on in and get the bones	`in-the-room`	
	`kill-the-bones`	
Go in and get the *goody*	`register-the-`*goody*	
	`in-the-room`	
	`pick-up-the-goody`	
Get the potion and set it off	`register-the-potion`	
	`pick-up-the-goody`	
	`use-a-potion` (chained)	
Scroll's ready	`scroll-is-ready`	
On your left! *and similar*	`look-out-relative`	`rotation`
On the left! *and similar*	`look-out-absolute`	`direction`
Use a knife	`use-a-knife`	
Hit it with a knife when it goes light	`hit-it-with-a-knife-when-it-goes-light`	
Use a potion	`use-a-potion`	
No, the other one	`no-the-other-one`	

Figure 5.2

expressions (instruction buffer set) which Sonja use in the central system. The third one indicates the type of arguments of the instruction.

The instruction buffer set is directly given to the arbiters of the central system. Since the proposers generate candidates for behaviors, the instruction has an effect on the selection of proposed behaviors.

What makes Sonja deal with the situated expressions is the way of the usage of the instruction in the central system. The candidate behaviors which the proposers generate restrict the referent of the situated expressions because the only information related to the candidates encounters the situated expressions at the arbiters.

The situated instructions are different from commands to a system. While the commands indicate the actions of the system, the instructions just have effects on the

selection of the actions. The relation between the instruction and the action is indirect. Sonja uses the instruction at the arbitration of action conflicts.

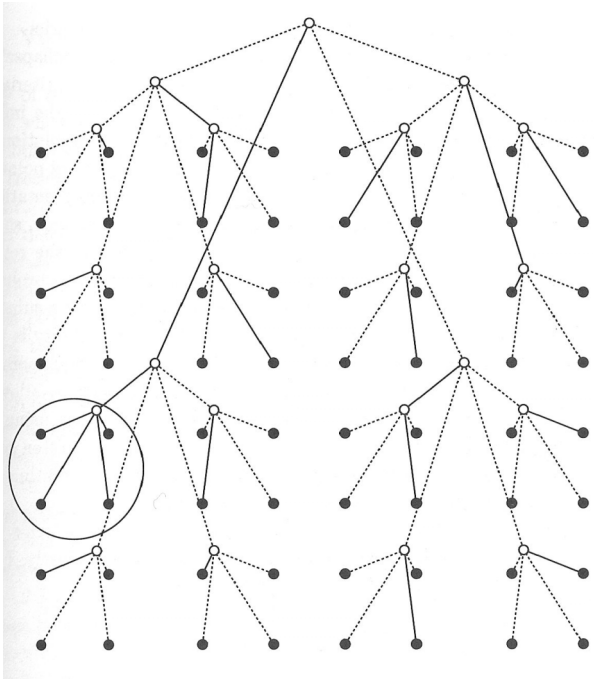
Let's remember the example of "go in and get the amulet!" There were three amulets in the example. At the situation, the proposers generate three actions each of which moves Amazon to each amulet. When an arbiter selects the actions, the instructions have an effect on it. Since the proprsers generate the candidates, the instructions never relate to the other actions. This mechanism of interpreting instructions achieves situated interpretation.

5.1.9 Vision System

The subsections from 5.1.9 to 5.1.13 explain the visual system which the visual architecture of Sonja employs. The vision system has the following functions and features.

- Visual attention
- Visual search
- Intermediate objects
- Specific operators

5.1.10 Visual attention

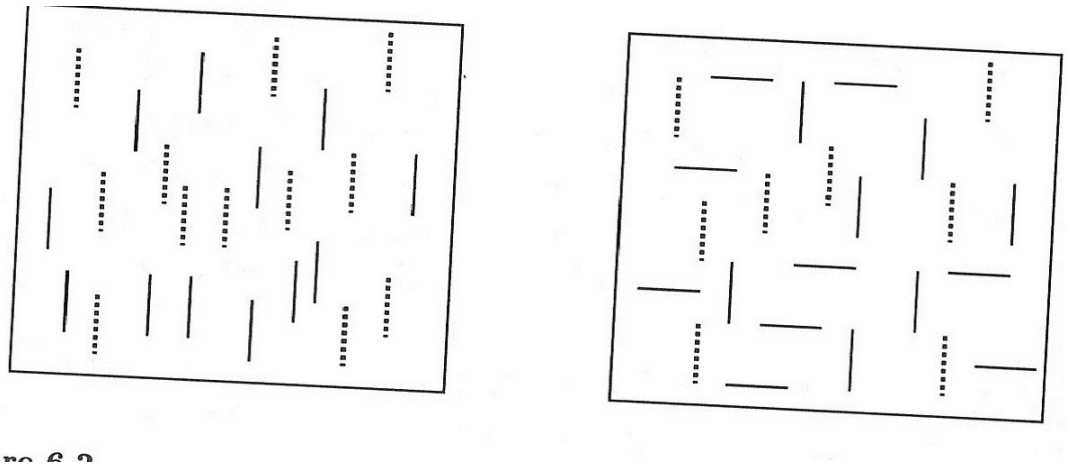


This figure shows a conceptual structure of visual attention. The matrix of the black dots corresponds to the retina of the visual architecture. Each black dot covers a small rectangle region of a game screen and is managed with the tree structure by the visual attention.

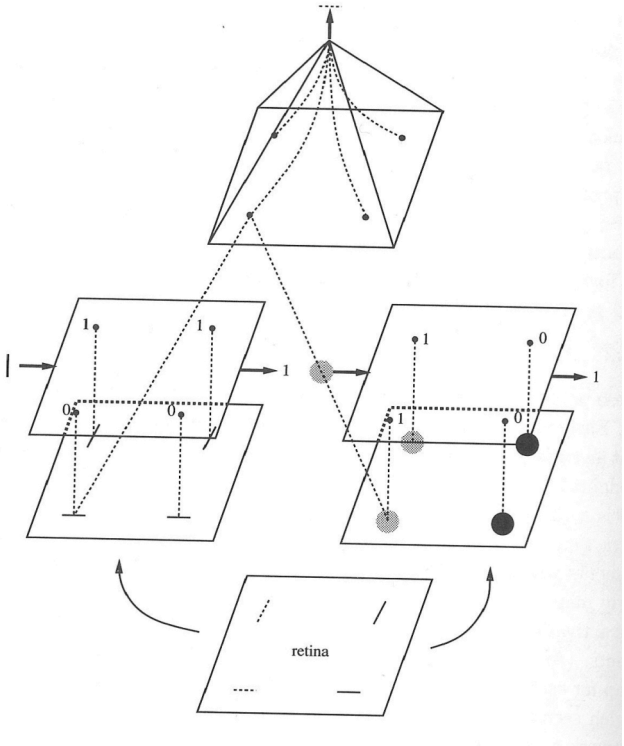
The visual attention directs Sonja's attention to the relevant information on the game

scene. In the figure, the solid lines from the root node (the white dot) express the attention Sonja directs. The four dots included by a circle are the target region of the visual attention. Also, the tree structure has intermediate nodes (white dots) which confirm whether or not the regions assigned to the nodes have properties which the central system requires. The region of visual attention is selected in the bottom-up manner along the tree structure toward the root node.

5.1.11 Visual search



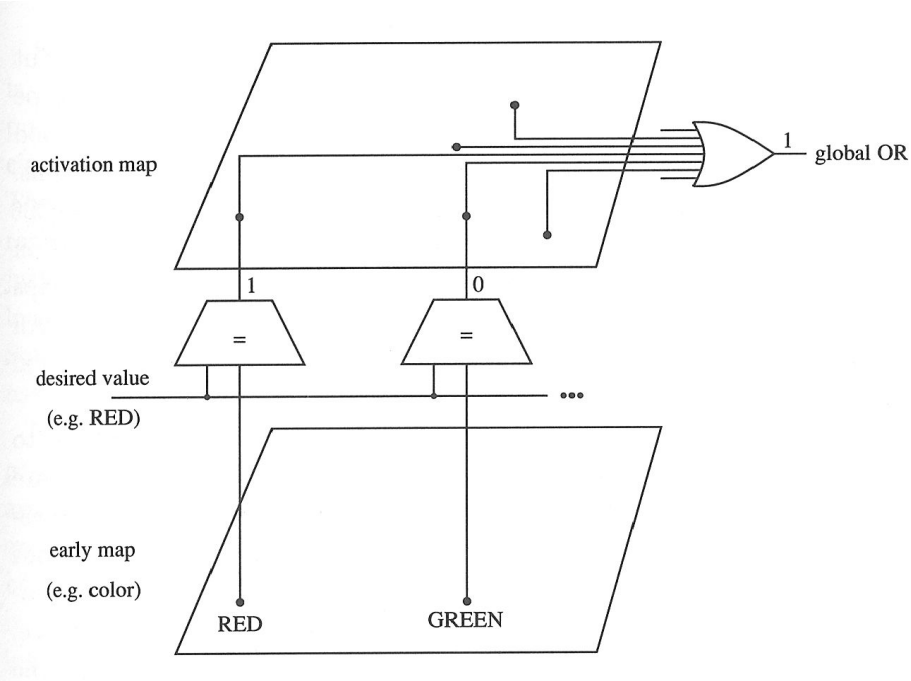
This subsection explains visual search of Sonja. These figures show the example of visual input. There are two types of lines in the left figure: solid vertical lines and dashed vertical lines. Also, there are three types: solid vertical lines, dashed vertical lines, solid horizontal lines. The features of lines are used for selecting a visual object by visual attention.



This figure shows the selection of visual object. There are four types of lines in the retina. The tree structure above the retina is a tree of visual attention. Also, the figure shows that there are two types of visual structures above the retina; one is early map (two rectangles above the retina) and the other is activation map (two rectangle above the early maps). The features which the vision system extracts are assigned to the early maps one by one. The left early map in the figure has charge of extracting the feature of direction (vertical or horizontal) of lines. The right one extracts the feature of their color (solid or dashed).

Also, the central system sets constraints to the activation maps for finding objects which it requires. Those are dashed vertical lines in the example. The left activation map is given a constraint for finding vertical lines. The right one is given a constraint for finding dashed lines. The results of finding the features are printed on the activation maps with 0 or 1. If there is a line which has the assigned feature on the activation map, the map outputs 1. Otherwise, it outputs 0. This method of the output corresponds to parallel searches on the retina. It makes the decision of Sonja faster than searching a target serially.

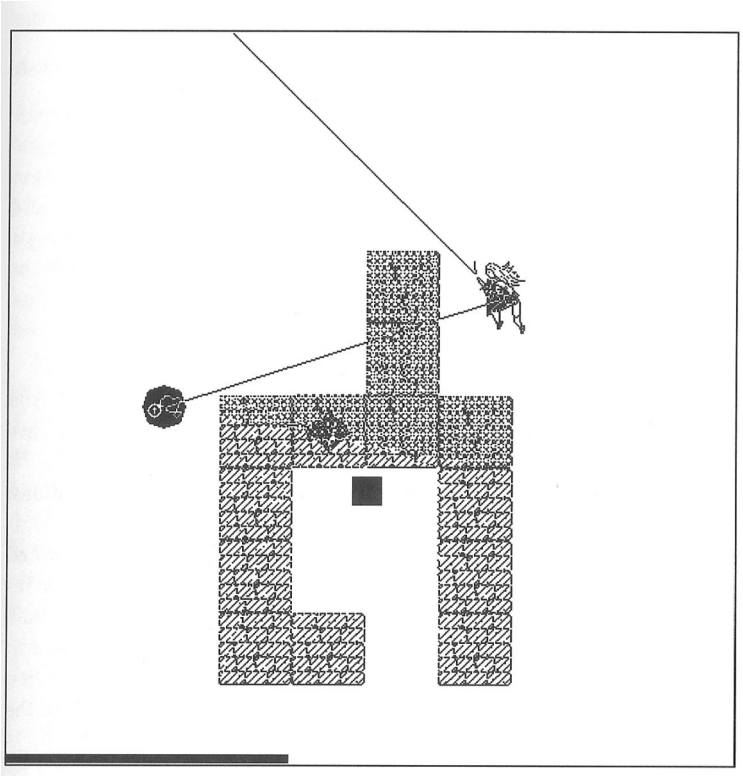
In addition, the figure shows that two features of a line (the left lower part of the retina) are connected to an intermediate node of the tree structure although almost other links are omitted in the figure.



This figure focuses on the relation between the activation map and the early map. There is an early map for extracting the feature of color. It has just extracted the features of red and green from the retina. Also, the central system set a desired value

(red in the example) to the activation map. The activation map has gates to check whether or not there are desired values on the retina. The global OR generates the output of the activation map in response to all results from the gates. The use of the global OR makes the activation map inform the central system of the existence of the desired value immediately.

5.1.12 Expressions of Game Scene



After recognizing objects from a retina, the visual architecture uses markers, lines, rays, and activation planes to express the location of the objects on 2D coordination. Those graphical items form a recognized scene of Amazon. The central system refers to the items to determine the behavior of Amazon.

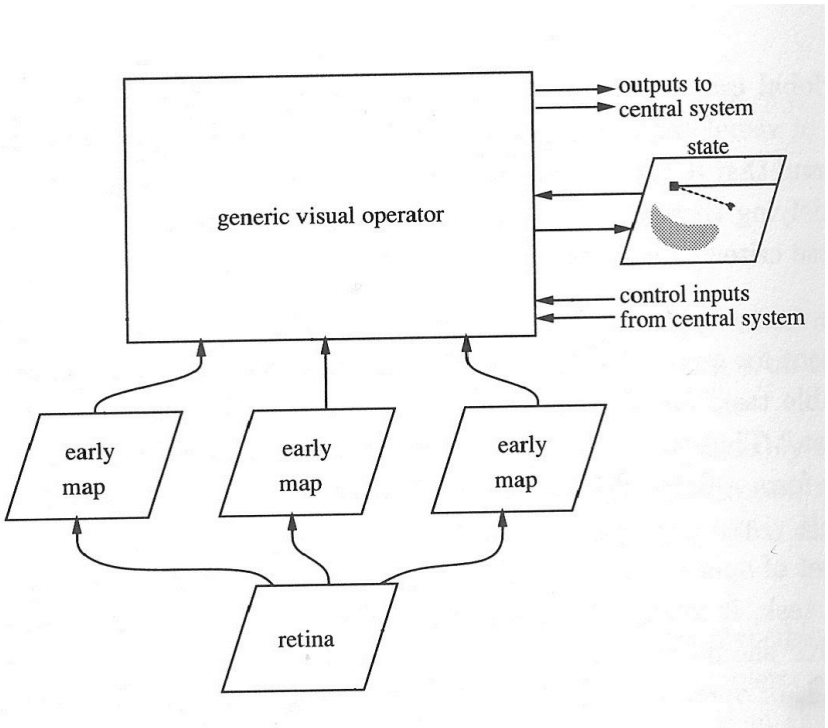
There are a right triangle, an octagon, a square, a diamond, a ray, a line, and an activation plane in the figure. The means of the items are listed in the table of the next page.

By referring to the table, we can understand the situation of the above figure. Amazon which is marked with a right triangle tries to go to an amulet which is marked with the octagon. The goal location are expressed by the line between Amazon and the amulet. A part of the wall has a grey region. The region is an activation plane which expresses an obstacle. Also, the obstacle of the wall is marked with a diamond. A ray from Amazon expresses a way to which Amazon should move to reach the goal. The ray is a result of

Display	Entity	Global variable
Markers:		
right triangle	*the-amazon*	`*amazon-marker*`
cross	*the-target*	`*target-marker*`
pentagon	*the-monster*	`*monster-marker*`
hexagon	*the-pile-of-bones*	`*bones-marker*`
septagon	*the-stairwell*	`*stairwell-marker*`
octagon	*the-goody*	`*goody-marker*`
nonagon	*the-fireball*	`*fireball-marker*`
square	*the-obstacle*	`*obstacle-marker*`
up triangle	*outside-the-doorway*	`*doorway-outer-marker*`
down triangle	*inside-the-doorway*	`*doorway-inner-marker*`
diamond	*something*	`*opportunistic-marker*`
Line:		
	amazon-to-goal	`*goal-line*`
Rays:		
	heading-direction	`*heading-ray*`
	demon-direction	`*demon-ray*`
Activation planes:		
increasing hatch pattern	*the-obstacle*	`*obstacle-plane*`
decreasing hatch pattern	general purpose	`*temp-plane*`
grey wash	*the-room*	`*room-plane*`

planning a rout avoiding the wall. Also, there is a square inside the wall. It expresses an obstacle.

5.1.13 Interaction between the central system and the visual system



This figure expresses a relation between the central system and the visual architecture. The generic visual operator generates the state of a current game by referring to the early maps. The central system uses the generated state to control Amazon (control inputs from central system in the figure) and to determine where the central system focuses on (outputs to central system).

The control commands are as follows.

- navigate-to-goody
- navigate-to-monster
- navigate-to-stairwell
- navigate-to-door
- navigate-to-bones

Those commands navigate Amazon to these goals like goody, monster, stairwell, door, and bones. Since the actual locations of the goals are managed as the graphical items by the visual architecture, the central system does not need to use complete forms of the commands.

5.1.14 Interaction with Sonja

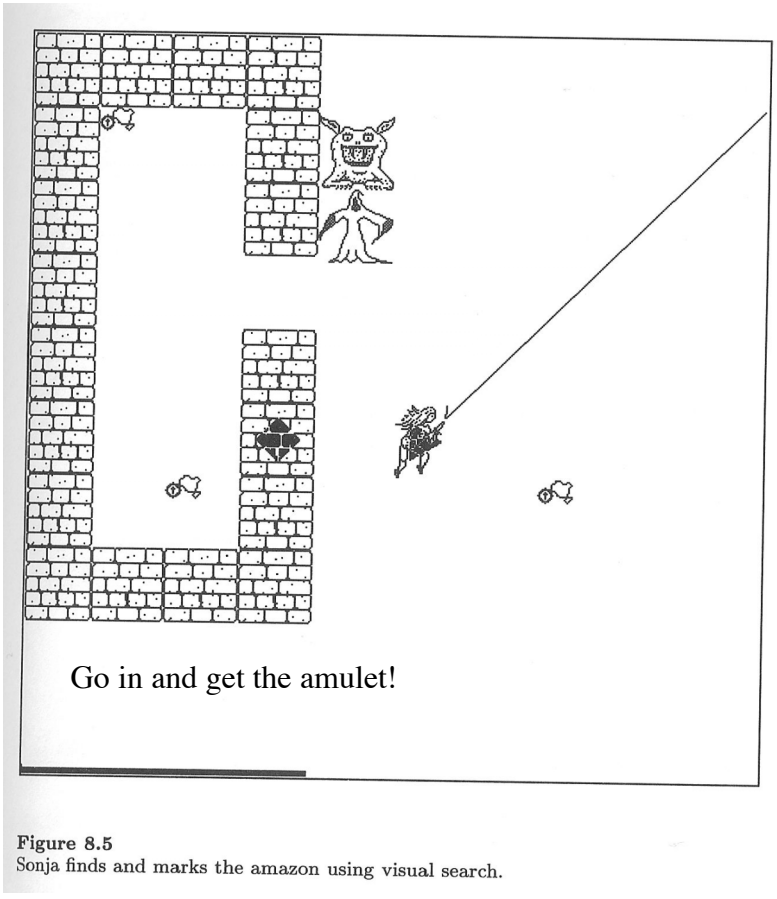


Figure 8.5
Sonja finds and marks the amazon using visual search.

This subsection shows an example of interactions between a human and Sonja. The

first figure is the same one as I already used as an example of Sonja's behavior. The visual architecture notices the location (the right triangle) and direction (the ray) of Amazon and an object (the diamond) whose types the architecture is not conscious of. Also, it does not realize that there are amulets, a ghost, and a monster. In other words, it directs its attention to only Amazon and the wall.

At the situation, a human give a command "go in and get the amulet!"

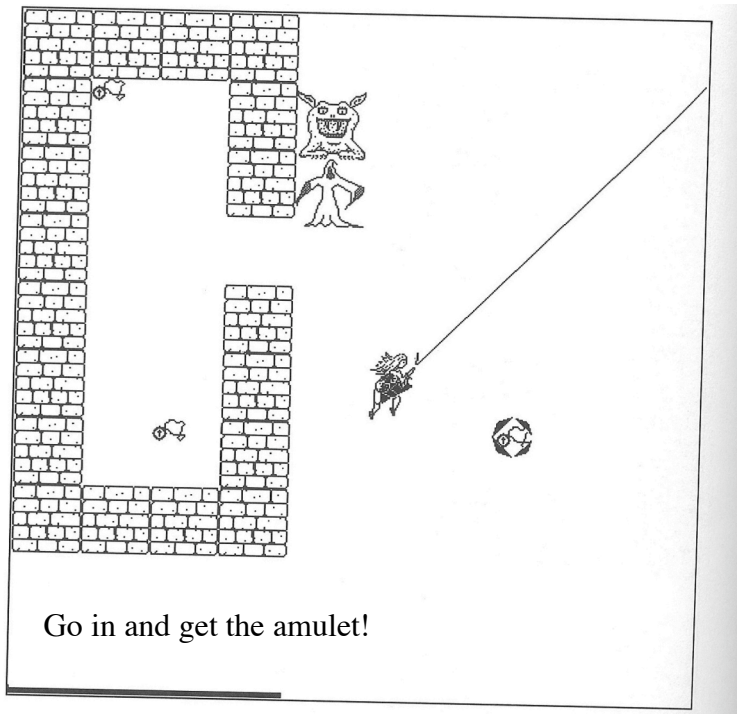


Figure 8.6
Sonja finds an amulet with visual search. However, this amulet is not in a room as required.

Since the visual architecture have not focused on amulets, the visual search searches an amulet. The above figure shows that the visual search has just found an amulet in front of Amazon. However, the amulet does not satisfy the given command because it indicates the amulet inside a room by the expression "go in."

The next figure shows that the visual search have just found an amulet inside the room. To find the amulet, the visual search uses activation planes to recognize the area inside the room.

There are two types of activation planes which used for recognizing the area; one (increasing hatch pattern) is used for the recognition of a wall, and the other (grey wash) is for the recognition of the inside area. By spreading the area of the increasing hatch pattern inside the edges of the wall, the pattern fills the wall itself. After recognizing the wall, the visual search spreads the grey wash inside the increasing hatch pattern. As a result, it obtains the area of the room as the area of the grey wash.

Then, it finds the amulet inside the gray wash. Also, the central system sets the amulet to the goal of Amazon. The goal is expressed by a line between Amazon and the amulet.

Moreover, the visual search also finds the ghost (the pentagon) and an end point (the diamond) of the wall while it turns the ray around Amazon to find the goal of Amazon.

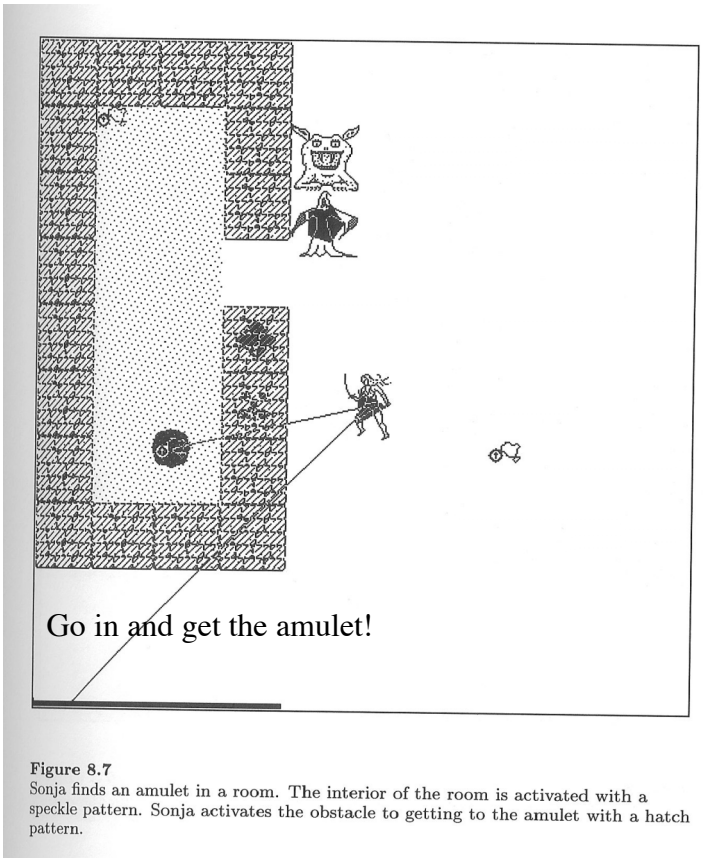


Figure 8.7
Sonja finds an amulet in a room. The interior of the room is activated with a speckle pattern. Sonja activates the obstacle to getting to the amulet with a hatch pattern.

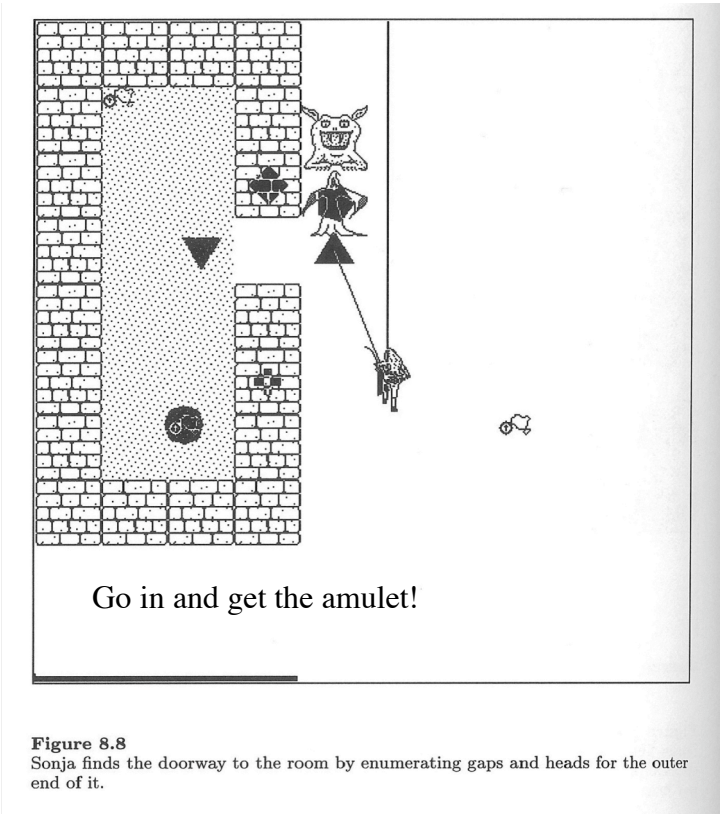


Figure 8.8
Sonja finds the doorway to the room by enumerating gaps and heads for the outer end of it.

The next figure shows that the visual architecture finds the doorway to the room. The door is found with the gap between the wall. Since it has already found the lower end of the wall, it just finds the other side (the diamond) of the end of the wall in the figure. As a result, it finds the doorway (an up triangle which denotes the outside of the room, and a down triangle which denotes the inside of the room) from the gap. The central system make Amazon head for the door (the up triangle).

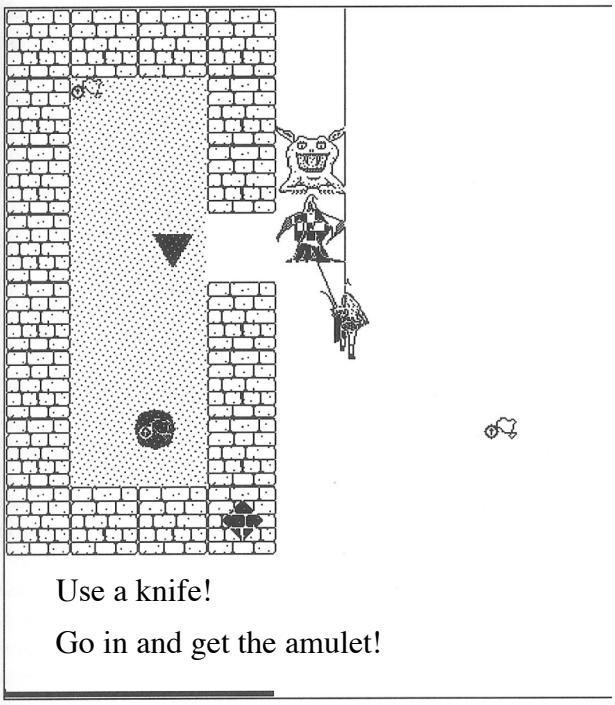


Figure 8.9
Sonja kills the ghost when it gets close enough to be a threat. A shurike immediately above the amazon.

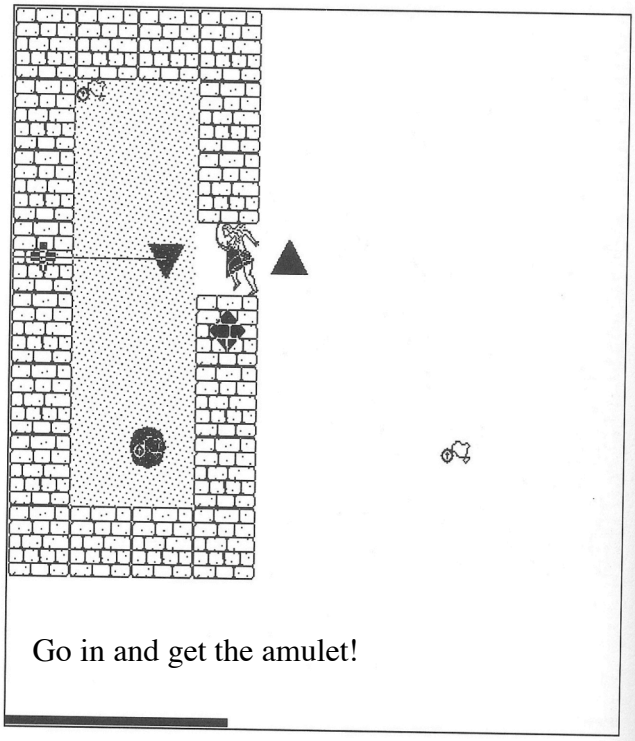


Figure 8.10
Sonja enters the room by passing the doorway markers in turn.

On the way to the door, the ghost also goes to the door and Amazon. The location of the ghost activates combat behaviors of Sonja. The left figure shows that the behavior of throwing a shuriken is activated.

If a user gives a command “use a knife” at the situation, the central system selects the behavior of killing the ghost with the knife immediately because the central system prepares potential behaviors for combat when Amazon encounters enemies.

After killing the ghost and the monster, Amazon goes in the room through the doorway like the right figure.

5.1.15 Combat

Sonja uses more complex graphical items when it combats enemies. The following figure shows Sonja finds a demon (marked with a pentagon) by following the direction

Find a monster based on visual search

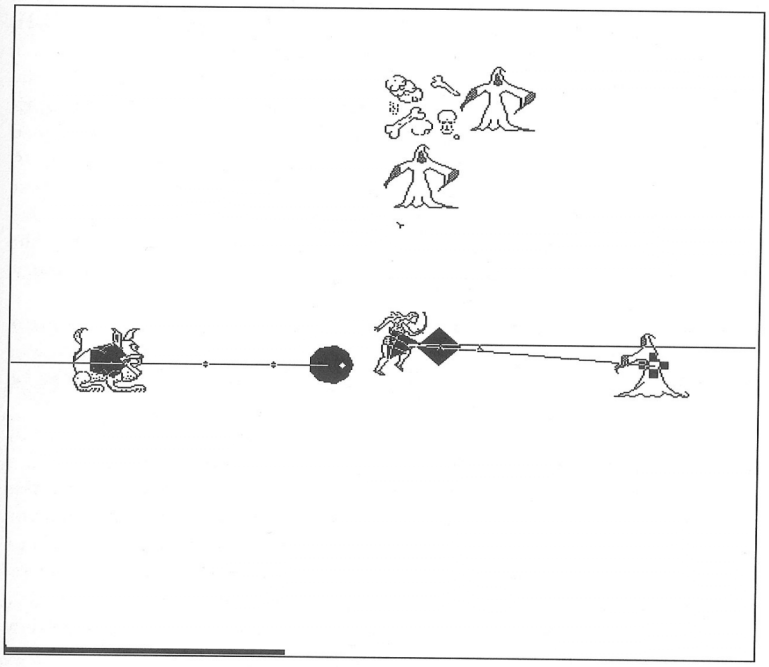


Figure 8.13
Registering a demon from its fireballs. Sonja has noticed the rightmost fireball and tracked it with the nonagonal marker. It has extended a ray from this marker in the direction opposite the fireball's motion and used that to find the demon. Having found the demon and tracked it with the pentagonal marker, Sonja will head for it in order to kill it.

of a fireball. At first, the visual architecture finds the fireball which is coming close to Amazon and marks it with a nonagonal. By extending a ray from the marker in the direction opposite the fireball's motion, it can find the demon. The central system selects behaviors of Amazon to kill the demon after finding it.

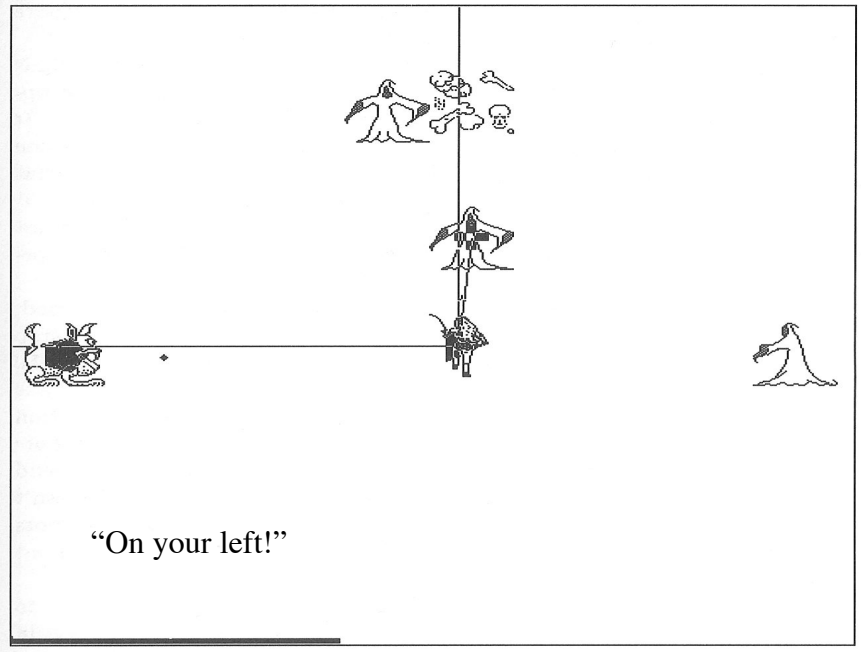


Figure 8.14
Registering a demon based on advice. The instruction "On your left!" has recently been given. Sonja has extended a ray in the specified direction and used it to find the demon. On this tick, Sonja has registered the demon; on the next tick it will turn the Amazon to face and shoot it.

The last figure in the previous page shows a case where Sonja finds a demon in response to an advice “on your left.” While a human gives the advice, the visual attention focuses on a ghost (marked with cross) in front of Amazon. However, the visual architecture turns the ray from Amazon in the direction to the left of Amazon after receiving the advice. Then, it gives a pentagon to the demon. The central system selects Amazon’s behavior which shoots the demon.

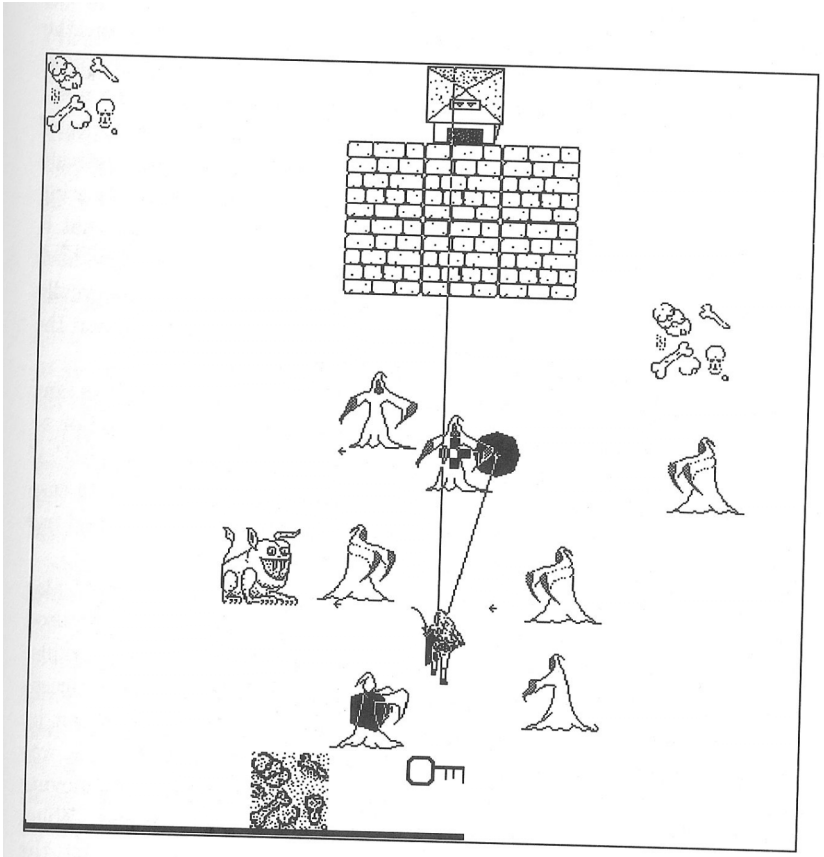


Figure 8.15
Opportunistic shooting. The nearest monster, the pentagon-marked ghost, is nonthreatening. Sonja has just noticed the knife marked with the octagon and turned toward it. The heading ray intersects another ghost, that marked with the cross, indicating that Sonja can shoot it opportunistically.

In the above figure, there are many enemies around Amazon. The visual architecture monitors the closest ghost marked with a pentagon among them. However, the ghost cannot attack Amazon soon. Instead of treating it, the central system tries to find the other relevant thing. The figure shows that Amazon has just found the knife marked with an octagon which another ghost holds. The knife activates the behavior of Sonja to reduce a threat from it. Then, the visual system turns the ray in the direction to the knife. While turning the ray, it finds the ghost which holds the knife and the central system sets the ghost as a target (marked with a cross).

5.1.16 Summary of Sonja

Sonja achieved a behavioral mechanism to deal with situated instructions. The interpretation of the instructions makes use of the effect of joint attention.

- Situated instruction
  - Effect on the selection of behaviors
  - Effect on visual attention
  - Low computational cost for the interpretation
- Joint Attention
  - Interaction on a video game
  - Sharing a game scene

5.2 Development of Infants' Joint Attention Skill

The section 5.2 explains how infants obtain the skill of joint attention during their developments. It is said that interaction between a mother and her infant is crucial for the development. Also, there are many studies of autism children to make clear the brain function related to joint attention because the autism children have communicative disabilities coming from the lack of the skill of joint attention. They do not have skills to read other's intention or states of the other's minds.

Joint attention is also the basis of social interaction. Since humans can establish joint attention, they can behave socially while being conscious of the other's attention. The fact causes a question whether or not joint attention is required for achieving social interaction between humans and robots.

5.2.1 Gaze Awareness

Researchers who studies language acquisition insist that the gaze awareness is a significant concept for understanding the process of the acquisition. Most important function in the gaze awareness is to draw the other's attention, which is called gaze drawing.

Gaze drawing indicates a cognitive phenomenon where a human draws the other's attention to something by using his/her direction or motion of gaze. It is sometime accompanied by pointing behaviors to assist the indication of the direction. Moreover, it becomes a trigger to start for two or more persons to share an experience.

We can say that joint attention heavily depends on gaze awareness.

There is an enigma about gaze drawing as follows.

- How does an infant understand mother's pointing behaviors of gaze drawing or her finger.

In other words, he/she must interpret the motions as the pointing behaviors. The study on the enigma makes clear the mechanism of joint attention.

5.2.2 Study on gaze interaction

Many researchers of developmental psychology have investigated what functions related to gaze infants acquire throughout their growth. The following list shows the overview of the functions at each stage of the growth.

- 6th month: Pay attention to an attracting feature of an object which is in his/her view
- 12th month: Response to geographical structure between mother's gaze and objects
- 18th month: Response to geographical structure based on a mental 3D map

6th-month-old infants direct their attention to an object only if the object is in field of their vision and is attractive to them. When they find an attractive object, they pay attention to it by following it with their eyes.

After 12th month, infants can find an object when a human directs their gaze to it. The gaze following is evidence that they can realize the relationship between human's gaze and the object. However, the object must be located between the infants and the human.

After 18th month, they can find the object even if it is behind them. They have a 3D mental map to recognize things. So the 18th-month-old infants turn their attention behind them when humans direct their attention behind the infants.

The developmental process indicates that infants obtain the skill of joint attention as a result of acquiring the skill of gaze following. Moreover, the process indicates that the gaze following is a combination of three cognitive skills: finding environmental feature, noticing geographical features (a relationship between pointing gestures or gaze and an object) and possessing a 3D mental map.

5.2.3 EDD and SAM

This subsection explains cognitive models for establishing joint attention. There are two models: Eye Direction Detector (EDD) and Shared Attention Mechanism (SAM). It is said that the brain has the same functions corresponding to those models.

EDD detects the direction of other's gaze. For example, humans can notice someone gazes at me. Also, SAM detects the establishment of joint attention. If you turn attention to the same thing with others or others turn their attention to the thing you focus on, SAM make you notice the establishment of joint attention.

It is also said that humans acquire the brain functions in the course of evolution to

maintain their social activities. In short, gaze is a cue to start social behavior and the joint attention is a basis to understand others focusing on the same thing.

5.2.4 EDD

EDD uses dyadic representation（二項表象）to express someone looks at someone/something. In short, it is a relation between two persons or a person and a thing. You can imagine the following variation of the dyadic representation.

- [agent -relation- self] , bidirectional
  - Ex. [mother -look-at- me], [I -look-at- mother]
- [agent -relation- proposition], one way
  - Ex. [mother -look-at- bus]
- [agent1 -relation- agent2], bidirectional
  - Ex. [mother -look-at- father], [father -look-at- mother]
- [self -relation- proposition], one way
  - Ex. [I -look-at- house]

The dyadic representations give a description of what someone looks at or where is an enemy and become the bases of expressions SAM uses in establishing joint attention. However, they are insufficient for establishing joint attention.

5.2.5 SAM

SAM uses triadic representation (三項表象) to identify whether or not the other pays attention to the same thing. It has a nested structure and expresses a relation between three items.

- [self -relation- (other -relation- proposition)]
  - bidirectional
  - [I -look-at- (mom -look-at- bus)]
  - [mom -look-at- (I -look-at- bus)]
- [self -relation- (other1 -relation- other2)]
  - bidirectional
  - [I -look-at- (mom -look-at- dad)]
  - [I -look-at- (dad -look-at- mom)]
  - [mom -look-at- (I -look-at- dad)]
  - [mom -look-at- (dad -look-at- me)], etc...

The first example shows a human pays attention to the same thing as another. That is that a person and his/her mother look at the bus. The relationships are bidirectional among “I” and “mon.”

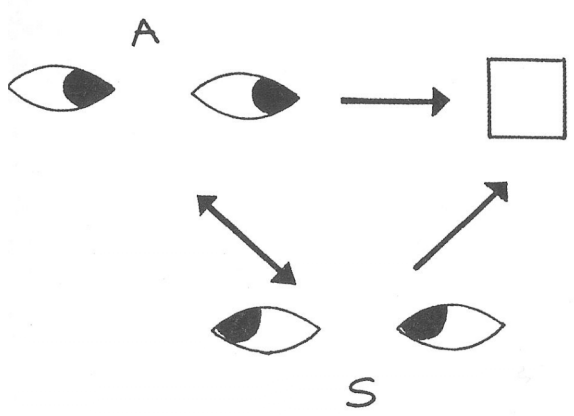
The second example shows that three persons "I," "dad," and "mon" are looking at each other. Also, the relationships are bidirectional among the three persons. However, the variations coming from the bidirectional relationships among them are more than that of the former example between "I" and "mon." There are six types of the variations.

Moreover, there is a more serious difficulty when you describe the triadic relations with those representations. The difficulty comes from the structure of the representations which allow nested structures to write down the relationship between three items. The nested structures mean that a relation has the other relation as one of its items. For example, you can write complex relationship as follows.

- Complex structure of triadic representation
  - [self -relation- (other -relation- (self -relation- proposition))
  - [other -relation- (self -relation- (other -relation- proposition))
  - [self -relation- (other -relation- (self -relation- (other -relation- .....)))

The first two examples show that the nested structures include two relations. Let's imagine that the relation corresponds to "noticing," self to "I," and other to "mon," proposition to "a cat" to make it easier for us to interpret the relations. The first example indicates "I notice that mom notices that I notice a cat." Also, the second expresses "mon notices that I notice that mon notices a cat." As the examples indicates, the expressions deal with other's state of mind.

The difficult point of the expression is that it allows much more complex expressions with the nested structure. The third example shows that. It indicates "I notice that mon notices that I notice that mon notices that…" In short, you can infinitely repeat the nested structure. It is not just a superficial phenomenon as a result of the describing method but suggests that the infinite nested structures are potentially included in a triadic relation.



We cannot describe the infinite nested structure with the expressions precisely. This is the reason why many researchers uses 2D graphical representation to expresses a

triadic relation. The figure in the previous page shows the 2D expression of a triadic relation between A, S, and an object. There are a bidirectional relation between A and S. Also, the relation between them and the object is one-way relation. The bidirectional relation expresses the infinite nested structure graphically.

Here, let us consider what psychological phenomena the infinite nested structure indicates. SAM, that is joint attention, is not only a phenomenon in which two or more humans pay attention to the same thing, but a phenomenon in which they notice where the others pay attention and whether the other notice where themselves pay attention. The reference of the other's state of mind corresponds to the infinite nested structure.

SAM establishes joint attention based on dyadic representation coming from EDD. The triadic representation is the result of the function of EDD and SAM as follows.

Gaze direction→EDD→Dyadic representation →SAM→Triadic representation

EDD generates dyadic representations from gaze directions and SAM generates the triadic representation from the dyadic representation. In short, gazing behaviors are crucial to establish joint attention.

5.2.6 SAM and ToMM

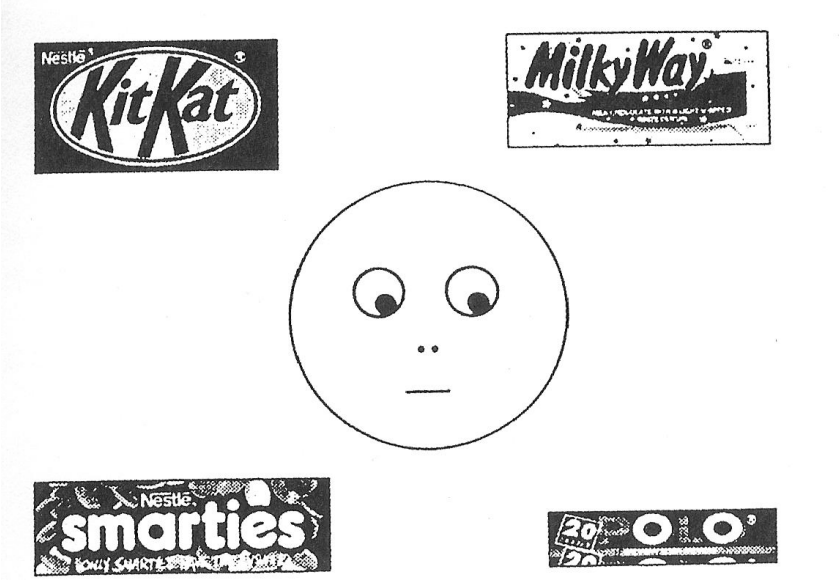
The structure of the triadic representation induces a human to have a viewpoint to consider other's intentions or goals. Since joint attention is established by observing gazing behaviors, we can say that humans infer other's intention or motivation to do something from other's gazing behaviors.

It is said that brains have Theory of Mind Mechanism (abbreviated to ToMM) for the inference of other's intention or motivation. SAM is a basis of ToMM by preparing the triadic representation. In short, ToMM also depends on the other's gazing behaviors because SAM does.

Developmental psychologists have investigated the behaviors of infants to verify the effect of gaze on the inferring other's intention. The experimental subjects were infants from 9^{th}-18^{th} months old. The infants watched adult's eyes when confirming his/her intention. In other words, infants made eye contact at the situation.

The experiment was designed to observe whether the infants make eye contact when an adult behave intentionally. In the experiment, he/she hided a thing intentionally when the infants hand reaches it. Also, the experiment prepared a control condition where he/she did not hide it at the same situation.

The results indicate that the infants made eye contact when he/she hide it intentionally. Otherwise, the infants did not. We can say that watching the other's eyes corresponds to a cue to infer other's intention.



Three or four years old kids can answer a question about Charlie who is a man in the above figure. They refer to the gaze direction of Charlie in inferring Charlie's intention. They can answer the following questions.

- What sweet does Charlie get?
- What does Charlie want?
- What does Charlie mention?

They indicated the lower right sweet in response to all questions. The result indicates that the kids selected the sweet by referring to the gaze direction of Charlie. Moreover, we can say that they inferred Charlie's intention from the gaze direction because these questions are related to intention of Charlie.

The previous discussions and the experiment highlight the relations between EDD, SAM, and ToMM as follows.

- EDD → SAM → ToMM

Many evidences of the relation have come form studies on autism.

5.2.7 Joint attention and autism

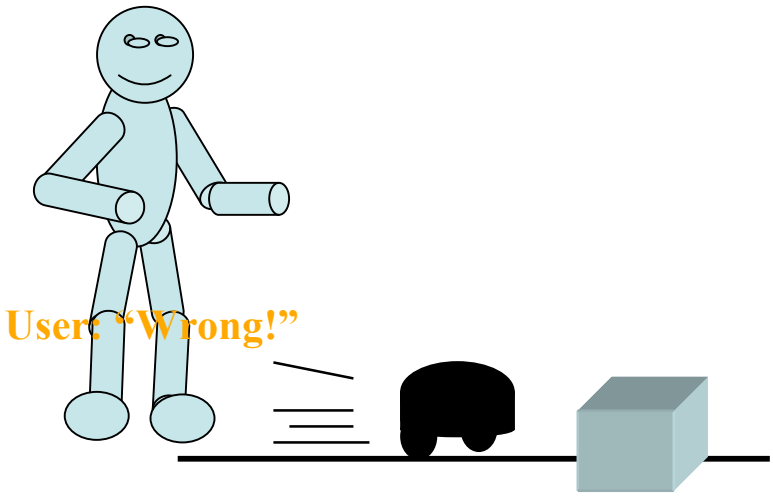
Autism children cannot infer the other's intention. Many researchers take methods to observe them to find what brain mechanism is crucial for ToMM because they seem not to have the mechanism.

Autism children have EDD but do not have SAM. They can recognize that someone is looking at them. But they cannot behave based on joint attention. It is also said that they can handle dyadic representation but cannot deal with triadic representation.

Moreover, autism children do not make eye contact. They are not good at finding out an item where someone directs his/her gaze. The evidence indicates the lack of SAM

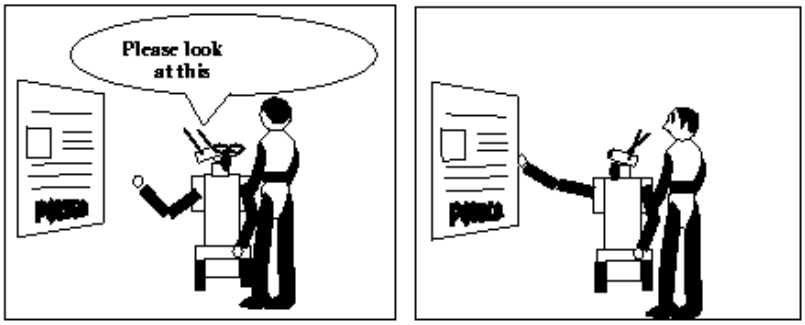
because they can identify the direction of other's gaze.

5.2.8 Joint Attention with Robot



Let us consider joint attention in human-robot interaction. This figure shows the interaction between a user and Linta-I. I think that you remember the method of the interpretation of a situated command by Linta-I. It interprets the command like "wrong" by referring to sensor information. However, there is no guarantee that the interpretation is correct. Joint attention must be established between the user and Linta-I for the interpretation.

However, it seems difficult for Linta-I to establish joint attention because it does not have eyes and arms to suggest the target of attention. We must consider what manifestation methods the robot requires to establish joint attention.



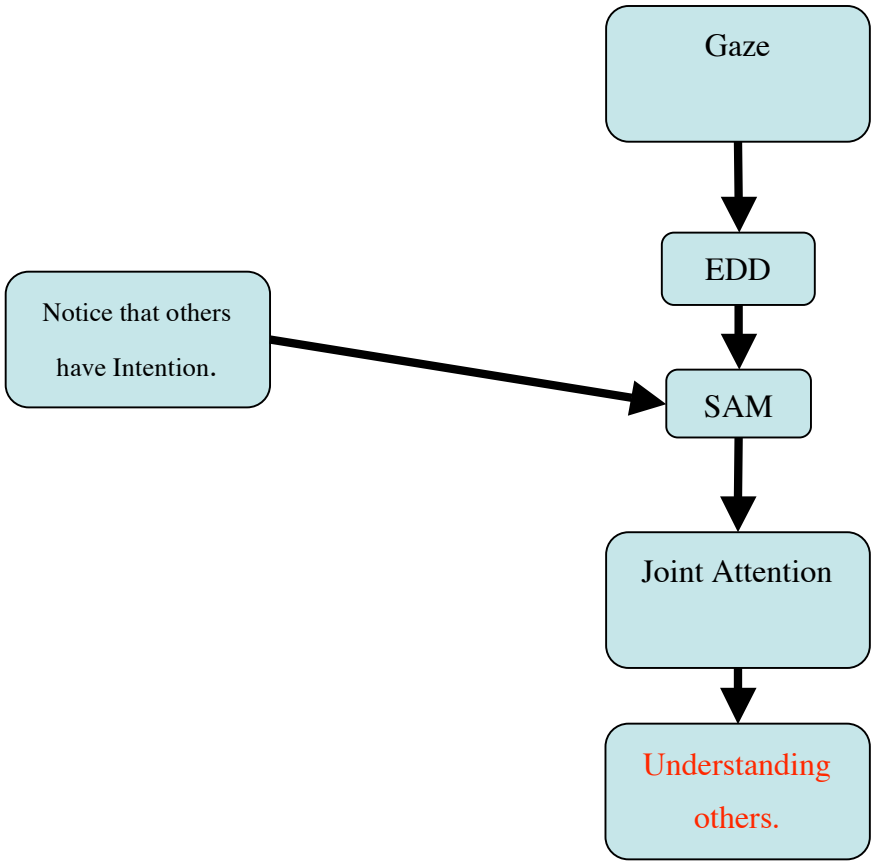
Easiest way to establish joint attention is to give the robot the same skill of physical expression as humans. For example, a humanoid robot can establish joint attention with human by making eye contact and pointing a target by its hand. The behaviors manifest the target of the robot's attention.

5.3 Theory of Mind Model and Infants' Joint Attention Skill

This section describes how ToMM works and when infants acquire the skills for ToMM.

Understanding others as intentional agents is the basis of initial stage of SAM development. If humans cannot understand other's intention, they cannot establish joint attention. Moreover, they cannot establish common understanding of situations in the real world.

5.3.1 Intention, joint attention, and understanding others



Understanding others as intentional agents means that others may pay attention to something intentionally and ignore the others. Since we know the other has such an intention, we interpret the other's behaviors such as gazing behavior in terms of his/her internal intention. The interpretations lead to understanding of others.

For example, let us consider the following two situations.

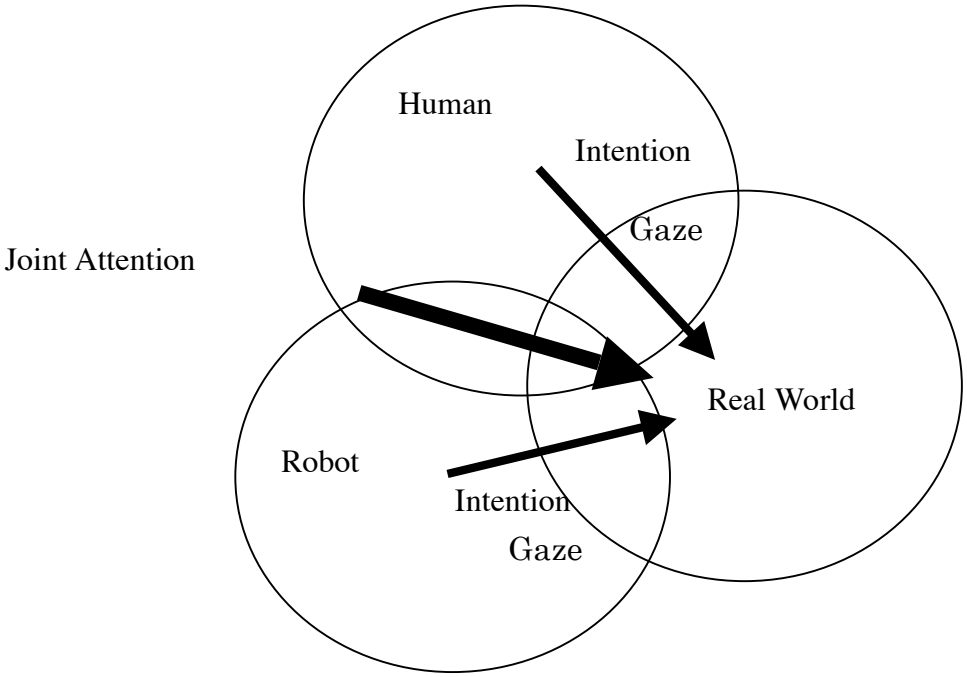
- A climber watches a mountain.
- A hiker watches a mountain.

The examples show almost same situation that a human watches a mountain. However, their intentions of watching it are different. The climber intends to confirm his mind of

conquering the mountain. On the other hand, the hiker intends to imagine expected joyful scenes at the mountain. Although others cannot observe the intentions directly from their behaviors, our brain functions make it possible.

The figure in the previous page shows relations between brain functions for understanding others. While the sequence from GAZE to SAM is the same as previous section, SAM requires understanding others as intentional agents when establish joint attention. In short, joint attention is a result of noticing other has intention and recognizing other's gaze direction.

The requirement of the ability of understanding intentional agents comes from the study on autism children and the development of infants. The autism children can recognize the direction of other's gaze but cannot understand that the other has intention. Also, normal infants acquire the skill of EDD at first. However, they cannot establish joint attention at this developmental stage. After acquiring the skill of understanding other's intention, they can establish joint attention.



Here, let us consider the same situation when a robot establishes joint attention with a human. When designing the robot, you must deal with intentions and gazes of the human and the robot. The robot must understand his/her gaze direction and infer his/her intention why he/she focuses on the event.

Moreover, the robot must express a target which it focuses on by using devices like eyes or alternatives to the devices. The intention of the robot is also significant for joint attention. The robot is not only designed to manifest its intention but also must make humans engage in inferring its intention because they do not always deal with the robot as an intentional agent.

The above requirements must be satisfied to establish joint attention between humans and a robot.

5.3.2 Infant's Joint Attention

As I explained at the previous subsection, the ability of understanding other's intention is crucial to establish joint attention. In spite of kids, they have the ability as follows if they are mature for establishing joint attention.

- Kids understand others in terms of intention.
- Kids understand that the others have intentions and they may differ from kids' intention.
- Kids understand that the others have intentions and they may differ from actual facts.

The remaining part of the subsection explains how infants acquire the ability for joint attention. I focus on 9th month, 18th month, 24th month old after their birth to explain the ability in terms of their development.

9th month old infants

Let us start explanations from 9th month old infants. There is a possibility that infants at the age are not aware that an adult directs attention to them. They look at the same thing that an adult watches simultaneously by chance. The shared attention is not joint attention because they do not infer the adult's intention to turn their attention intentionally.

After 9th month old, they obtain the skills for joint attention according to the following steps.

1. The infants look at a situation where the adult looks at the thing.
2. Gaze alternation: Infants direct their attention toward the two interesting things (the adult and the thing) by turns.

18th month old infants

18th month old infants have a skill to establish joint attention. They can learn something by mimicking adult's behaviors. Also, they take intentional communication with others. The infant's behaviors are the results of joint attention. Moreover, they understand the others as intentional existence.

The evidence of the skill comes from the fact that they direct their attention to an adult and a target by turns spontaneously without adult's prompt. For example, an infant follows an adult's gaze direction spontaneously, identifies the target, and

returns to the adult immediately to confirm that the adult's attention still remains there.

There are also experiments confirming the skill. I introduce an experiment as the evidence of infant's skill. That is behaviors based on joint attention.

The experiment observed infants' behaviors when an adult bent and pushed the power button on a wall with his/her head. When an experimenter showed the situation to infants, they mimicked the adult behavior. In short, they pushed the wall with his/her head. On the other hand, if the adult just pushed wall with his/her head, they did not mimic it.

The experiment indicates that the infants inferred the adult's intention to turn on the light with head. Then, they mimicked the behavior with respect to his/her intention. In other word, they inferred an adult's intention based on joint attention.

There are the other behaviors related to intention. For example, when an infant tells an adult his/her intention, the infant carries out spontaneous gaze alternation expecting adult's intentional responses while generating pointing behaviors. More significant fact is that an infant gazes at the adult's eyes instead of the adult's hand even though the hand achieves the desired response.

The above evidences indicate that infants at 12th--18th month old behave by recognizing an adult's attention and intention, and by understanding a difference between their intentions.

	9th month	12th month
Following attention Following gaze Sharing interaction	 Directed gaze following Passive sharing	 Spontaneous gaze following Confirming sharing
Following action Learning by mimicking	 Just mimicking	 Mimicking based on other's intention
Directs other's attention		Pointing behavior and gaze alternation

Let us compare the skill of 9th month old infants with that of 12th month's. The above table shows the difference between them. The second row shows comparison in terms of following other's attention. There are two items for the comparison: following gaze and sharing interaction. 9th month infants follow the other's gaze when the other prepares some situation which forces them to follow his/her gaze. On the other hand, 12th month old infants follow the other's gaze spontaneously.

Also, 9th month old infants already interact with others by sharing a situation. However, the sharing is established in a passive manner. On the other hand, 12th month old binfants share the interaction by confirming the sharing intentionally. For example, they frequently make eye contact with others in the interation.

Third row shows the comparison in terms of following action. Following action means mimicking behaviors. Although 9th month old infants just mimic the other's behaviors, 12th month old infants refer to the other's intention before starting mimicking them.

Forth row shows the skill of directing other's attention. 9th month old infants cannot draws other's attention intentionally. On the other hand, 12th month old infants can by using a pointing behavior and gaze alternation.

24th month old infants

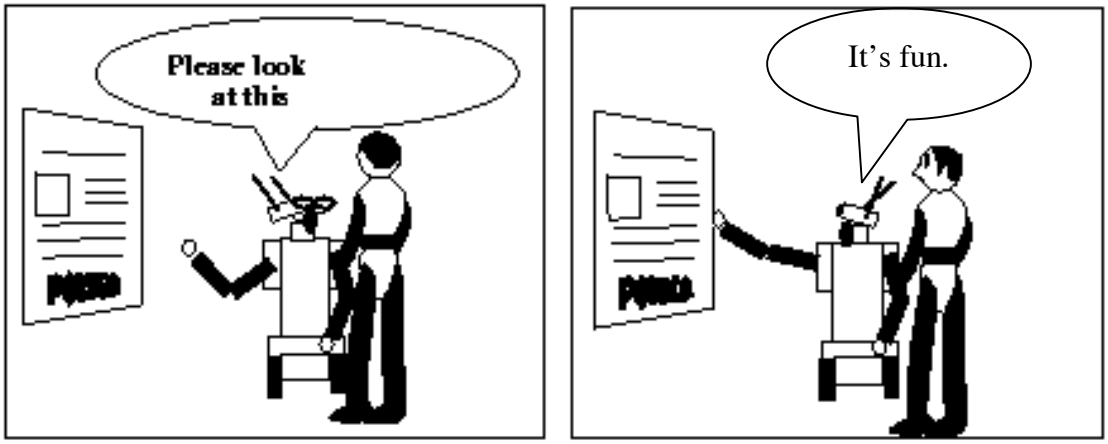
24th month old infants can deal with more complex interaction by using verbal expressions and joint attention. People establish joint attention when generating a verbal directive (言語的指示). "I intend that you intend to interact with X." shows the example of a verbal directive. The intention is directed to you, not to X. Dealing with the verbal directive requires joint attention.

24th month old infants already understand others as intentional agents. Based on the understanding, they learn a new word by finding an attention target. Moreover, they can draw the other's attention by using a language.

Moreover, 24th month infants notice that others' intentions are different from them. The knowledge is also used in learning a new word based on the other's attention, even though the infants direct their attention to a different thing.

Language acquisition of infants

Topic(話題)　　Predicative(叙述)



The effect of joint attention appears on infant's language acquisition. In particular, it

has an effect when infants give others a topic or a predicative. The figure in the previous page shows the example of a topic and a predicative. Giving a topic is to introduce a subject of a conversation. In the figure, the robot introduces a poster as a topic by a pointing gesture and an utterance “please look at this.” Giving a predictive is to give information about the subject. In the figure, the robot informs the human that the poster is fun by using the utterance.

Almost utterances which adults speak consist of a set of a topic and a predicative. However, the utterances of infants do not have the structure of the set. The next table shows how infants acquire the structure of an utterance throughout the stage of development.

Month	Communication skill	Topic	Predicative
0-6th	Primary intersubjectivity 第一次 間主観性		
6-12th	Secondary intersubjectivity 第二次 間主観性	Nonverbal	
12-18th	First word	Verbal	
18-24th	First sentence	Nonverbal	Verbal
24th-	Conversation	Verbal	Verbal

The table divides the developmental stage into five stages: 0-6th, 6-12th, 12-18th, 18-24th, and 24th month old. The communication skill indicates what skill infants have at each stage. The topic and the predicative in the table indicate what expressions infants use at each stage.

The communication skill of 0-6th month old is called primary intersubjectivity. Intersubjectivity corresponds to a psychological state that someone believes that the other actually has an ability of recognition. We cannot soon conclude that the belief is true when considering it deliberately. In the consideration, we can imagine that the other does not have the ability of recognition. On the other hand, we assume the belief in our daily life. It can be said that admitting intersubjectivity is the basis for a communication between humans. However, it cannot be said that infants have intersubjectivity from their birth. Infants gradually obtain it.

0-6th month old infants have primary intersubjectivity. At the stage, infants express their emotion directly. The expressions do not include a topic and a predicative.

6-12th month old infants have secondary intersbujectivity. At the stage, infants introduce a topic with nonverbal expressions. In other words, they tries to draw the

others attention to the topic.

12-18th month old infants can speak words and use it to introduce a topic. However, they do not give a predicative to the topic.

18-24th month old infants introduce a topic with nonverbal expressions while giving a predicative to the topic with verbal expressions. For example, they say “wet” pointing at their pants.

24th month old infants can use verbal expressions which are the set of a topic and a predicative. For example, they can say “my pants is wet.”

The developmental stages of acquiring language skills also indicate the development of the skills of joint attention. 12-18th month old infants have the skill of establishing joint attention. Moreover, 18-24th month infants can control the target of joint attention. Let us consider the example of “wet” again. They establish joint attention toward their pants with a nonverbal expression. At the same time, they make the other focus on the part of wet in their pants. The predicative “wet” controls the other’s attention after the nonverbal expression establishes joint attention.

5.4 Human-robot interaction and theory of mind model.

The previous sections have explained what structure the human’s joint attention and theory of mind model has. Also, they have explained what effects they have in a communication.

This section describes an example of human-robot interaction based on theory of mind model.

5.4.1 The Effect of ToMM

As an example of ToMM in human-robot interaction, this section shows a psychological experiment. In the experiment, a robot requests a human to do a task.

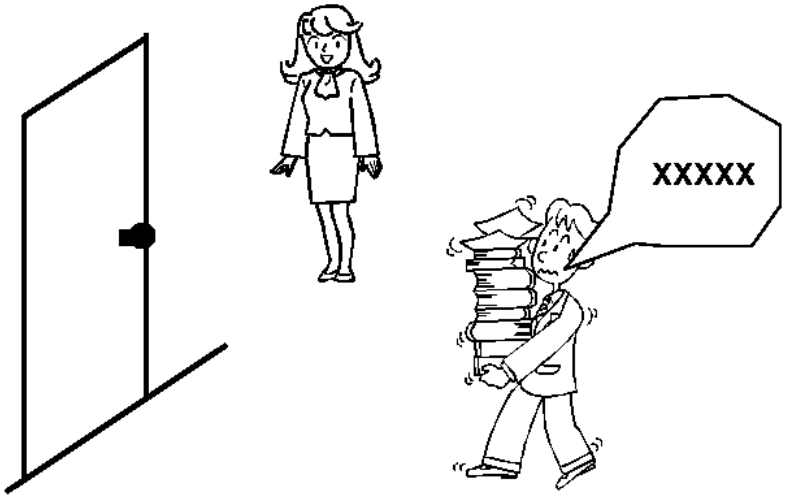
When the robot gives him/her the request, the following psychological phenomena must work well in the interaction. In other word, we must design the interaction between the human and the robot to deal with the phenomena.

- Joint Attention
- Mind reading
- The development of relationship

Here, there are new words: mind-reading and the development of relationship. Mind reading shows that a human infers the intention of the other’s behaviors or the state of the other’s perception. The relationship is important factor to communicate with someone. A robot must also establish the relationship with humans when

communicating with them.

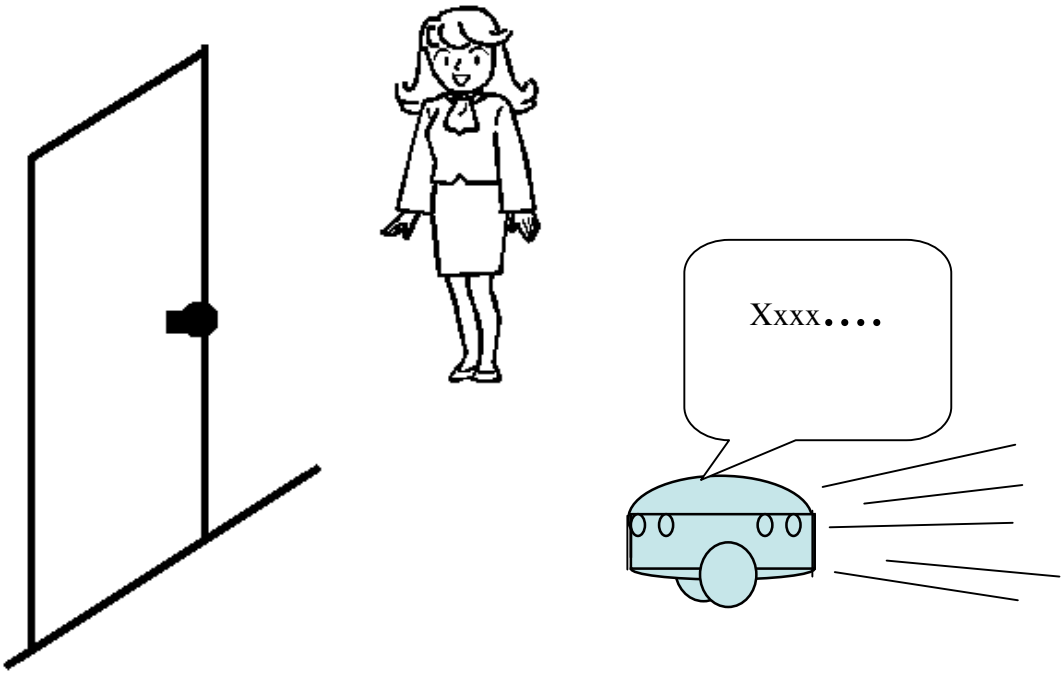
The experiment reveals the effect of the psychological phenomena in terms of ToMM.



Let us consider a situation shown in the above figure. This example includes effects of the three psychological phenomena. In the figure, the man tries to go through the door. However, he cannot open the door because his hands are full of books. The woman is his friend. If she looks at the situation, she will open the door even though he does not say anything. In actual, he just grunted in the figure.

At the situation, she must infer his intention that is to go through the door while establishing joint attention to his behavior. In short, she reads his mind and establishes joint attention. The relationship as a friend plays an important role in the situation. She actively reads his mind because he is her friend. Without the relationship as a friend, she might not behave like this. As the example shows, joint attention, mind-reading, and relationship are important in communications. Their effects are general in human's communication. You always read the other's communicative intention in a conversation. You seldom read intention of an unfamiliar person.

The figure in the next page shows a situation in which the robot tries to go through the door. We can consider the situation by analogy with the former example. What should the robot do to make her open the door? Is it sufficient for it to generate a request "please open the door?" The answer of the experiment in this section is no. The robot must develop a relationship with her and make her read the intention of the robot. Only verbal request does not sometimes work in human-robot interaction.



5.4.2 Hypothesis

At first, I will introduce a hypothesis for a psychological experiment.

Hypothesis: *By reading a robot's mind, a human can estimate the robot's intention with ease, and, moreover, the person can even understand the robot's unclear utterances made by synthesized speech sounds.*

This hypothesis means that the robot makes a human infer the intention of the robot when telling him/her something.

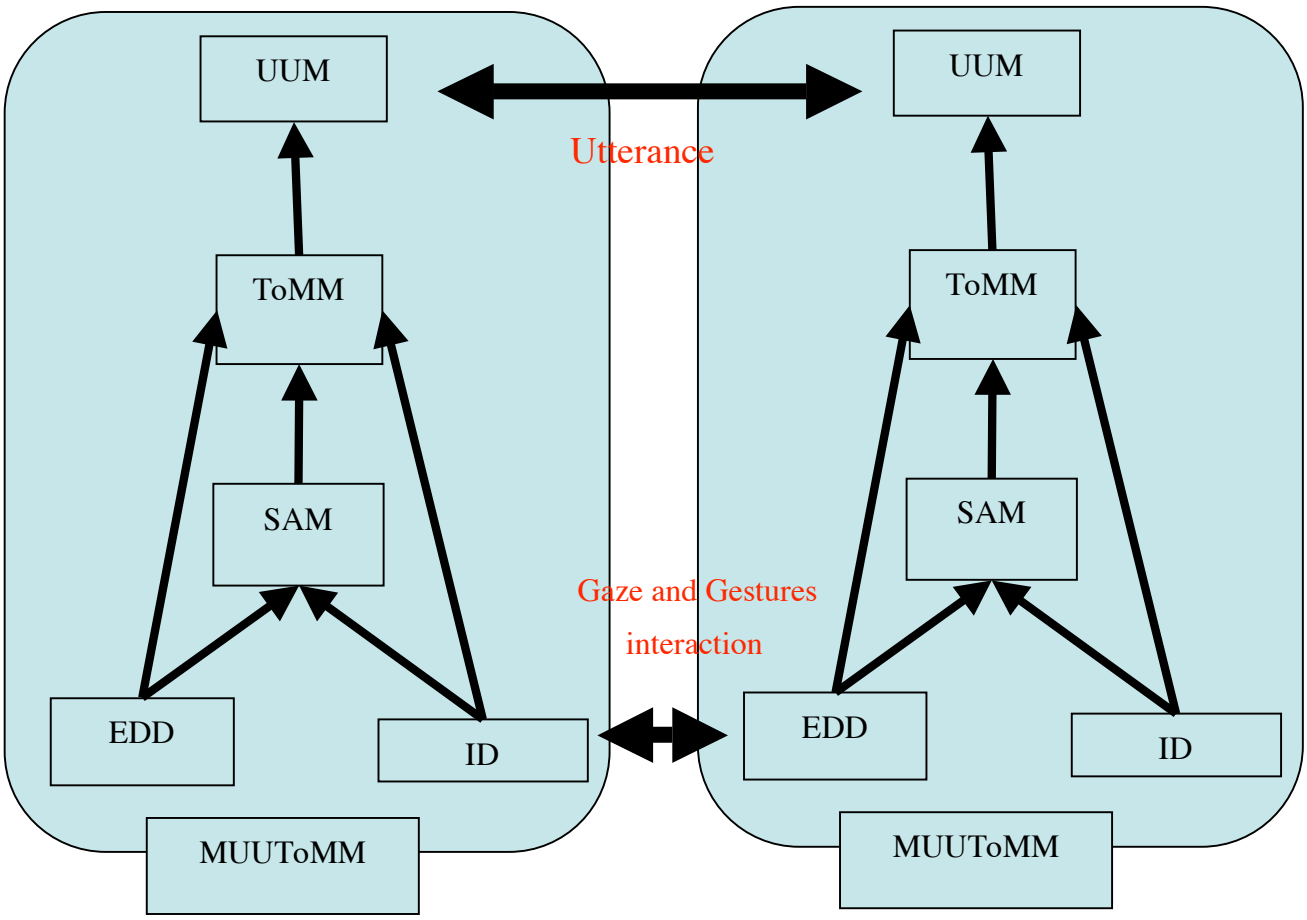
5.4.3 Model of Utterance Understanding based on ToMM

Before starting the explanation of the experiment, I introduce the model of utterance understanding based on ToMM which is abbreviated to MUUToMM. The figure in the next page shows MUUToMMs of two persons.

MUUToMM has ToMM which consists of EDD, ID, and SAM as I already explained. Although I have not explained ID explicitly, ID is the abbreviation of Intention Detector which detects other's intentions. Let us remember that SAM works based on EDD and the intention of others.

MUUToMM indicates that UUM (a utterance understanding model) works on the result of ToMM.

Let us consider interaction between two persons in terms of MUUToMMs. There are two types of interaction in the model: an utterance and the set of a gaze and a gesture. EDD monitors the other's gaze direction and pointing gestures. Also, ID detects the existence of the intention in the gaze and the gestures. ToMM infers the other's



intention and perception while establishing joint attention with SAM. Since joint attention is established between them, they can assume that the inferred intention and perception are shared between them.

UMM interprets an utterance based on the shared intention and perception. Also, MUUToMM indicates that UMM fails in the interpretation if EDD or ID does not work correctly. In short, the interaction is not established with only utterances.

Let us apply this model to the example of a man bringing books. Since she is his friend, her EDD and ID start to monitor his gaze and his behaviors. Then, SAM establishes joint attention to his behavior bringing many books and ToMM detects his intention to go through the door. In the example, he says an unclear utterance “XXX.” UUM interprets the utterance based on the intention. At last, she notices that he want her to open the door.

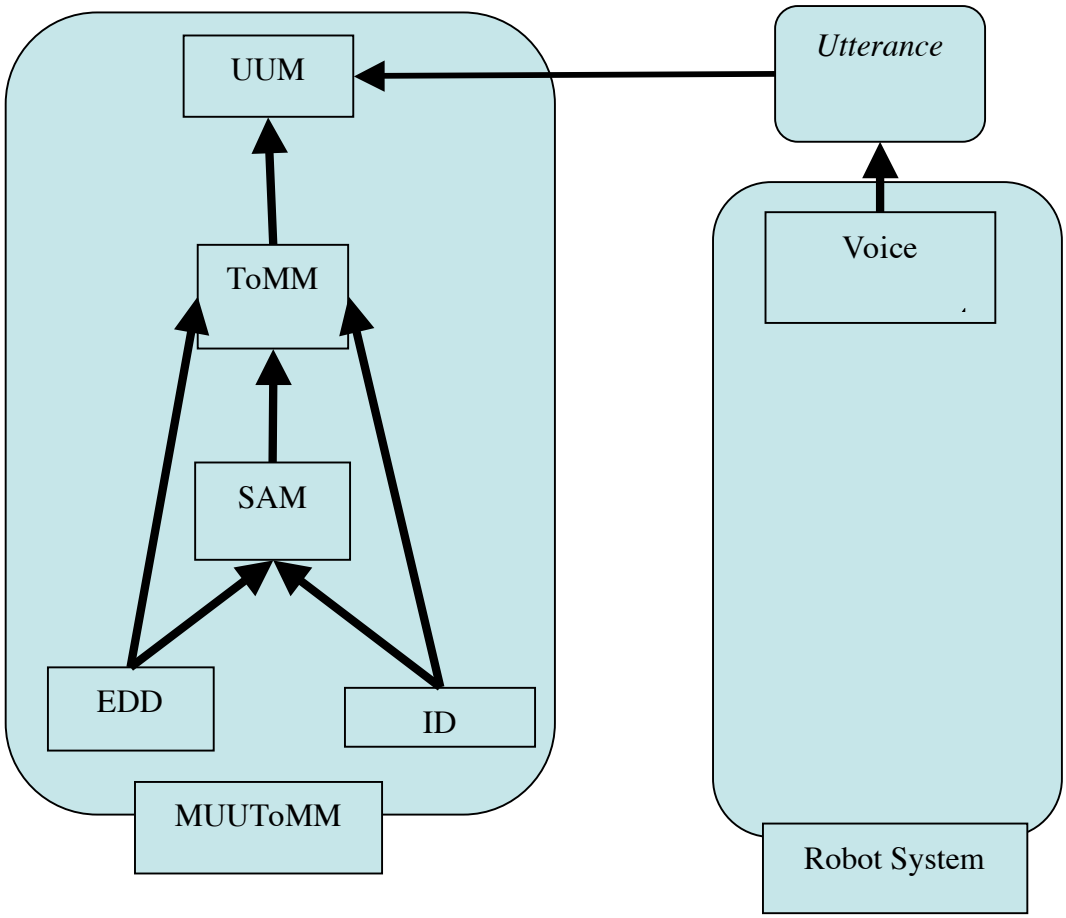
5.4.4 Robot System

Let us consider a robot system to achieve a communication based on MUUToMM. The difficulty of achieving the communication comes from the lack of robot’s appearance enhancing ID, EDD and SAM of humans. Without the appearance, humans do not regard robots as autonomous beings with intention.

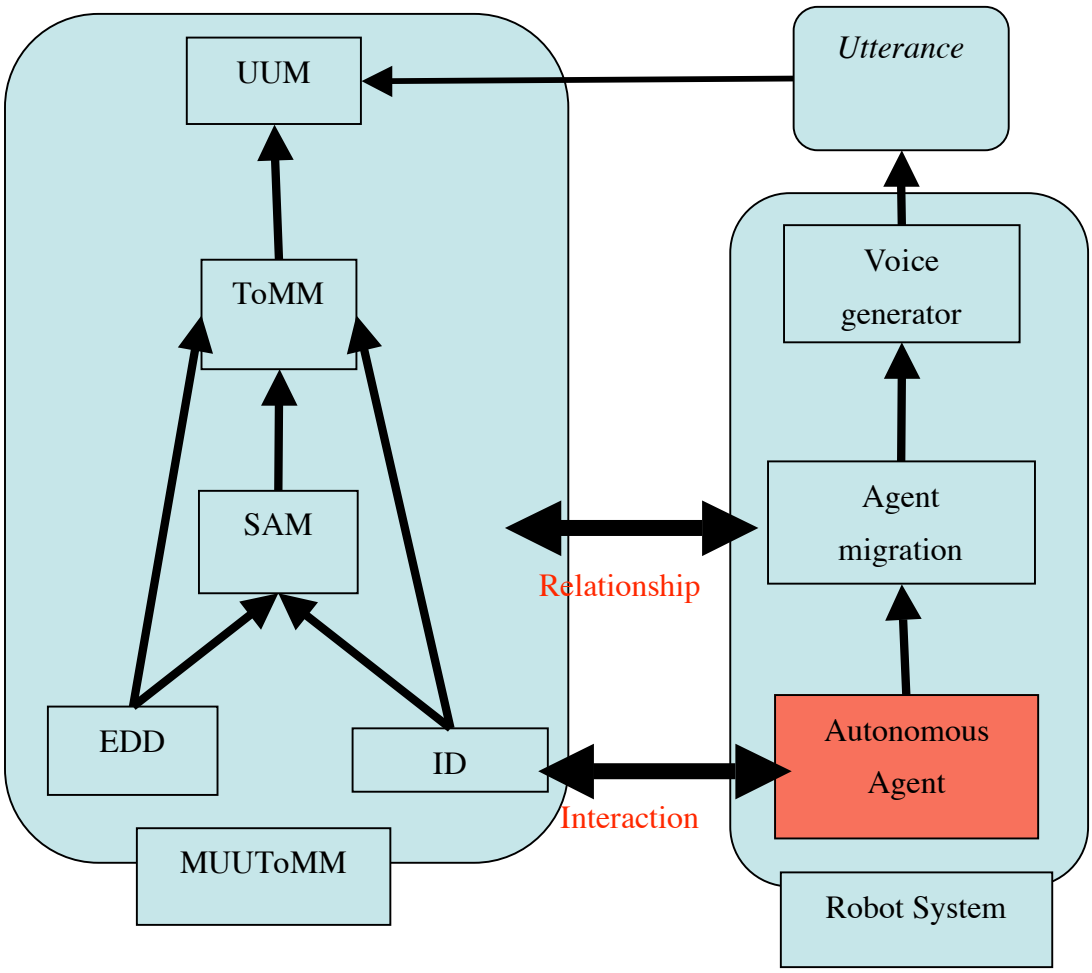
This section introduces an agent mediated communication interface which compensates for the lack of the appearance. The interface employs a CG character which I call an agent. The agent appears on mobile PC and behaves as if it is an autonomous being. A user can interacts with it by giving foods or stroking it. The interaction causes the development of a relationship between the user and the agent.

The agent mediates the relationship between the user and a robot. The robot of the system also has a display. The agent migrates from the display of the mobile PC to that of the robot. The agent migration gives the robot relationship with the human because a relationship is already established between the user and the agent

5.4.5 Understanding a Robot's Utterances



Let us consider the agent mediated communication interface in terms of MUUToMM. The figure shows the model. The left figure shows MUUToMM of a human. The right figure indicates the model of the robot. At first, the robot does not establish a relationship with the human. At the situation, EDD, ID, SAM, and ToMM of the human are not activated. In short, he/she does not infer the intention of the robot.



Without knowing robot's intention, he/she cannot understand an utterance from the robot.

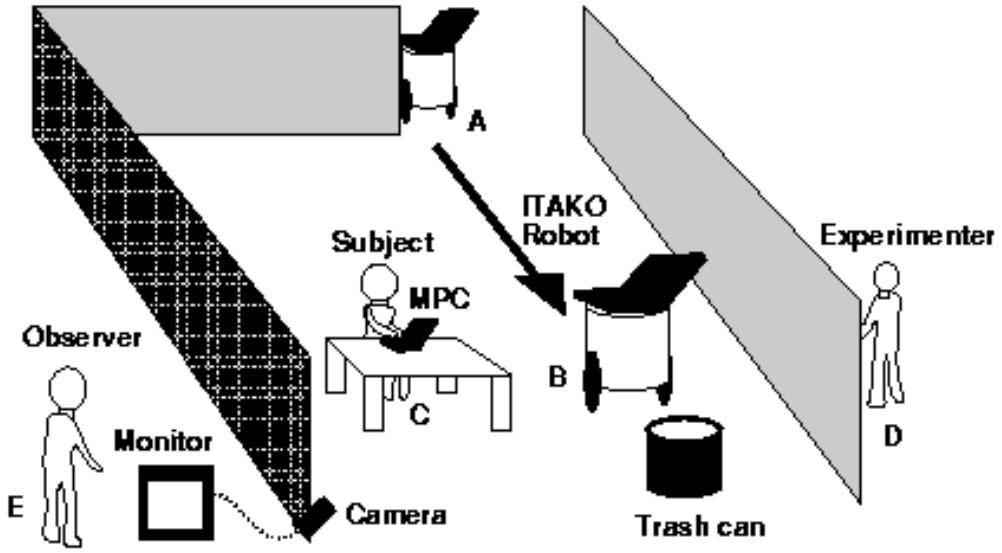
The figure in this page indicates the model of the humans and the robot after the agent migrates to the robot. The interaction on mobile PC establishes a relationship between the user and the agent. The relationship also migrates to the robot when the agent migrates. The interactive relationship activates EDD and ID of humans. As a result, the human infers the intention of the robot with ToMM. At last, the human can understand the utterance from the robot based on the inferred intention.

Here, the robot does not have eyes to activate EDD of the human. However, you will notice that the front direction of the robot plays a role as eyes because he/she refers to the front direction of the robot as the target of the robot's attention.

5.4.6 Experiment

Let us start explaining a psychological experiment which verifies MUUToMM in human-robot interaction. The experiment was held at a room like the figure in the next page.

The scenario of the experiment is as follow.



1) An experimenter asks subjects to sit at C in the figure and to interact with an agent shown on the mobile PC (MPC in the figure). The task of the subjects is to evaluate the interaction with the agent. Moreover, the experimenter did not tell them the existence of a robot.
2) After giving them the instruction, the experimenter went to place D from the room.
3) While the subjects interacted with the agent, the robot appeared suddenly and moved to a trashcan.
4) The robot asked them to move the trashcan at place B.
5) An observer observed the behaviors of the subject at place E to find what they responded to the requests from the robot.

There are two experimental conditions when the robot gave them the requests at scenario 4).

1. The agent migrates on the robot when the robot generates the requests.
2. The agent does not migrate.

The observer observed the difference of subject's behaviors between the conditions.

There are twenty subjects in the experiment. They were divided into two groups equally at random. One group was given the first condition and the other was the second one.

5.4.7 Hypothesis and Prediction

The experiment confirms the following hypothesis. We brought about the hypothesis corresponding to the model proposed at 5.4.5.

Hypothesis: *By reading a robot's mind, a human can estimate the robot's intention with ease, and, moreover, the person can even understand the robot's unclear utterances made by synthesized speech sounds.*

We can anticipate the following predictions in the experimental scenario in terms of the hypothesis.

1. The subjects will regard the robot as an autonomous entity with intention, which is the result of the functions ID and EDD.
2. The subjects will first look at the robot and then turn their eyes to the trashcan, which is the result of SAM.
3. The subjects will be able to estimate the robot's intention with ease, and this will facilitate their understanding of the robot's utterance , which is the result from ToMM and UUM.

5.5.8 Experimental Outcome

	With agent migration	Without agent migration
Utterance Understanding	8/10 persons	3/10 persons
Moving the trashcan	8/10 persons	1/10 persons

The table shows the result of the experiment. It indicates that eight subjects among ten could understand the request while the agent migrates. On the other hand, only three subjects could. Moreover, all subjects who could understand it moved the trashcan while the agent migrated. On the other hand, only one subjects moved without the agent migration.

From the results, we can conclude the following facts.

1. The subjects regarded the robot as an autonomous entity with intention (ID, EDD).
2. The subjects first looked at the robot and then turned their eyes to the trash can (SAM).
3. The subjects estimated the robot's intention with ease, and this facilitated their understanding of the robot's utterance (ToMM, UUM).

5.5.9 The Effect of Eye-contact

The experimental results come from the effect of mind reading. The agent migration induces the mind reading in the experimental setting. The primary functions for the mind reading are EDD, SAM, and ToMM. In particular, the relations between them are significant.

Although the agent migration induces ToMM in the experiment, eye contact and behaviors related to gaze are most basic behaviors for activating ToMM in human-human communications. There is a question what factor induces joint attention

and mind-reading phenomena for humanoid-robot.

5.5 Joint Attention Mechanism

This section explains a joint attention mechanism to establish joint attention between a human and a robot. It uses eye contact and pointing behaviors. Also, I explain a generation mechanism of robot's utterances which reflects the degree of establishment joint attention. The following items are the topics of this section.

- The effect of eye-contact
- Pointing behavior
- The degree of embodied expressions and language expressions
- The degree of development of joint attention and language expressions

5.5.1 The effect of eye-contact

As section 5.4 explained, joint attention is the result of functions EDD, SAM, and ToMM. In particular, eye contact and gaze behaviors are significant for activating EDD, SAM, and ToMM. That is, the gaze information is detected by EDD and SAM establishes joint attention by recognizing the other as an intentional existence.

The first part of this section introduces a joint attention mechanism using gaze behaviors of a robot.

5.5.2Gaze drawing

I have developed a communication system named Linta-III by using joint attention mechanism. The joint attention mechanism employs gaze behaviors of a robot to draw human's attention to somewhere.

Also, I have conducted a psychological experiment to confirm the effect of eye-contact on human-robot interaction. Strictly speaking, I observed whether the interpretation of utterances changes according to the gaze of the robot when humans understand the utterance of the robot. In particular, I selected utterances referring to physical world objects as targets of the experiment because joint attention have an effect on the reference.

5.5.3Utterance Referring to Physical World Object

This subsection shows notations which Linta-III uses to deal with utterances. The following notations expresses an example utterance "please remove this." KORE WO DOKETE KUDASAI in Japanese. The example includes a demonstrative pronoun "this" (KORE).

ask(R,P,move(P,O)),

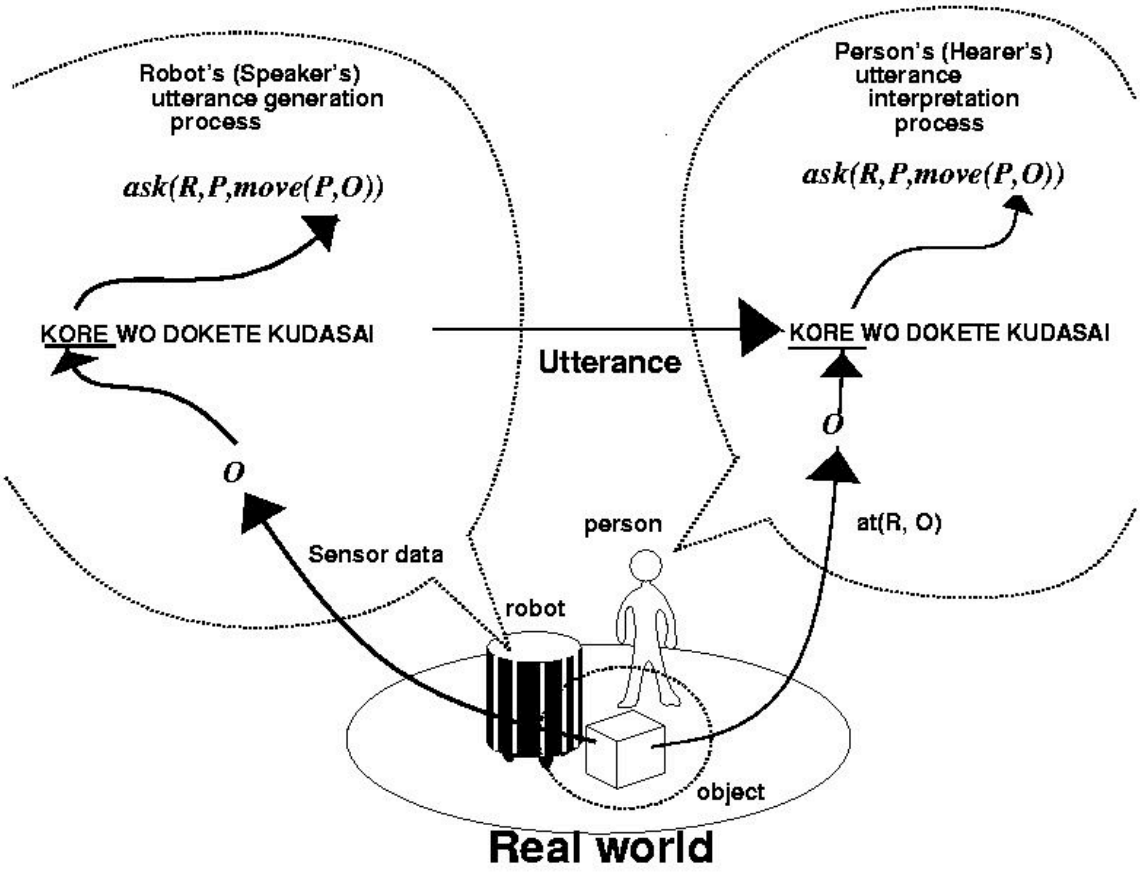
f(robot)=R

f(this)=O,

f(hearer)=P.

(P is omitted in the imperative form(命令形).)

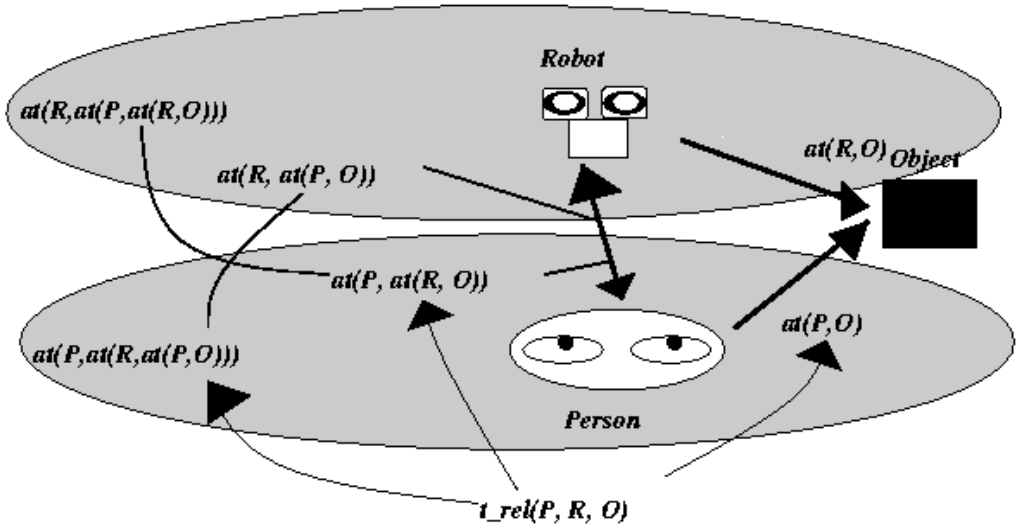
ask(R,P,move(P,O)) expresses the meaning structure of "please move this." *R, P,* and *O* are real thing existing in the real world. The word "this" is bound with the real thing *O* by the binding function *f*. Also, the hearer and a speaker (the robot) are bound by *f*. I explain the interpretation of utterances based on the expressions.



The figure shows the process of interpreting the utterance "please move this." The human is a hearer and the robot is a speaker. The robot finds the object *O* with sensors and intends to ask someone to move *O*. To generate a request to the human, it selects a word referring to *O*. As a result, KORE WO DOKETE KUDASAI (please move this) is generated.

The human must pay attention to the object *O* to interpret the utterance correctly. The attention corresponds to joint attention between the human and the robot. By connecting the word KORE (this) with *O*, he/she obtains *ask(R,P,move(P,O))* from the utterance.

5.5.4Triadic Relation



This subsection discusses triadic relation emerging on the experiment. The above figure shows the triadic relation between a human, a robot, and an object. There are several attentions in the situation. *at(P,O)* and *at(R,O)* correspond to human's attention toward the object and robot's attention to it respectively.

Moreover, *at(R,at(P,O))* and *at(P,at(R,O))* show that the robot notices that the person pays attention to the object and that the human also notices that the robot pays attention to it. There are more complex expressions in the triadic relation. Those are *at(R,at(P,at(R,O)))* and *at(P,at(R,at(P,O))).* They express that the robot notices that the person notices that the robot pays attention to the object and that the human notices that the robot notices that the human pays attention to it.

When joint attention is established, at least the three types of attention must be established. That is because joint attention is a phenomenon that the two persons not only pay attention to the same thing but also are aware of the other's attention.

However, the figure is still different from actual joint attention. The figure takes an ideal viewpoint to express joint attention. That is, the viewpoint can observe both states of the human and the robot while we have only subjective viewpoint. In actual, we identify joint attention from only our own viewpoint by imaging the other's attention.

Taking the subjective viewpoint into consideration, there are two types of triadic relation in the figure: *t_rel(R,P,O) and t_rel(P,R,O). t_rel(R,P,O)* is a triadic relation from a robot's viewpoint. *t_rel(P,R,O)* is the one from a human's viewpoint.

Robot: *t_rel(R,P,O)* ←*at(R,O),*
at(R,at(P,O))
at(R,at(P,at(R,O)))

Person: *t_rel(P,R,O)* ←*at(P,O),*
at(P,at(R,O))
at(P,at(R,at(P,O)))

The above two rules show what attention must be established to establish each subjective triadic relation. In short, if human has the three attentions, he/she thinks that joint attention is established. Even though the robot does not have the three attentions, the human imagines the establishment.

Also, if the robot has the three attentions, we can say that the robot establishes joint attention. The attentions can be established by an attention mechanism and EDD.

Robot: *t_rel(R,P,O)* ←*at(R,O),* ← Attention Mechanism
at(R,at(P,O)) ← EDD
at(R,at(P,at(R,O))) ← EDD

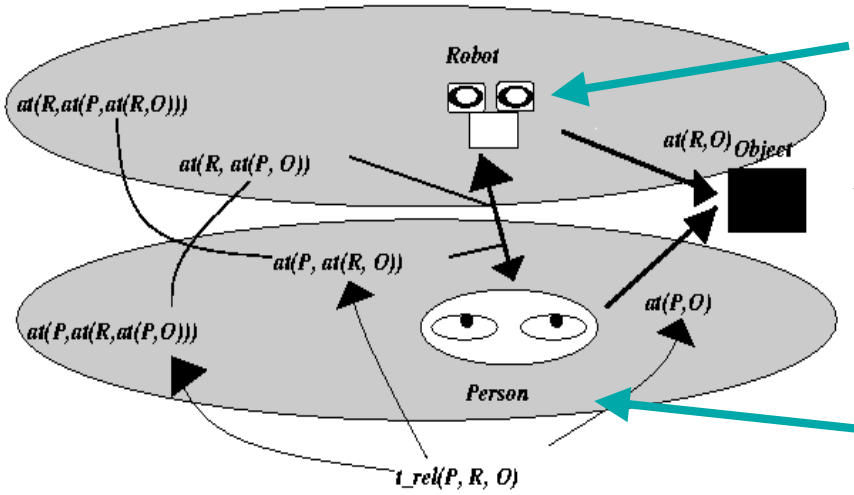
The attention mechanism can select relevant information from an environment. The function of the mechanism establishes *at(R,O)*. EDD recognizes the direction of human's gaze. The robot can infer the two attentions with the recognized gaze information.

However, the implementation of EDD is impossible as far as vision mothods are based on a camera mounted on a stable position. Since the cameras mounted on a robot moves frequently, we cannot use the recognition methods for EDD. In short, we cannot give a robot *t_rel(R,P,O)*. The robot obtains only *at(R,O)*.

Fortunately, we can establish joint attention between a human and a robot even though the robot cannot establish joint attention. As I discussed above, the humans imagine the establishment of joint attention when they have three attentions. Since the robot has only *at(R,O)* with the attention mechanism, joint attention has the following asymmetric structure in the actual human-robot interaction. Linta-III establishes this type of joint attention with the joint attention mechanism. It is sufficient for human-robot interaction to give only person joint attention.

Robot: *at(R,O)*

Person: *t_rel(P,R,O)*←*at(P,O),*
at(P,at(R,O))
at(P,at(R,at(P,O)))



Obtaining Real World Information with Attention Mechanism

The selection of sensor information with voluntary attention and reactive attention

Drawing Person's Attention

This figure shows what functions the robot needs to establish joint attention. That is the attention mechanism and a mechanism for drawing person's attention. The attention mechanism achieves *at(R,O)*. The mechanism for drawing person's attention achieves *at(P,O)*, *at(P,at(R,O))*, and *at(P,at(R,at(P,O)))*.

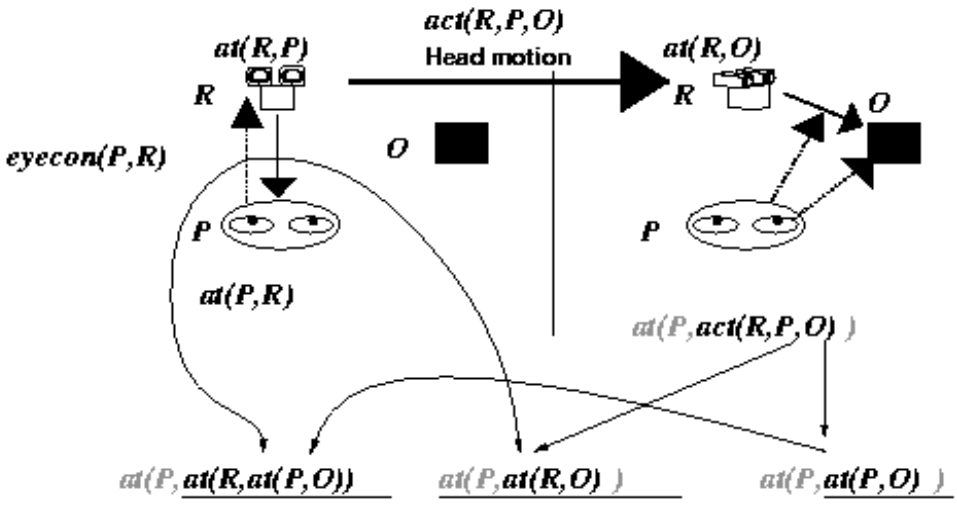
5.5.5 Attention Expression

You can achieve a mechanism for drawing person's attention with many types of robotics and computer technologies. The remaining subsections focus on the communication strategies for humanoid-robots. Since humanoid-robots have a figure similar to humans, we can apply the same strategy as humans to drawing human's attention. That is attention expression.

The attention expression is a gaze motion or a hand gesture. Humans frequently use the gaze motion or the hand gesture to draw other's attention. I introduce joint attention mechanisms based on the attention expression.

Eye-contact

Attention Expression

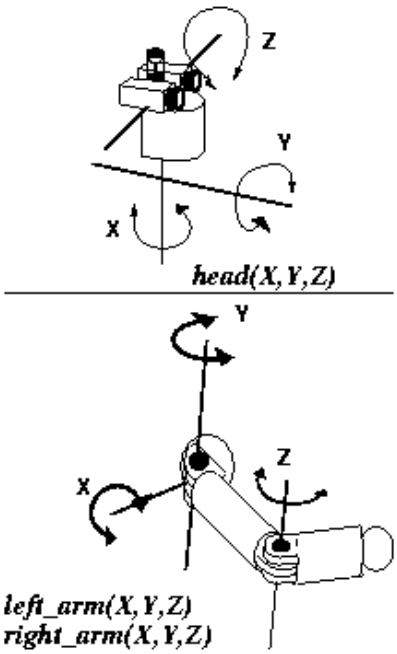


The figure in the previous page shows a process that the robot achieves three attentions of a person with attention expression. The attention expression used in the figure is gaze motions of a robot. The gaze motion consists of eye contact and an attention expression toward an object. The two gaze behaviors result in a robot head motion *act(R,P,O)* which expresses that the robot *R* turns its head between the person *P* and the object *O*.

At first, the human notice the robot action *act(R,P,O)*. Then, he/she notice the existence of the object, which is expressed as *at(P,O)*. Moreover, he/she also notice that the robot pays attention to the object, which is expressed as *at(R,O)*. However, he/she cannot obtain direct information about *at(R,O)*. The eye contact plays a role for him/her to confirm the belief of *at(R,O)*.

Also, he/she notice that the robot notice *at(P,O)* while he/she carries out eye contact with the robot. Throughout the interaction, the human has joint attention from his/her subjective viewpoint.

5.5.6Everyday Robot



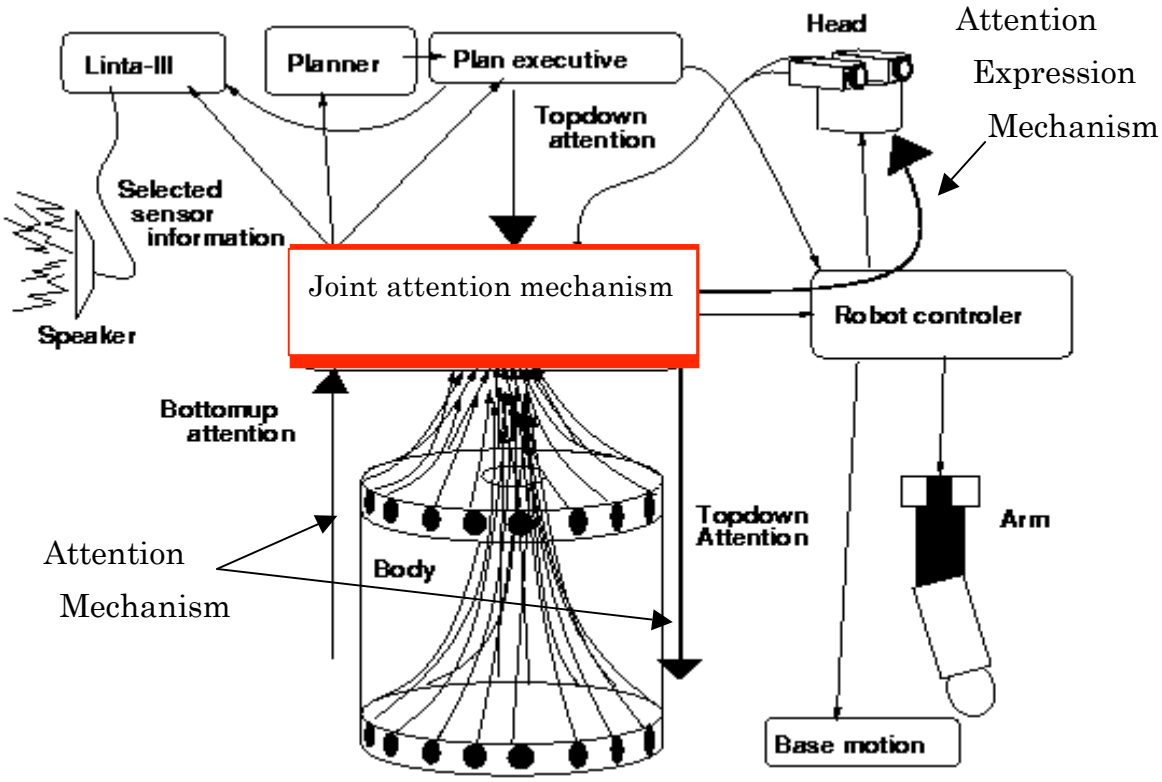
I employed a humanoid robot named Everyday Robot for the basis of joint attention mechanism. It has three DOF head and two hands whose DOF is also three. It also has sonar ring around it to measure the distance between it and an obstacle.

5.5.7 Joint Attention Mechanism I

The joint attention mechanism implemented on Everyday Robot is based on the model of joint attention introduced at subsection 5.5.5. That is the attention expression of the

head motion.

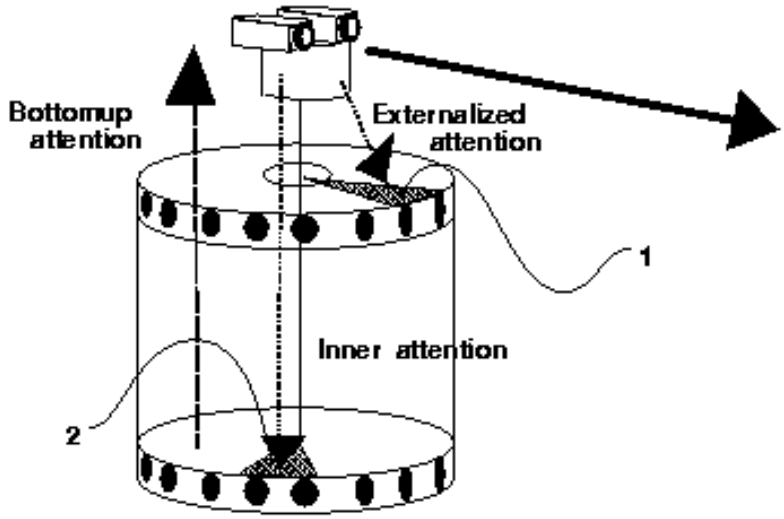
The joint attention mechanism consists of an attention mechanism and an attention expression mechanism.



This figure shows the structure of a system implemented on Everyday Robot. The primary parts of the system are Linta-III and the joint attention mechanism. Linta-III is a dialogue system which refers to information about the real world by using the joint attention mechanism.

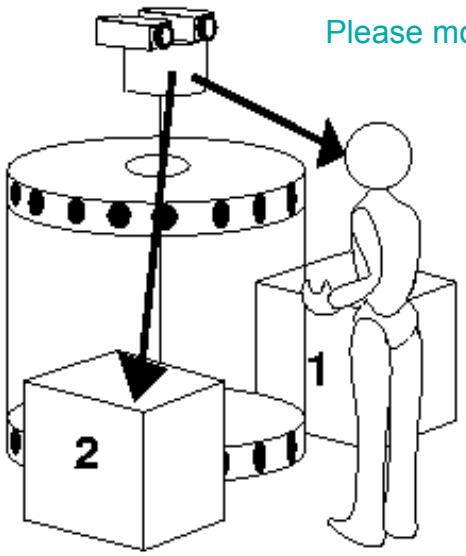
The joint attention mechanism includes the attention mechanism and the attention expression mechanism. The attention mechanism selects relevant sensory information with top-down attention when a plan executive gives the joint attention mechanism a command of top-down attention. Also, the attention mechanism directs its attention to an event which happens around Everyday Robot. The attention corresponds to bottom-up attention.

The attention expression mechanism is highly significant for this section. It turns the head of Everyday Robot based on the motion strategy of *act(R,P,O)* introduced at subsection 5.5.5. By carries out *act(R,P,O)*, it can draw human's attention to the target object and establish three attentions of human. In short, it establishes joint attention toward the object when the attention mechanism finds a relevant object.



This figure shows a scene that the attention expression mechanism turns the head of Everyday Robot to the direction of relevant sensory information. There are two pieces of relevant sensory information 1 and 2 in the figure. The attention expression mechanism manifests its attention toward 1. On the other hand, 2 is dealt as internal attention. Since the head does not turn to 2, humans never notice that Everyday Robot focuses on the direction 2. In short, the attention expression mechanism selects relevant sensory information intentionally for telling humans its attention. Moreover, the attention expression mechanism turns the head when an event occurs around Everyday Robot.

5.5.8 Experiment



Condition1 Eye-contact＋Attention Expression

Condition2 Eye-contact

Condition3 Attention Expression toward the box2

This experiment verifies the ability of the joint attention mechanism by confirming the effect of gaze motion (eye-contact and attention expression). The figure in the previous page indicates an experimental setting. There are two boxes 1 and 2 in the left and right sides of Everyday Robot. Experimental subjects stand in front of Everyday Robot. It directs its attention to the box 2 and tries to ask them to move this by saying "please move this." If humans do not notice that it focuses on the box 2, they cannot understand the utterance. The attention expression mechanism must make them notice the box 2 with the head motion.

There are three conditions in the experiment.

- Condition1 Eye-contact+Attention Expression toward the box2
- Condition2 Eye-contact
- Condition3 Attention Expression toward the box2

Condition 1 carries out *act(R,P,O)* to draw humans' attention to the box 2. On the other hand, Conditions 2 and 3 carries out one of the actions. Under these conditions, I observed the behaviors of the subjects. That is what they move in response to the request of the robot "please move this."

There are thirty subjects. They were divided into three groups equally. Each group corresponds to one of the conditions. According to the groups, each subject was given one of the conditions. The instruction which I gave them is "please obey a robot utterance."

5.5.9 Experimental Outcome

Condition1	Condition2	Condition3
9/10persons moved at the correct object. The 9 persons determined the target of "This" according to the head motion	2/10persons	3/10 persons Only 2 persons refers to the head direction.

The above table indicates the outcome of the experiment. The nine out of ten subjects moved the box 2 when they given condition 1. On the other hand, only two subjects moved it under condition 2 and only three subjects did under condition 3.

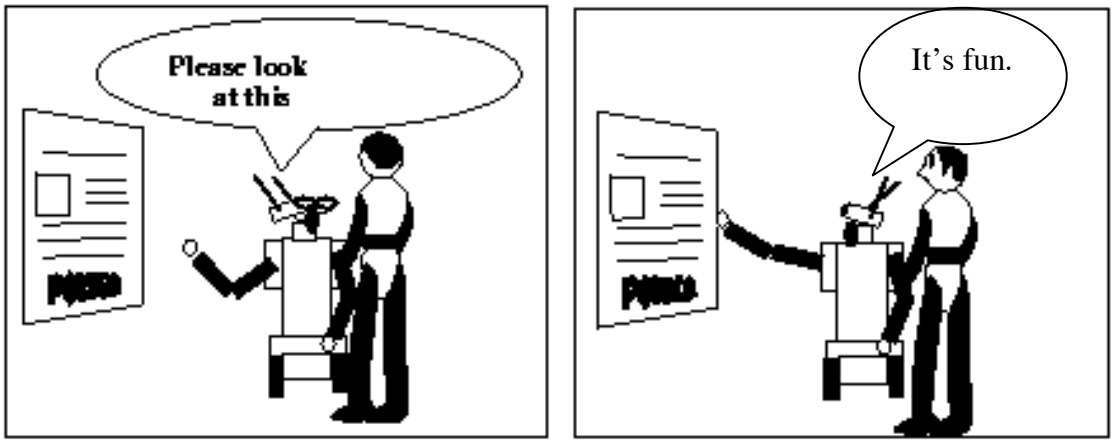
I also took questionnaire to know what information they used to identify the referent of "this." I gave the questionnaire to only subjects under condition 1 and 3 because the

condition2 did not give them information related to box 2. The answers indicate that all subjects who moved box 2 under condition 1 referred to the head motion. On the other hand, only two subjects referred to the head motion under condition 3.

These results indicate that eye contact and attention expression encouraged the development of joint attention.

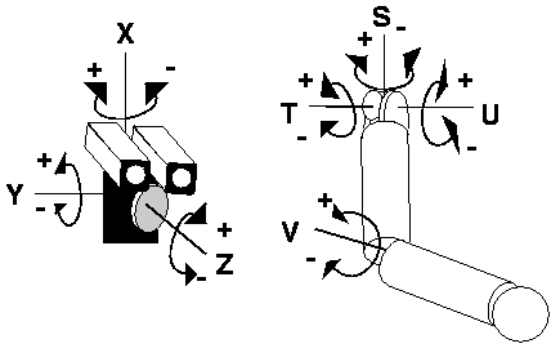
5.5.10 The Effect of Pointing Behavior

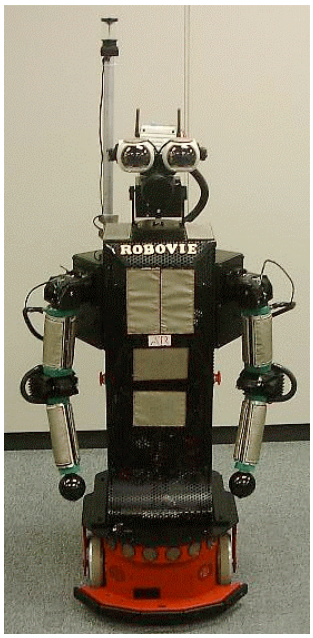
The study of infants' joint attention indicates that gaze detection is significant to construct joint attention. Also, pointing behavior is significant for us to draw the other's attention. We consider the relation between pointing behavior and eye contact.



I have frequently used the figure in this handout. In the figure, the robot uses a pointing behavior and eye contact to draw human's attention.

5.5.11 Communication Robot Robovie

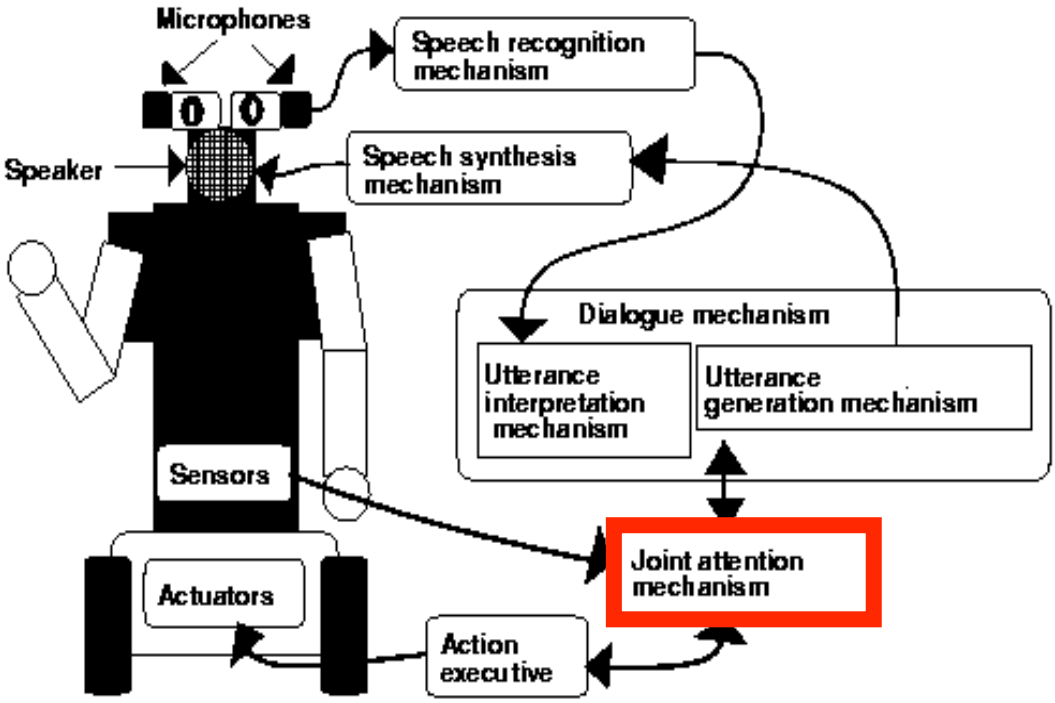




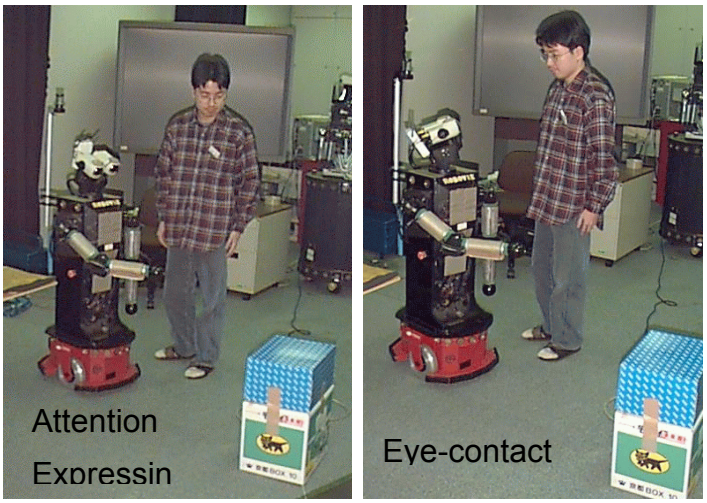
To investigate the relation between pointing behaviors and eye contact, I employed a

communication robot named Robovie and develop a new joint attention mechanism on it. Robovie has 3 DOF head and two 4 DOF hands. It has ultra sonic distance sensors around it and touch sensors on its arms, a stomach, and head. Moreover, it has a set of stereo vision sensors and an omi-directional vision sensor.

5.5.12 Joint Attention Mechanism II



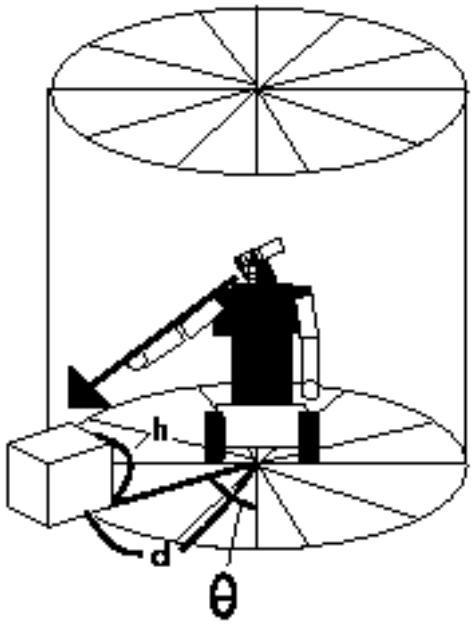
This figure shows a joint attention mechanism which is different from that implemented on Everyday Robot. The new joint attention employs pointing behaviors. It manifest its attention with the pointing behaviors and gaze motions.



In short, Robovie expresses its attention its pointing behavior and its head direction

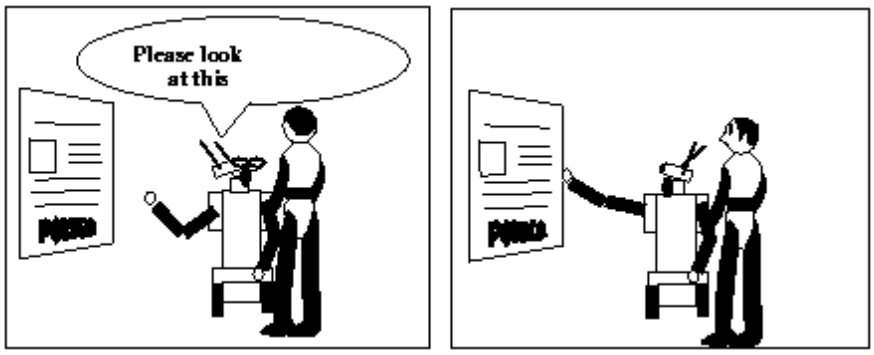
while it carries out eye contact.

5.5.13 Attention Coordinate



To generate pointing behaviors and gaze motions, the joint attention mechanism has an attention coordinate like the figure. The attention coordinate is a polar coordinate. When Robovie finds an object with its sensor, an angle from the front of Robovie and a distance between Robovie and the object are expressed on the coordinate. Then, the arm pose and the motion of Robovie's head are generated based on the information.

5.5.14 Experiment



I have conducted an experiment to verify the relation between eye contact and pointing behaviors. I prepared two conditions to confirm the fact as follows.

- Condition 1 Pointing behavior with eye-contact and attention expression
- Condition 2 Pointing behavior without them

The target of the attention is a poster hanged on a wall like the figure. When Robovie

proceeds to the poster, it generates an utterance “please look at this.” The behaviors of Robovie vary depending on a given condition. That is, it points at the poster with its arm while moving its head, or it points at the poster with only its arm. In the second condition, Robovie never turns its head while generating utterances.

I observed where subjects looked in response to “please look at this.” If they focus on the poster correctly, they can understand that the word this refers to the poster. Otherwise, they cannot.

There were twenty subjects. They were divided into two groups equally. Each group corresponds to one of the conditions. In short, Robovie’s behaviors were different depending on the condition given to each subject.

I gave the subjects an instruction “please obey robot’s utterances.”

5.5.15 Experimental Outcome

Condition1: with eye-contact and attention Expression	Condition2: without head motion
All subject (10 persons) looked at the poster.	Only one person (10 persons) looked at the poster. The other 9 looked at the hand.

The above table indicates the results of the experiment. In condition 1, all subjects noticed that the word “this” referred to the poster and they also looked at the poster. On the other hand, only one subject looked at the poster in condition 2. The other nine looked at the hand and have also understood that the word “this” indicated the hand or Robovie.

The difference of the result comes from the only difference of gaze motions. The fact indicates that the gaze motions have more power than pointing behaviors to manifest the attention of a robot.

5.5.16 The Degree of Embodied Expressions and Language Expressions

The previous subsections have explained two types of joint attention mechanisms. However, joint attention cannot be always established when the mechanisms carries out attention expressions. Moreover, a robot must use a different language expression depending on the degree of established joint attention. Since the attention expressions heavily depend on embodied expressions, the following subsections deal with a relation between the degree of embodied expressions and language expressions

5.5.17 Utterance Generation based on Joint Attention

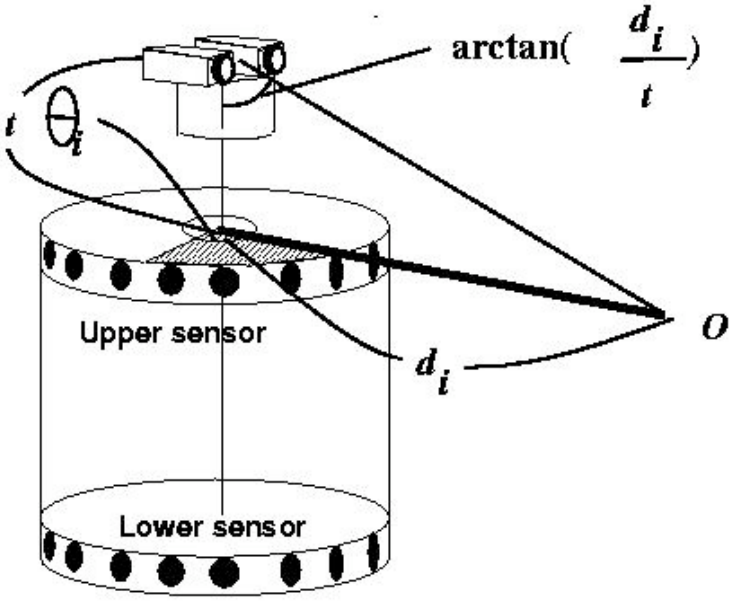
The joint attention mechanism establishes asymmetric joint attention as I explained at 5.5.4. The joint attention is not the same as the one between humans. Conversations on the joint attention are also different from humans.

The interaction on the asymmetric joint attention corresponds to robot oriented interaction. The communication skill of a robot is lower than humans. However, we can establish human-robot interaction by utilizing higher recognition skill and intelligence of a human. The achieved interaction is just the robot oriented interaction. The robot introduces a story of a situation and draws human's attention to it. As far as humans play along with the robot, they achieve smooth interaction.

The asymmetric joint attention is also an example of the robot oriented interaction. The robot selects relevant sensory information and draws human's attention to it. After drawing his/her attention, the robot generates an utterance. If he/she does not play along with the attention drawing, the interaction is never established.

Since the interaction depends on whether or not a human plays along, a robot must take into account the degree of attention expression and joint attention when generating an utterance. At first, I deal with the theme of the degree of attention expression.

5.5.18 The Degree of Attention Expression



The degree of attention expressions is defined by the posture of the robot head because there is not any other information except for the posture when humans identify the robot's attention. Even if the robot has a hand, the attention expression by the gaze motion is much more significant than the hand as the experiment result at

5.5.15 indicated.

The degree of the attention expressions is defined the measurement shown in the figure of the previous page. To express the degree, I employ *t_val(P,R,O)* which suggests the degree of joint attention toward an object *O* between a human *P* and a robot *R*. The value is a set of *at_val*s.

- $t\_val(P,R,O)=(at\_val(R,P),at\_val(R,O))$

The *at_val* denotes the degree of an attention expression. *at_val(R,P)* is the degree of the attention expression toward a human and *at_val(R,O)* is that toward an object.

The value of *at_val* is determined depending on the following constraints.

$$-90^\circ \le \theta_i \le 90^\circ$$
$$30^\circ \le -\arctan(d_i / t)$$
$$d_i \le 1.5m$$
$$d_i \le 4.5m$$

Here, *i* denotes a human or an object. The relations between the value of *at_val* and the constraints are as follows.

- When all conditions are satisfied, the value is "c." The constraint means that humans easily notice the attention.
- When three conditions are satisfied, the value is "p." Humans probably notice it.
- When less than equal two conditions are satisfied, the value is "u." Humans hardly notice it.

For example, let us consider the following situation.

- The location of a person

  $\theta_i = 0^\circ, d_i = 1m, t = 1.7m$
- A object's location

  $\theta_i = 40^\circ, d_i = 3m, t = 0.5m$

At the situation, we obtains the following the degree of joint attention.

- $t\_val(P,R,O)=(at\_val(R,P),at\_val(R,O)) = (c,p)$

5.5.19 Utterance Generation depending on the degree

Depending on the degree of joint attention, a robot can select an appropriate utterance expression. If joint attention is established completely, the robot can omit a word which refers to obvious event in a physical world. An utterance generation mechanism must take two steps to achieve the adaptive generation: constructing joint attention and confirming the degree of the attention.

The following sequence is the generation of an utterance under joint attention.

1. A robot directs its attention to physical world event *O* spontaneously with

attention mechanism.

2. A planner determines *O* must be mentioned and generates intentional attention from the planner.
3. Planner generates utterance content.

   $p1(R,P,p2(P,O))$, $f(\boldsymbol{R})=R$, $f(\boldsymbol{P})=P$, $f(\boldsymbol{O})=O$
4. Eye-contact and Attention Expression to *O*
5. Checking the degree of the attention expression
6. Generating an utterance based on the degree

Here, $p1(R,P,p2(P,O))$ is the content of an utterance. *p1* denotes a predicate, a first argument denote agent of the predicate, second and third arguments are objects of the predicate.

5.5.20 Checking the degree

$t\_val(P,R,O)$	*O*
(c,c)	An utterance uses a demonstrative pronoun
(c,p)	An utterance uses direct expression with adjectives
(p,c)	An utterance uses a direct expression
(p,p)	An utterance uses a direct expression with adjectives
$(u,\text{-}),(\text{-},u)$	No utterance (- denotes DON'T CARE)

The table indicates what utterance a robot generates according to the degree of joint attention. The sixth row suggest that the robot does not generates any utterances under the degree of $(u,\text{-})$ or $(\text{-},u)$. In other word, if $t\_val(P,R,O)\neq$(-,u), (u,-), the robot generates an utterance. $(u,\text{-})$ or $(\text{-},u)$ means that the attention expression toward a human or toward an object is hardly noticed by humans. Since it is impossible to establish joint attention under the situation, the robot does not generate an utterance.

(c,c) indicates that the robot can perfectly carries out attention expression toward both of a human and an object. In the situation, the robot generates an utterance by using a demonstrative pronoun.

(c,p) indicates that the attention expression toward a human is perfect but that toward an object is difficult. In the situation, the robot use a direct expression and adjectives to refer to the object in the utterance.

(p,c) indicates that the attention expression toward a human is difficult but that to the object is easy. Since it is easy for him/her to notice the object after realizing that the robot says something to him, the robot uses only direct expression to refer to the object.

(p,p) indicates that both attention expression are difficult. In the situation, the robot also uses a direct expression and adjectives to refer to an object.

5.5.21 Example Interaction

Let us consider an example of interaction based on the degree of joint attention. We assume that Everyday Robot finds a object at the following poition.

$$(d_o, \theta_o) = (s_j, 360j/24) = (0.5, -10)$$

Then, it has intention for a person to remove it as *ask*(*R,P,move*(*P,O*)). The expression means that Everyday Robot asks a human to move an object. Here, the locations of the human and the robot are as follows.

$$(d_p, \theta_p) = (1,50), (d_o, \theta_o) = (0.5, -10)$$

Based on the location, Everyday Robot carries out eye contact and attention expression. In the situation, the degree of joint attention becomes *t_val*(*P,R,O*)=(*c,c*) because both locations satisfy all conditions. As a result, it connects a demonstrative pronoun "this" to the object by the function f(this)=*O*. At last, Everyday Robot generates an utterance "Please move this."

Let us consider another situation where the location of humans is different from the former example as follows.

$$(d_p, \theta_p) = (1,100), (d_o, \theta_o) = (0.5, -10)$$

In the situation, the degree of joint attention becomes *t_val*(*P,R,O*)=(*p,c*). According to the degree, Everyday Robot selects a direct expression "obstacle" by using a binding function f(obstacle)=*O*. Then, it generates an utterance "please move the obstacle."

The next example shows that the object is also in the different location as follows.

$$(d_p, \theta_p) = (1,100), (d_o, \theta_o) = (4, -10)$$

In the situation, the degree becomes *t_val*(*P,R,O*)=(*p,p*). According to the degree, Everyday Robot selects a direct expression with an adjective such as f(front object)=*O*. At last, it generates an utterance "Please move the object in front of me."

5.5.22 The Degree of Development of Joint Attention and Language Expression

The several previous sections deal with the degree of joint attention in terms of attention expression. On the other hand, it takes some steps to establish joint attention. For example, the robot must turn its head to a target and to a human to make him/her notice the target while generating pointing gesture. The remaining part of this section explains the degree of joint attention in terms of the process of its development.

5.5.23 The Stage of Joint Attention Process

I introduce the concept of the stage of joint attention process which expresses how degree joint attention is established in the course of achieving the joint attention. Utterances generated by a robot must also depend on the degree.

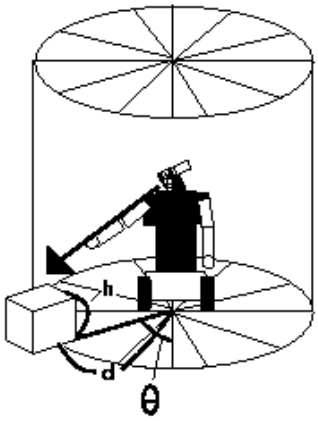
To expresses the degree, I employs a notation expressing sensory information as follows.

$$f[(\theta,d,h)]$$

Here, *f* denotes *p, o, e, r.* Each symbol expresses a person, an object, a poster, Robovie. Moreover, since a person directs his/her attention to *a*, I use a specific expression for sensory information of a person as follows.

$$p[(\theta_p,d_p,h),a]$$

Each parameter comes from the attention coordinate of Robovie.



The following items indicate the development of joint attention and its expressions.

- When a person and a poster are near Robovie,

$$p[(\theta_p,d_p,h),-],e[(\theta_e,-,h)]$$

Since Robovie cannot identify where the human direct his attention, the slot for a human's attention does not have any value.

- After Robovie carries out Eye-contact, the person may notice Robovie. The situation is expressed as follows.

$$p[(\theta_p,d_p,h),(r[(0,0,0)])],e[(\theta_e,-,h)]$$

In the expression, the slot for a human's attention includes *r*[0,0,0]. It indicates that the human notice Robovie. Also, r[0,0,0] indicate that the robot is in a center of the attention coordinate.

- After Robovie carries out attention expression toward the poster, the person may also notice the poster. This state corresponds to joint attention.

$$p[(\theta_p,d_p,h),(r[(0,0,0)],e[(\theta_e,-,h)])],e[(\theta_e,-,h)]$$

5.5.24 Utterance Generation Rule

This subsection explains rules for generating an utterance depending on the degree of joint attention development. I show the rules as follows.

1. $v(-,f)$ → $v(r,-)$

2. $v(-,f), p[-,-]$ → Hello, $v(-,f)$

3. $v(-,f), p[-,(r,-)]$ → $v(-,f)$

4. $v(-,f), p[-,(r,f)]$ → $v(-,dp)$/v(-,-)

The rule has a structure of *utterance content, degree* → *altered utterance content.* The first rule indicates a situation where there is not a human around Robovie. In the situation, it cannot get the degree of joint attention development because it is impossible to establish joint attention without a human. Then, Robovie translates the content of utterance into an utterance expression whose subject is a robot and uses an intransitive verb.

The second rule indicates a situation where there is a human but he/she does not notice the robot. In the situation, Robovie generates a greeting and generates the utterance by using a direct expression.

The third rule indicates a situation where the human notices the robot but does not notice an object. To make him/her notice it, Robovie generates an utterance by using a direct expression.

The fourth rule indicates a situation where joint attention is established between a human and Robovie. Since he/she already notices the object, Robovie generates an utterance by using a demonstrative pronoun.

Let us consider an example of generating utterances depending on the rules. When Robovie is stuck by an obstacle, it has an utterance content $move(p,o)$. Also, if there is no human around it, it selects the first rule. The content is translated by the rule as follows.

$move(p,o) \rightarrow \neg proceed(r)$

Then, it generates an utterance "I cannot proceed!"

Next, we assume that there is a person but he/she does not have a relation with the robot. That is p[-,-]. In the situation, Robovie has an utterance content $look(p,e)$. The second rule is used to generate an utterance in the situation. In short, the following rule are used.

$look(p,e)$,p[-,-]→Hello, $look(p,e)$

Then, Robovie generates an utterance "Hello, please look at the poster!"

When eye-contact with a person is achieved, the degree of the joint attention development becomes p[-,(r,-)]. *Also, the content is* $look(p,e)$. In the situation, The third

rule is used as follows.

look(*p,e*),p[-,(r,-)]→*look*(*p,e*)

Then, Robovie generates an utterance “Please look at the poster!”

When joint attention is achieved, the degree becomes p[-,(r,e)]. In the example, we assume that Robove has an utterance content *is-a*(*e,fun*) which means that the poster is fun. Since joint attention is already established, the fourth rule is selected.

is-a(*e,fun*),p[-,(r,e)]→is-a(dp,fun)

Then, Robovie generates an utterance “It is fun!” by using demonstrative pronoun “it.”

6.Sociality of Systems

This chapter considers the sociality of interactive/communication systems. Many researchers have ignored the effect of the sociality while designing their systems in terms of only engineer. That is, they have focused on developing the functions of the system. However, the sociality actually has effects on interaction between humans and systems. The chapter explains the effect as follows.

- The Effect of Sociality on Human-Computer Interaction
- The Sociality of Agent Character
- Social Cues
- The Sociality of Robots

Let us start to consider the effect of sociality in human-human interaction. Humans refer to the other's position in society, the other's looks, and the quality of voice. They are sometime affected by the other's kind or unfriendly response.

The sociality is optional when we interact with others. The works which we take there is to giving or receiving massage or utterances each other. However, we change the mode of interaction according to the sociality in actual.

For example, we feel authority when meeting a man wearing a uniform of a security guard. Moreover, we tend to believe utterances given by them without evidence. The behaviors of humans refer to the social position of others.

Systems (computers, agents, and robots) must consider the effect of their sociality when interacting with humans. As the human-human interaction indicates, we can expect that the sociality of the systems will have effects on us. In actual, this chapter shows you the evidence of the effects. Moreover, you learn what design gives the systems the power of sociality.

6.1The Effect of Sociality on Human-Computer Interaction

Computers also have social effects on interaction with humans. A book entitled Media Equation has introduced many findings related to the sociality of computers. The findings indicate that humans treat a computer as a human.

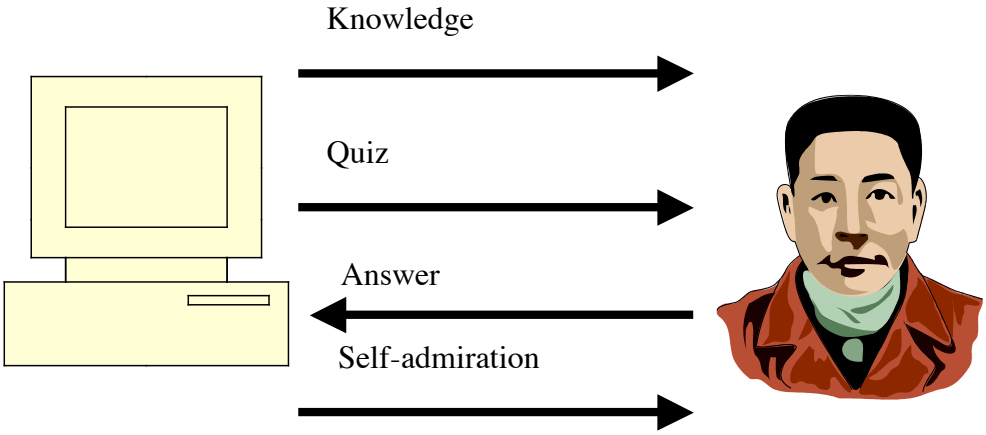
The authors of the book are Byron Reeves and Clifford Nass who have conduced many psychological experiments to reveal the sociality. Here, The word 'media' corresponds to videos, voices, computer applications, and autonomous agents.

Humans behave socially in response to others in a communication. For example, they show their appreciation when someone does something for them. Humans also respond to media socially as if they interact with humans even though they are aware that it is just a machine which does not have mind. However, there is scientific evidence which

supports that the social response actually occurs. The evidence suggests that social design is also important for human-robot interaction.

I show you an example of Media Equation. Humans consider that actors/actresses in a movie have minds. However, they are just images on films. Humans assume a mind on the images. The ability of humans is called anthropomorphic ability （擬人化能力).

Experiment related to politeness



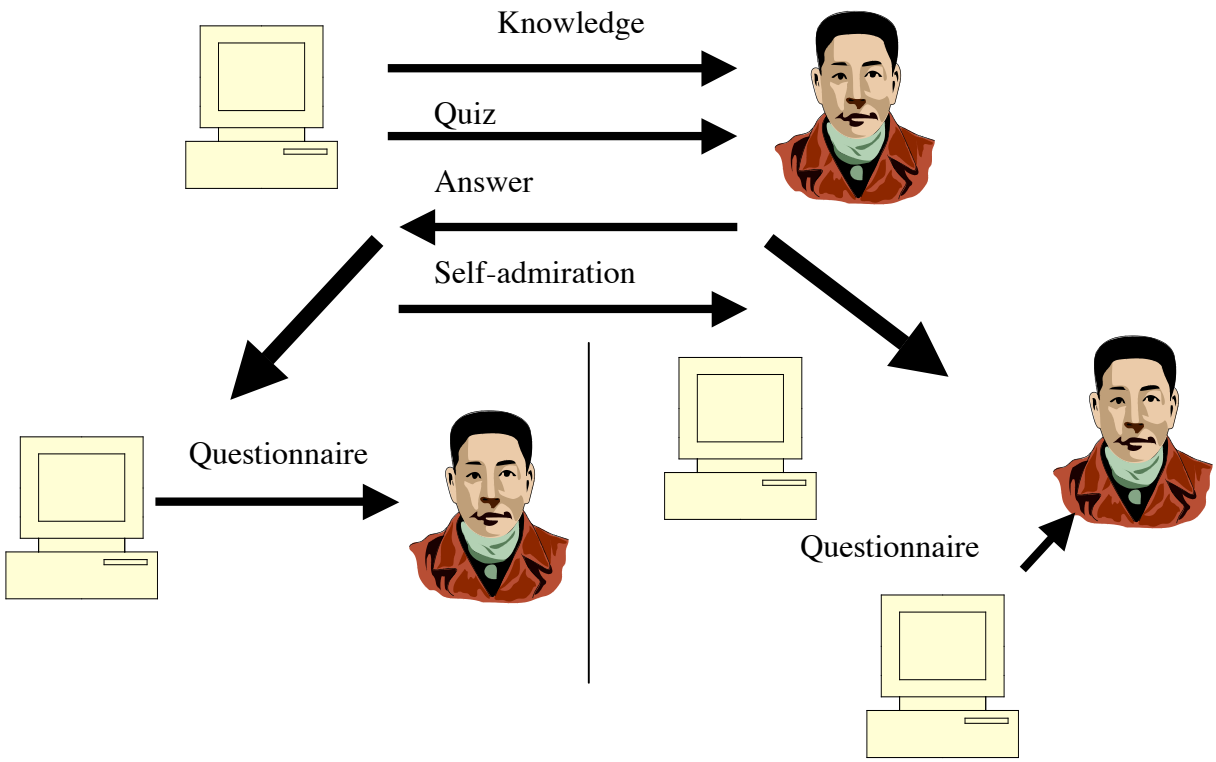
I introduce an experiment which suggests the sociality of a computer. There are many types of sociality in interaction between humans. This experiment picked politeness for one of the sociality to apply it to human-computer interaction. The experiment have verified whether humans interact with a computer politely.

The above figure indicates a sequence of interaction between a human and a computer. The point of the sequence is that the computer admires itself. The exact sequence consists of four sessions. At first, the computer gives him/her knowledge about a topic. After this, it gives him/her several quizzes related to the topic. He/she must answer the quiz. After three sessions, the computer admires itself.

After the interaction, an experimenter gives a questionnaire “are the computer/software’s knowledge useful for you?” If humans behave politely, they gives a good evaluation for the computer which gives them the knowledge about the quiz event though the given knowledge is not so useful for them.

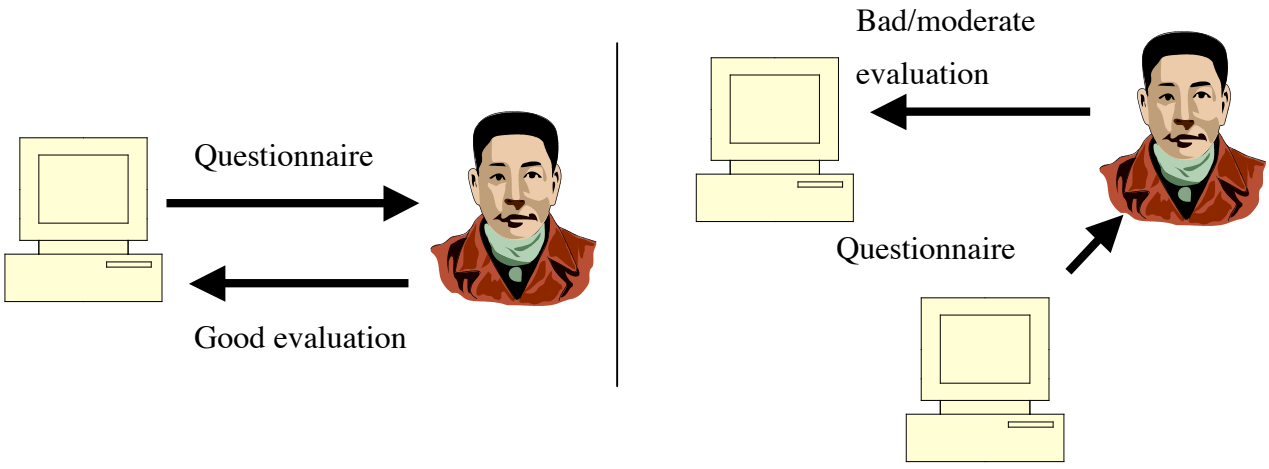
The experiment prepared two experimental conditions when the experimenter gives the humans a questionnaire. The conditions are shown in the figure in the next page. In the first condition, the computer used in the sequence of interaction gives humans the questionnaire. In the second condition, another computer which is different from the one used in the sequence gives humans the questionnaire.

We can expect the following results of human’s response if a person treats computers socially.



・ He/she flatters the computer and gives good answer.
・ He/she tells the other PC the truth about the computer.

The next figure shows the predictions. The responses in the predictions indicate that the humans behave politely if the computer asks them the evaluation of itself. On the other hand, they does not if the other computer asks them the same evaluation.



Here, we have a question about the social response of humans. Does the social evaluation actually occur? Computers do not have emotion. Computers do not have any personalities. If the predictions are true, it is funny that we respond socially to the mindless machine.

Before showing the result of the experiment, I explain the experiment in detail. Test subjects (被験者) are given 20 sorts of knowledge about today's American affairs. For

example, 30 % of 10 years old American kisses at the first date. Then, the subjects select a response to the knowledge from “I know it well,” “I almost know it,” and “I do not know it.” The computer gives bits of knowledge about it depending on the response. However, it gives the same knowledge regardless of the response in actual.

After giving the knowledge, the computer gives quizzes to him/her. It tells which is right and which is wrong after the subjects answer each quiz.

At last, the computer admires itself like “I think that I’ve done best jobs.”

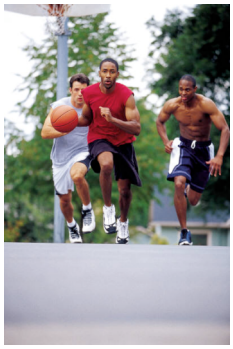
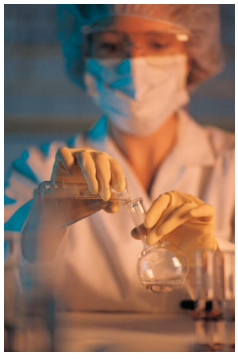
The evaluation of the computer is done along the experimental conditions shown in the previous page. The experimenter employed adjectives with which the subjects can express the ability of the computer. The adjectives are correct, analytical, useful, fair, friendly, or considerable. The subjects determine the degrees of each adjective as evaluations of the computer.

Experimental results indicated humans’ social response. They gave good evaluation to the computer which gave them the knowledge but they gave bad or moderate evaluation when the other computer asked. The result suggests politeness toward the computer even though it did not have emotion or autonomous abilities.

Experiment related to Specialist

I introduce the other example of the sociality in interaction. The sociality comes from human’s position title as a specialist. The title as a specialist gives significant factor not only for human-human interaction but also for human computer/robot interaction.

What is a specialist? We consider that he/she has proper knowledge. And he/she can make appropriate comments on his/her field. Based on these assumptions for a specialist, we behave socially in response to them.

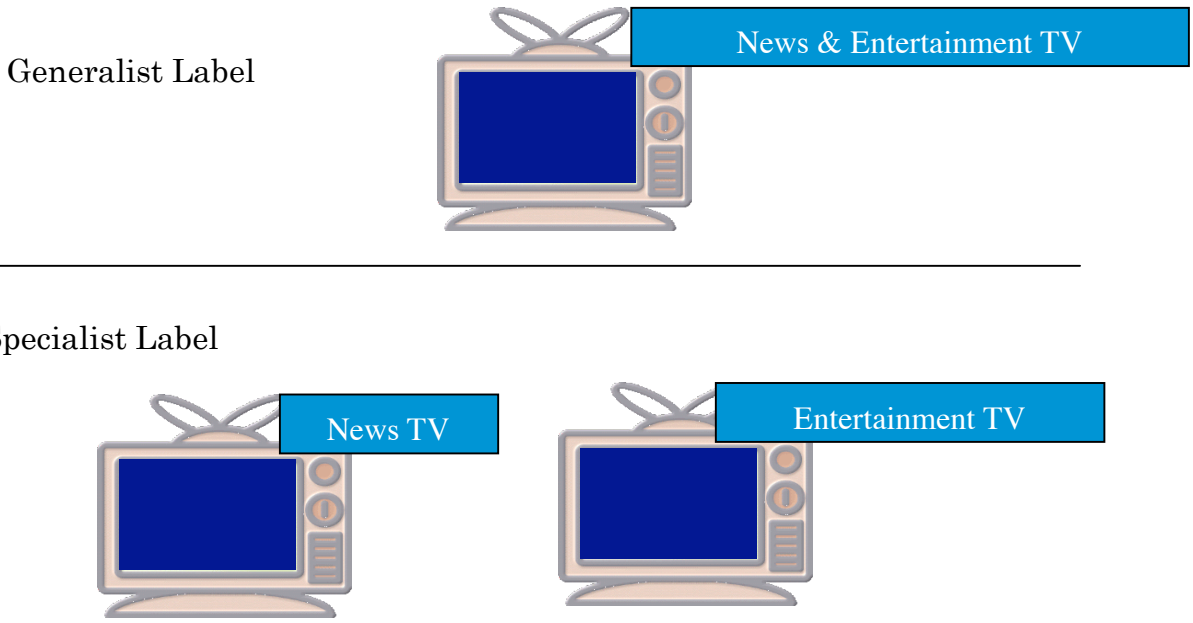




There is a question about a specialist. That is how people identifies who is the a specialist. The answer is a label. We believe the label of others. The label corresponds to just a title which they have. We do not confirm the label.

The reason why we believe the label in interacting with others is because it is easier to refer to the label than to confirm it. The confirmation requires much more cost than

interaction.

I introduce an experiment which verified what effect the title as a specialist gives media. The following figure shows the method of the experiment. The experiment compared two types of TV. A TV has a label as a generalist. The label shows “News & Entertainment TV.” The other TV has a label as a specialist. It shows “News TV” or “Entertainment TV.” The experimenter merely attached the labels on each TV.



If humans respond socially to TV based on the label, they evaluate the TVs differently even though all TVs give the same contents. To confirm the question, the experiment prepared the same contents for all TVs and compared the effect of the labels.

The experiment employed words to evaluate the contents of TV. It must be noticed here that the target of the evaluation is contents of each TV. TVs themselves are not the target.

There are two sets of words for the evaluation. One is a set of words for News contents: quality of NEWS, appropriateness, significance, usefulness, enjoyment, and seriousness. The other is for Entertainment contents: quality, appropriateness, enjoyment, and degree of relaxation.

The results of the evaluation indicated that the subjects regarded NEWS on News TV as more reliable than those of the generalist TV. Also, they regarded entertainment contents on Entertainment TV as higher quality than the one of the generalist TV. Since the contents of all TV are same, the difference of the evaluations came from the difference of the labels.

The label as a specialist actually has an effect on the humans as the experiment has shown. However, they cannot be conscious of the effect of the label. The experiment

also confirmed this fact. When the subjects were asked the factor of the contents' reliability after the evaluation, they would not select the label factor. The result indicates that the label gives an unconscious factor to interaction.

The effect of the label gives us a significant point to design a communication system. What a person identifies for Media/a system has a significant effect on interaction with it. For instance, if a robot has a label as a tour guide specialist, humans interpret the behaviors of the robot in terms of a guide robot. In short, we must design social relationship between a person and a machine to give it a kind of a proper label.

Capping game software (しりとりソフト)

The last evidence of the sociality of a system is a Capping game software. A person input a word to it. It outputs a word which starts from the last Japanese character of the input word. It treats only noun. It rejects a word which finishes by Japanese character NG.

Y.Yamamoto have conducted an experiment with the simple software. He measured the degree of pleasure in interacting with the software. He confirmed whether the pleasure varies depending on whether humans consider that a person play with them throughout the software. When they thought that they interacted with someone, the pleasure increased. On the other hand, when they noticed that they interacted with just a software, the pleasure decreased.

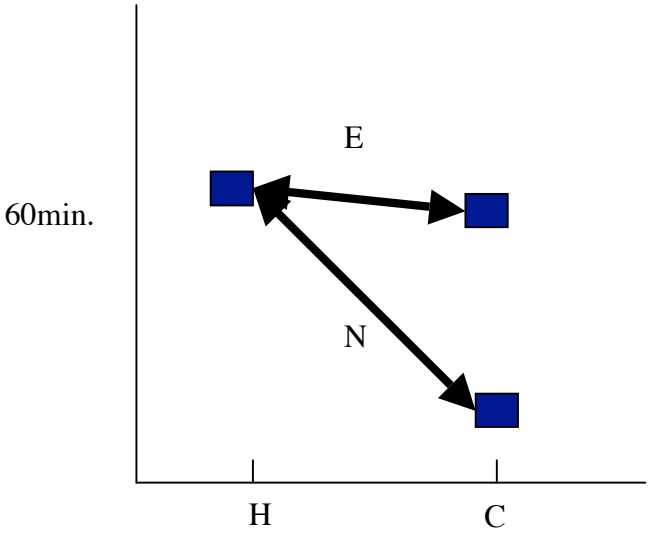
The precise setting of the experiment was as follows. There ware 24 test subjects (SFC students). The experiment prepared four experimental conditions.

- H: "You play a network capping game with a person."
- C: "You play capping game software."
- E: "This is an experiment."
- N: "This is just a game."

The conditions were used by combining H or C with E or N. So, each subject was given one of instructions based on the conditions HE, CE, HN, or CN. For example, the

instruction was “This is an experiment. From now, You play the network capping game with a person.” when a subject was given HE condition.

Moreover, an experimenter told subjects that they could stop the capping game when they were bored. The experiment measured the time of the interaction. The expeiment employed the time as a measurement of the degree of the pleasure.

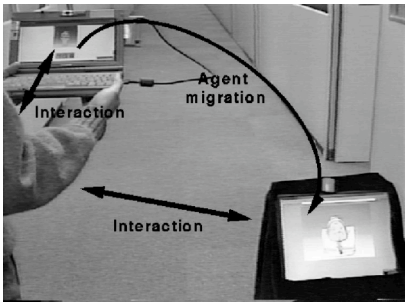


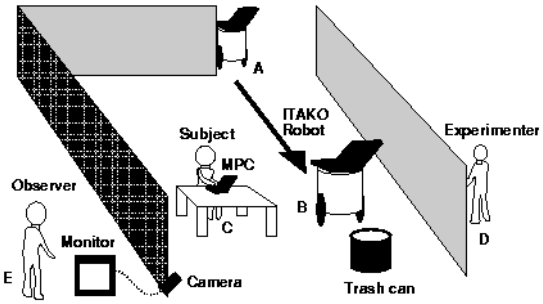
This graph is the result of the experiment. The horizontal axis denotes the difference of experimental condition between H and C. The vertical axis denotes the time from the start to the end of the interaction with the software. There are two line in the graphs; they correspond to E or N. There are interaction (交互作用) between E and N.

Let us focus on the result of N because there is the significant difference of the time between H and C in N but there is not in E. HN and CN used the same software. The difference between them was only the instructions given to the subjects. These were “you play capping game software” and “you play a network capping game with a person.” In short, persons’ identification of the software (a human or software) affects the degree of the pleasure of the interaction. Moreover, we can say that the difference between H and C corresponds to the difference of labels. The result indicates that a label or an appearance give a system kinds of factor related to sociality.

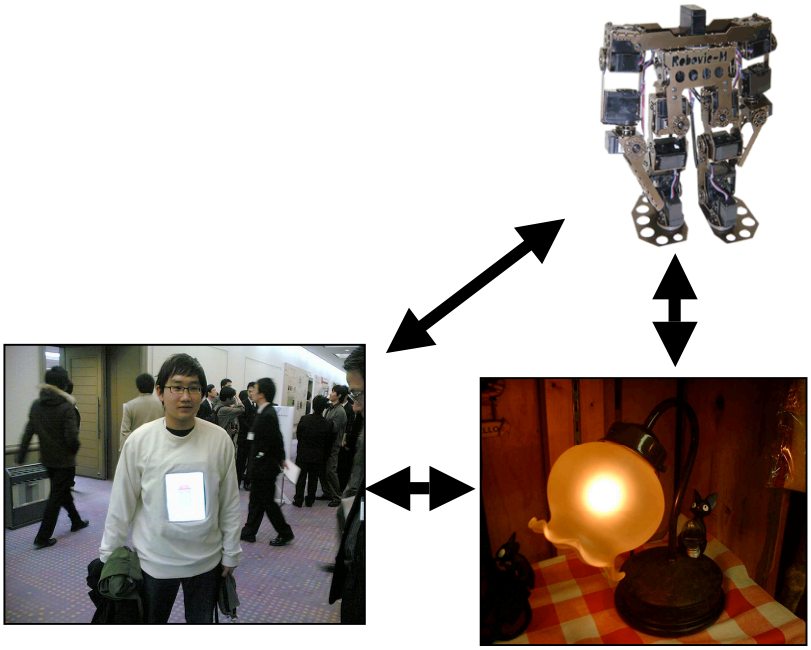
6.2 The Sociality of Agent Character

Autonomous CG characters have much more social effect than the computers or Media. This section explains the factor related to CG agents.





The agent can establish a relationship with humans. Even if the agent migrates to different media, it can maintain the relationship. The feature of the relationship was already explained at section 5.4. The above picture and figure show the experiment where the relationship established between humans and the agent on a mobile PC could be also established on the robot after agent migration to the robot.



Although the experiment employed the visual aspect of the agent, the agent can use the other modalities to maintain the relationship. The above figure shows the use of consistency in the tone of an agent. The CG agent is also shown on a display attached on a wearer and establishes a relationship with a man in the left picture. The agent can migrate to a lamp or a humanoid robot. The human notice the relationship with the devices when he hears the same tone of agent's voice.

The appearance of the agent plays a role as a basis of the agent sociality. In other word, the appearance corresponds to labels of Media. The humans establish a relationship with the agent by referring to the appearance.

On the other hand, continuous interaction maintains social relationship with humans when the agent migrates to the other device.

6.3 Social Cues

This subsection explains what factor a humanoid robot can use to establish a relationship with humans. It is difficult for a robot to achieve social interaction with a human. I have already shown several psychological experiments related to the social interaction.

The experimental results have indicated that humans behave socially toward the computers even though it does not have emotion. The label of a system also affects interaction. The invariance of relationship between human and agents gives a robot the relationship when the agent migrates to the robot.

The results give us knowledge about what cues a robot can use to establish social interaction.

The first one is label as a social robot. You can give a robot the outfit of sociality. For example, each outfit of a pet robot, a guide robot, and a guard robot gives the robot a kind of sociality. They induce humans' social response.

Also, continuance interaction establishes the relationship between a human and a robot. Humans have a relationship with a pet robot while they plays with it everyday.

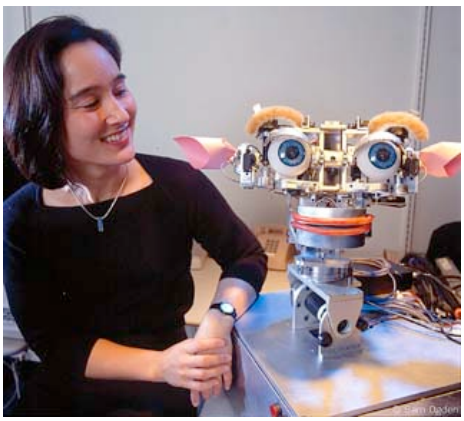
Embodied interaction is also significant when a robot establish the relationship. As the experiment at section 5 has shown, eye-contact and gestures induces communicative relationship between the human and the robot. Moreover, synchronization of gestures also induces the relationship. Humans feel a relationship with the others when gestures of the others are synchronized with those of themselves.

These factors are called social cues when they induce humans' social response.

6.4 The Sociality of Robots

This subsection introduces an example of a social robot. The name of the robot is Kismet. The photograph in the next page shows its figure. It was developed by Cynthia L. Breazeal.

The motivation of the development of Kismet is to make a sociable robot like human beings



The design of Kismet is based on the study on developmental psychology. She pursues sophisticated interaction where people can interact, play, and teach Kismet as naturally as they would an infant or very young child. And it deals with attention and joint attention by using gaze awareness. Its behavior induces human to imitate its behavior. The interaction based on imitation also establishes a relation between them.

The interaction between humans and Kismet is designed as a metaphor of infant-caregiver interaction. It interprets and sends readable social cues such as gaze direction, facial expression, body posture, and vocalizations.

The social robot requires the following items to achieve social interaction.

- Social environment (social persons)
- Real-time performance: natural interactive rate
- Establishment of appropriate social expectation: consideration of software and hardware limitations
- Self-motivated interaction

The first item indicates that there must be social humans around the robot when it achieves social interaction. Although you may think that this requirement is trivial, in actual situation, humans do not behave socially in response to a robot. The robot can achieve social interaction when it interacts with social humans.

The second requirement indicates that the robot must generate its response by natural response time. If the response is too slow, humans cannot engage in its interaction.

The third requirement indicates that the robot must generate behaviors which humans expect as social response to their actions to the robot. The design of the behaviors must be reflected in the limitations of the software and the hardware of the robot.

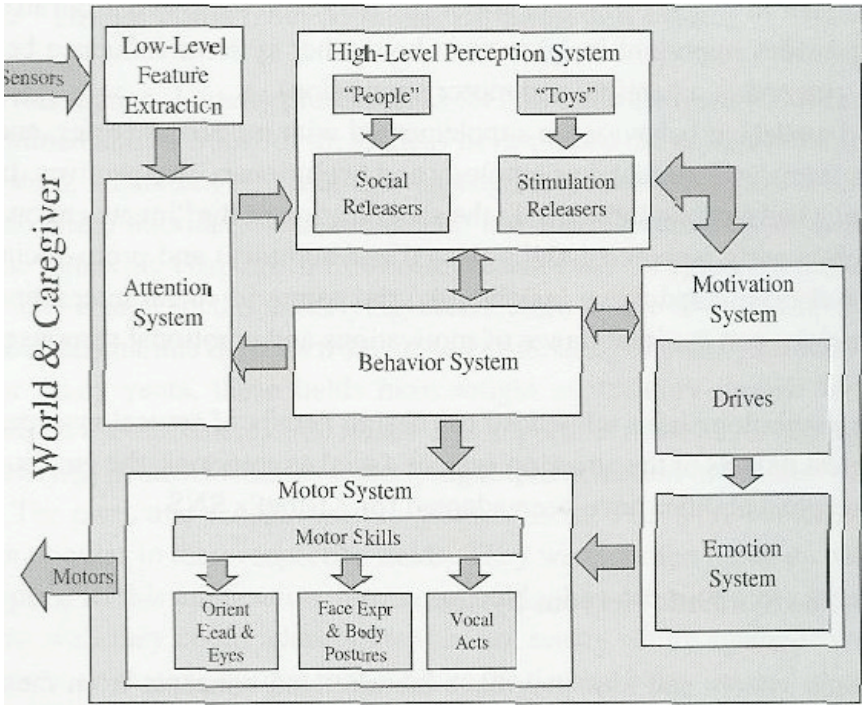
The fourth requirement indicates that the robot must behave autonomously to pursue its own motivation.

Also, there are more requirements for achieve the social robot as follows.

- Regulation of social interactions by intention manifestation
- Readable social cues
- Interpretation of human's social cues
- Competent behavior in a complex world:SSA
- Believable behaviors
  - ✧ convey intentionality, promote empathy, be expressive, and allow variability.

The first requirement indicates that the robot must express its intention to manage ongoing social interaction. The second indicates that it must generate social cues which humans easily understand. The third indicates that it must also recognize human's social cues. The fourth indicates that it must deal with complexity of the real world like SSA. The fifth indicates that it must generate believable behaviors.

6.5 Synthetic Nervous System



This subsection explains an architecture called a synthetic nervous system which Kismet employs. The above figure shows the structure of the architecture. I explains each module of the architecture in the remainder of the subsection.

The low-level feature extraction system detects visual and auditory cues by using several techniques of computer vision.

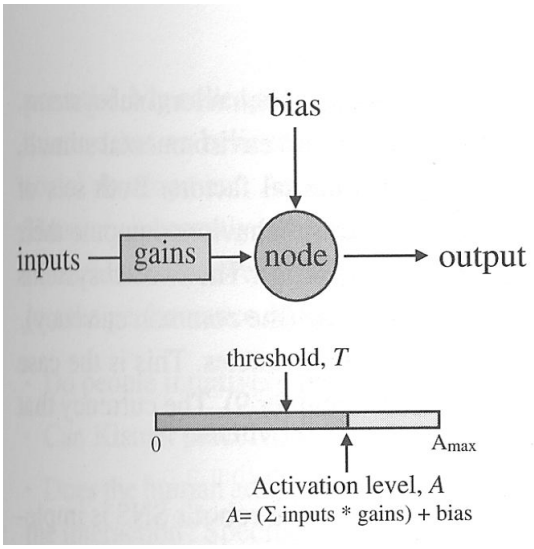
The attention system selects relevant sensor information from information extracted by the low-level feature extraction system.

The high-level perception system obtains relevant information for a behavior system from sensor information. People and toys correspond to the information for the behavior system.

The motivation system generates spontaneous motivations which have effects on the selection of Kismet behaviors. There are two types of motivations: Drive and Emotion. Drive gives motivation or needs to Kismet. Emotion promotes people's empathy for Kismet. People have the feeling when Kismet behaves emotionally. There are six basic emotions: anger, disgust, fear, joy, sorrow, and surprise. Also, it has arousal-based response: interest, calm, and boredom.

The behavior system consists of self-interested and goal directed entities. An arbitration mechanism selects some behaviors in response to the motivation and the environmental information of Kismet.

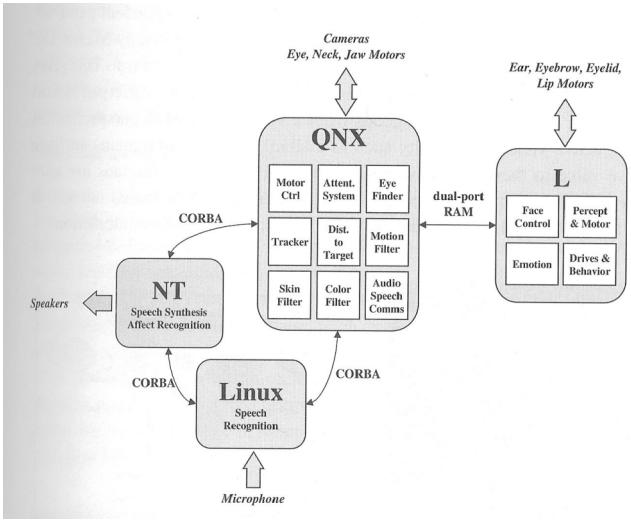
The motor system generates facial expressions, control its voice, and its head and eyes.

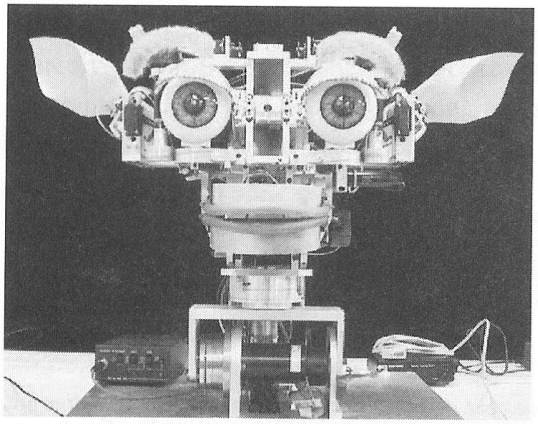


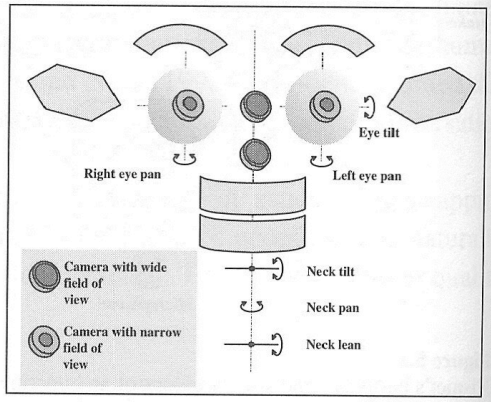
Computational components which the synthetic nervous system employs have a structure like the above figure. The node have a biased activation level which is a result of a summation of inputs from other nodes. The node generates an output when the activation level has a value beyond a threshold T. Each system consists of the network of the nodes.

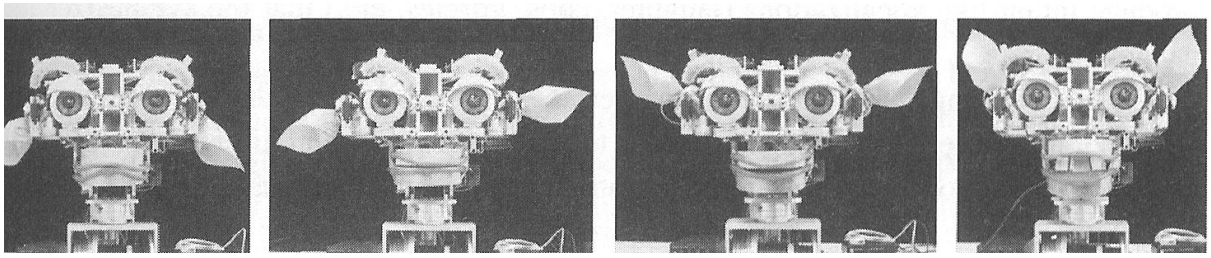
6.6 Hardware

This subsection explains hardware which Kismet uses. The figure in the next page shows computational structure of Kismet. It consists of four computers. They have different Operating System: L, QNX, NT, Linux. Each system of the synthetic nervous system is distributed on the computers. Refer to the figure to identify the exact









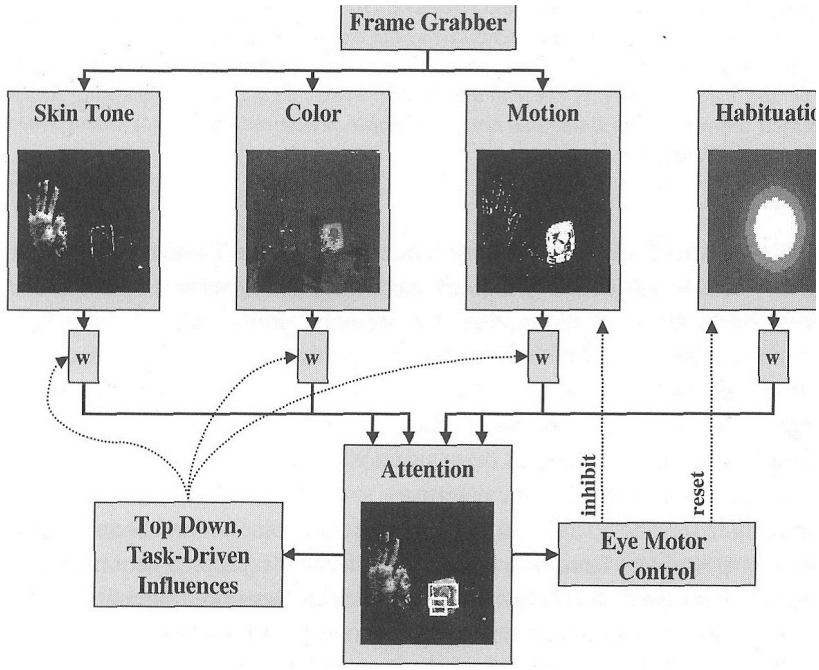
locations of each system.

The center picture and the figure show a hardware structure of Kismet. It has two 2DOF eyes, movable eyebrows, movable ears, movable mouth, and a 3DOF neck. Also, it has two types of camera. One captures an overall scene in front of Kismet. The other focuses on a relevant target in the overall scene.

The four pictures in the lower side of them show examples of facial expressions Kismet has done.

6.7 Vision System

The following figure shows a vision system Kismet uses. It corresponds to the



low-level feature extraction system in the synthetic nervous system.

The frame grabber captures an image from camera Kismet has. It analyzes what features the image includes by three independent modules. The first one detects the location of a skin color. The second finds an object by analyzing colors on the image. The third finds an object by analyzing motion on the image. The results of the analyses are given weights by the attention system, which corresponds to top-down attention. Then, the attention system focuses on some regions in the image. In the figure, it focuses on the face of a human.

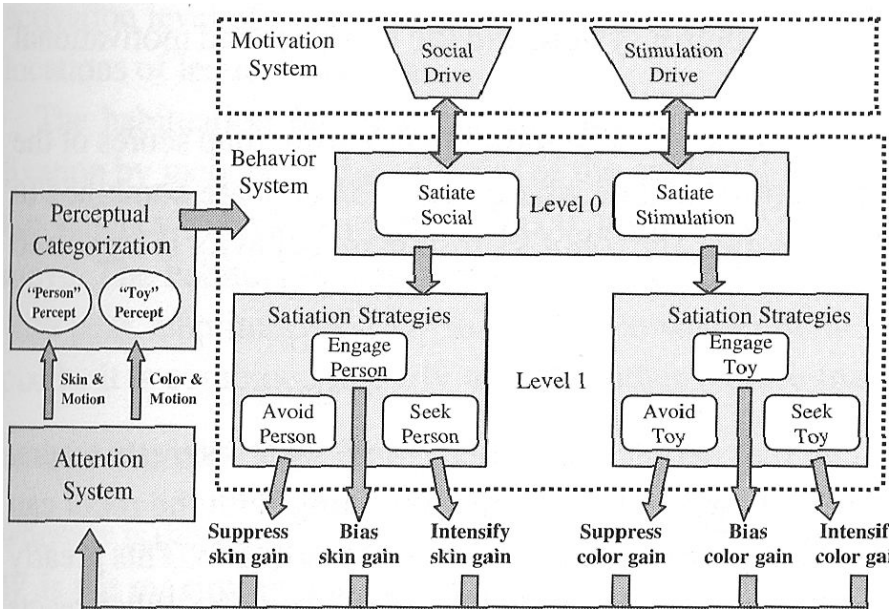
In addition, there is a habituation module. The module remembers the location the target of attention. This information is used when the frame grabber loses the image of the target. The target is sometime lost by noise or changes in luminance of an environment. At the situation, the reason why camera cannot capture the target is because there are some errors in capturing it. The habituation module compensates for the errors by indicating the existence of the target virtually.

The eye motor control controls the direction of Kismet's camera while generating two types of signals. One is an inhibition of the motion module. The aim of the inhibition is to prevent the motion module from working because it cannot identify any motions while the camera moves. The other is a reset signal toward the habituation module. The reset signal prevents the habituation module from outputting a ghost in an image after the camera reaches a new direction.

The attention system works as a filter for the captured image. The picture in the next page shows the results of the filter. The left figure shows a situation where Kismet has

Seek-people Seek-toy

a motivation to seek a human. Based on the motivation, the attention system focuses on a visual target which has features of a human's face. In the right figure, Kismet has a motivation to seek a toy. So, it focuses on the features related to a toy.



This figure indicates how the target of attention is selected. Overall behaviors of Kismet are based on motivation which the motivation system generates. The behavior system selects a policy to interact with a target based on the motivation. For example, if it has a motivation to keep the interaction with a human, the behavior system select a behavior named "engage person." The behavior gives the attention system a command to track the location of humans. On the other hand, if it is bored to interaction with him/her, it selects a behavior named "avoid person." The behavior

stops for the attention system to tack him/her.

The visual features selected by the attention system is given to the perceptual categorization which is done by the high-level perception system. Based on the given features, it identifies a human and a toy. Since the attention system selects the visual feature prior to the perceptual categorization, the identification is done under the control of motivation.

At last, the identified target is given to the behavior system and the selected behavior also induces changes in the motivation.

Kismet can recognize the gaze direction of humans. Since it employs easy way for the recognition, the accuracy of the detection is lower than that of a system which actually recognize the direction of eyes. The recognition method Kismet employs is to find a center position between two eyes. Also, it detects the region of a face. By identifying where the center position is in the face region, it extracts the direction of a human's gaze.

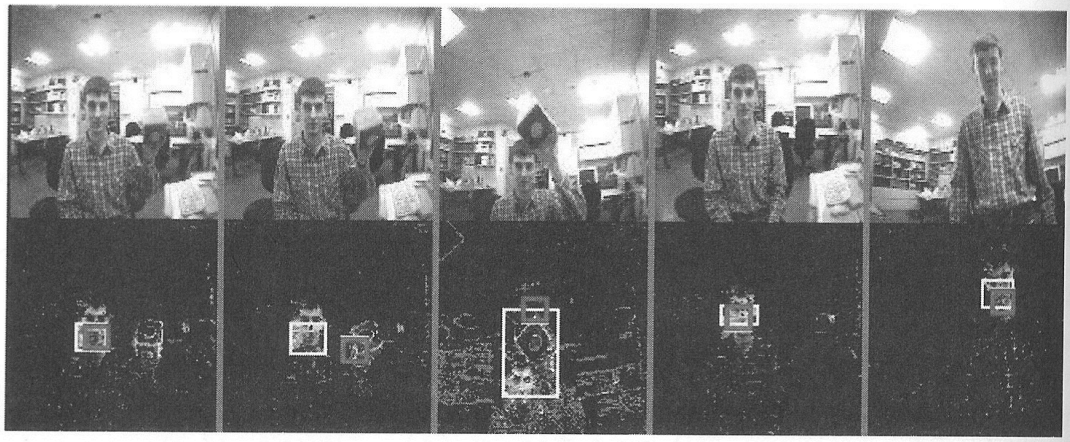


Figure 6.9

The figure in the previous page shows an example of attention shifts. At first, Kismet recognizes the location of a human's face and a toy. Also, the attention system focuses on the toy. The second picture shows that a grey rectangle moves with the toy. The grey rectangle denotes a target of the attention. The third picture shows that humans move the toy in front of him. At the situation, the attention of Kismet shifts from the toy to the human. After that, it tracks the location of the human.

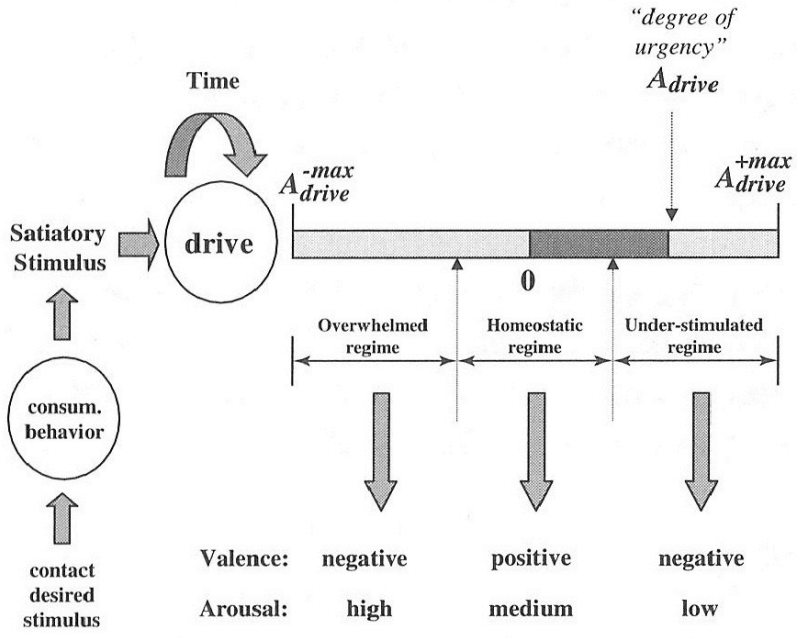
6.8 Motivation System

The motivation system consists of two types of motivations: homeostatic regulation (恒常性制御) and emotion. The homeostatic regulation is called drive in the motivation system. The emotion is called just emotion.

The homeostatic regulation maintains critical parameters. For example, temperature, energy level, and amount of fluids are the targets of human's homeostatic regulation. Kismet also has the same parameters to be maintained. If the parameters are not satisfied, the lack of them gives motivations to the robot as humans have a motivation to eat something if they are hungry.

The emotion has an effect on selecting behaviors of Kismet. For example, if it has a fear emotion, it selects a behavior for escaping from a human.

6.9 *Drive*

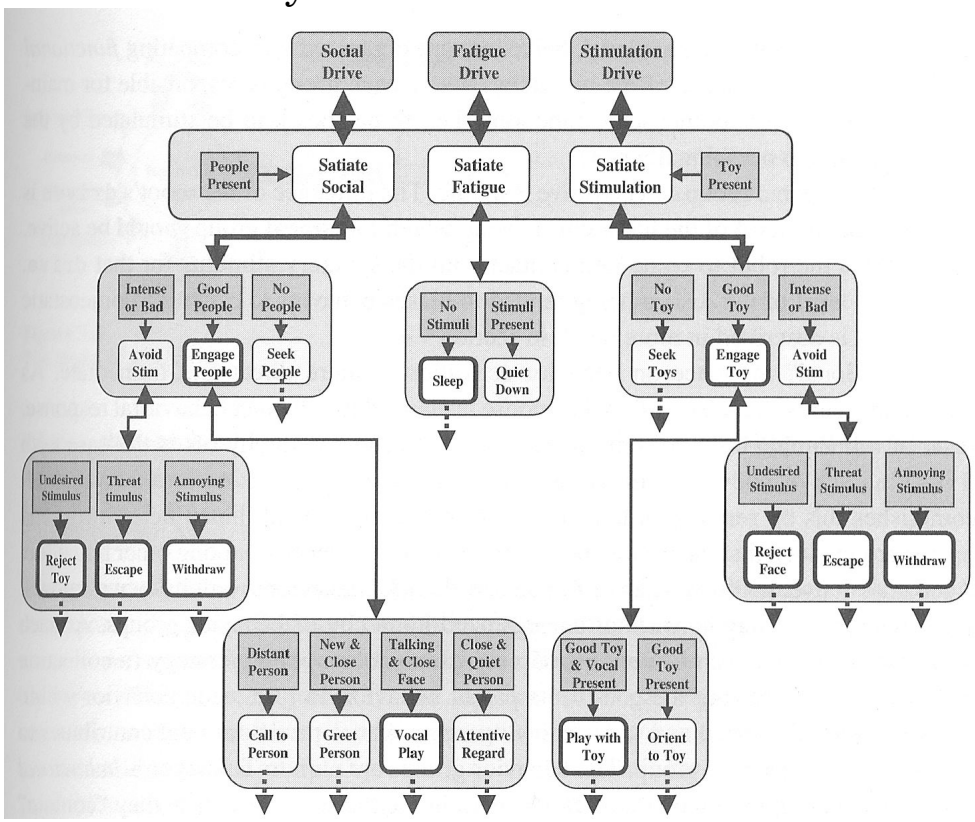


Drive consists of nodes like the above figure. The node has a drive value from a minus value to a plus value. The scale of the drive value is divided into three parts: an overwhelmed regime, a homeostatic regime, and an under-stimulated regime.

The drive value changes in the course of time evolution. When the value is in the under-stimulated regime, it activates behaviors to achieve the motivation of the drive. After desired stimulus is satisfied, the stationary stimulus informs the drive of the satisfactory stimulus. Then, the drive value takes a value in the homeostatic regime. The homeostatic regime activates behaviors to maintain the desired stimulus. The value decreases while it is getting the desired stimulus. At last, the value goes into the overwhelmed regime. The overwhelmed regime expresses the state of boredom and activates behaviors to escape from the stimulus.

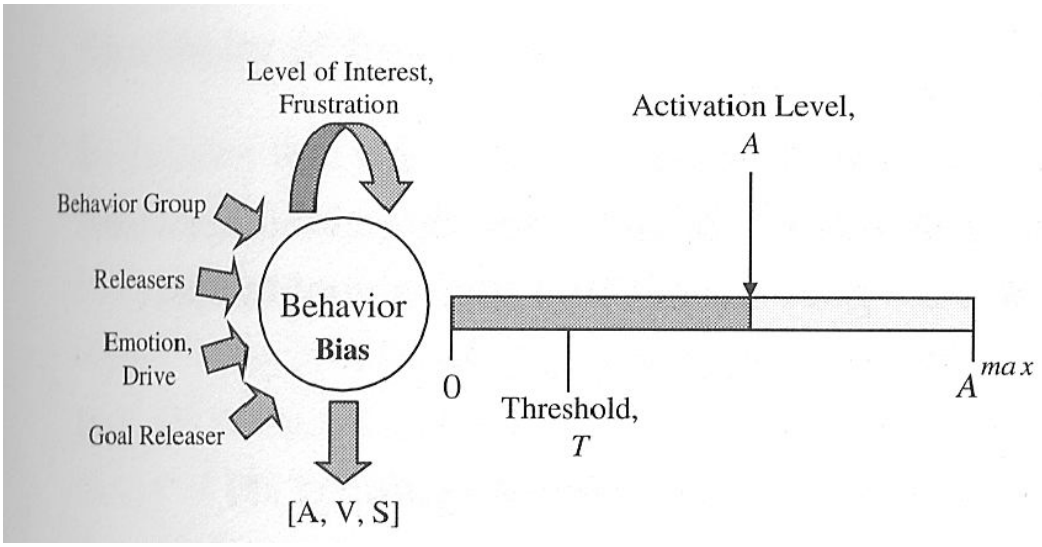
Kismet has three types of drives: a social drive, a stimulation drive, and a fatigue drive. The three regions for the value of the social drive has are finding person (the under-stimulated regime), continuing interaction (the homeostatic regime), and avoiding eye-contact (the overwhelmed regime). The value of the stimulation drive are categorized as turning its head (the under-stimulated regime), fixing its head(the homeostatic regime), boring (the overwhelmed regime). The value of the fatigue drive are awaking (the under-stimulated regime), awaking (the homeostatic regime), sleeping (the overwhelmed regime).

6.10 Behavior System

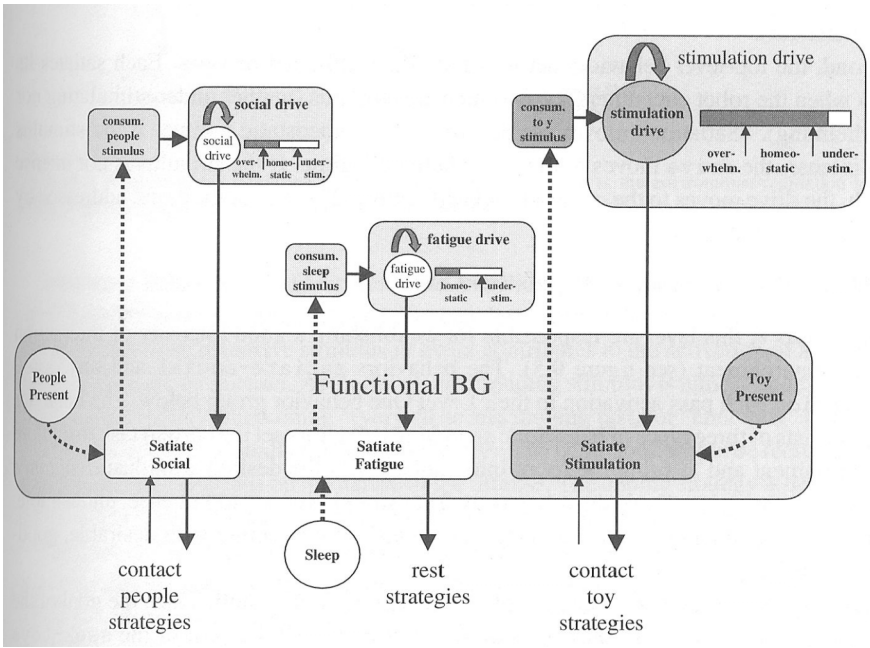


The above figure shows the behavior system and the motivation system. The top layer of the figure indicates the motivation system. There are three layered behavior system

under the motivation system. The value of the drive has an effect on behaviors in the first layer. The motor command is generated through these layers.

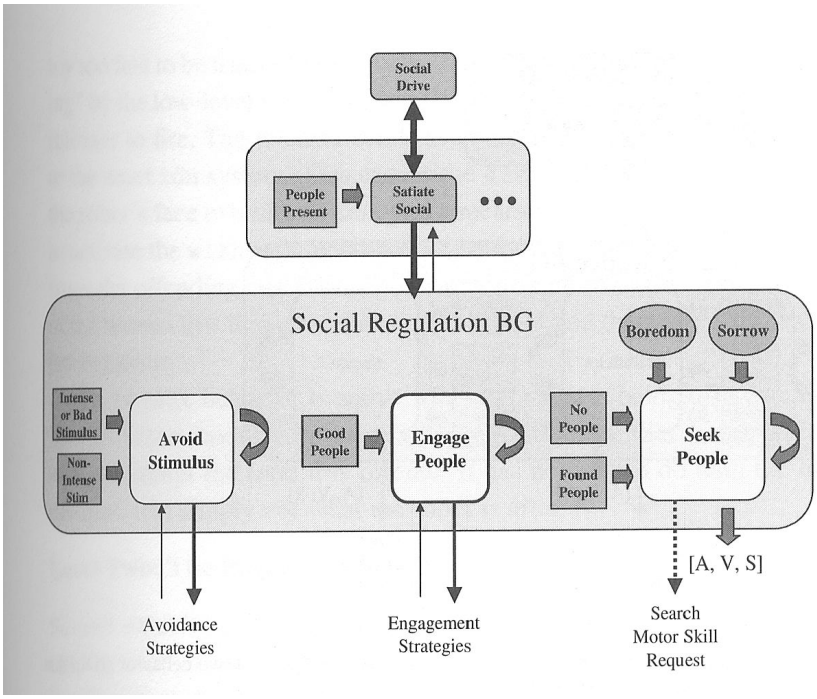


This figure shows a node of a behavior. It gets information from the other behavior group, releasers, the emotion, the drive, and a goal releaser. Based on the information, it generates an activation level. If the level is beyond the threshold, the behavior is activated.



I explain the detail of the behavior system along each layer. The above figure indicates the behavior system of level zero. In the figure, the stimulation drive has a value of an under-stimulated regime, which means that Kismet want to interact with a toy. The satiate stimulation behavior is activated by the drive and tries to find a toy.

The figure in the next page shows behaviors in the level one. In the figure, the social drive is activating the satiate social behavior, which means that Kismet want to



interaction with a human. Then, the seek people behavior is selected in the level one. The behavior changes its action in response to the existence of a human (no people or found people) and emotion (boredom or sorrow). And it generates a request to search motor skill in the motor system.

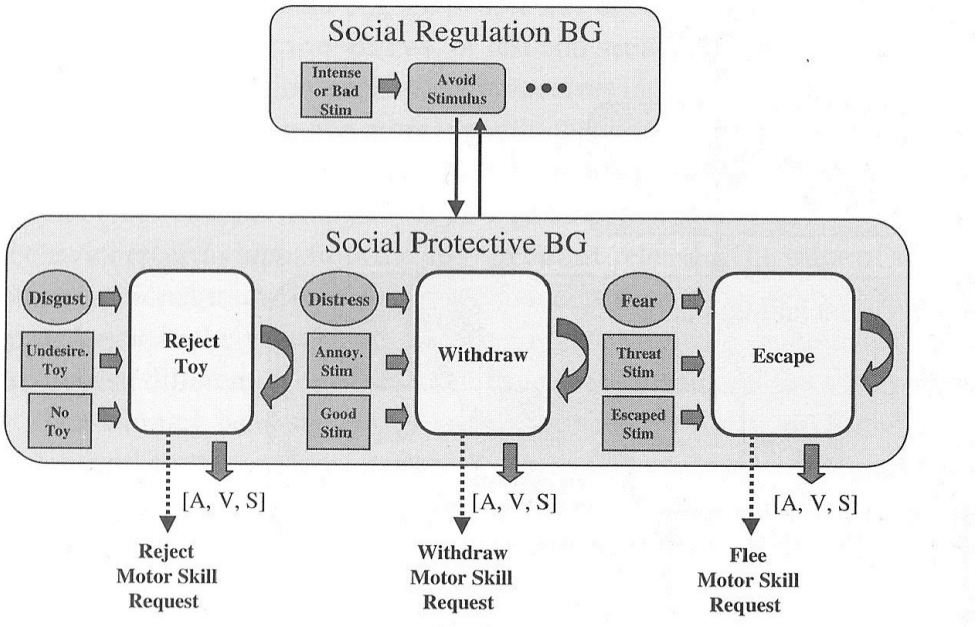


Figure 9.6

This figure shows an example of level two. In the figure, the avoid stimulus behavior in the level one is activated, which means that Kismet does not interact with a toy much more. In response to the behavior, social protective behavior group in the level

two is activated. There are three types of behaviors: a reject toy behavior, a withdraw behavior, and an escape behavior. What behavior is activated is depending on the state of emotion, a category of stimulus.

6.11 Overall Interaction

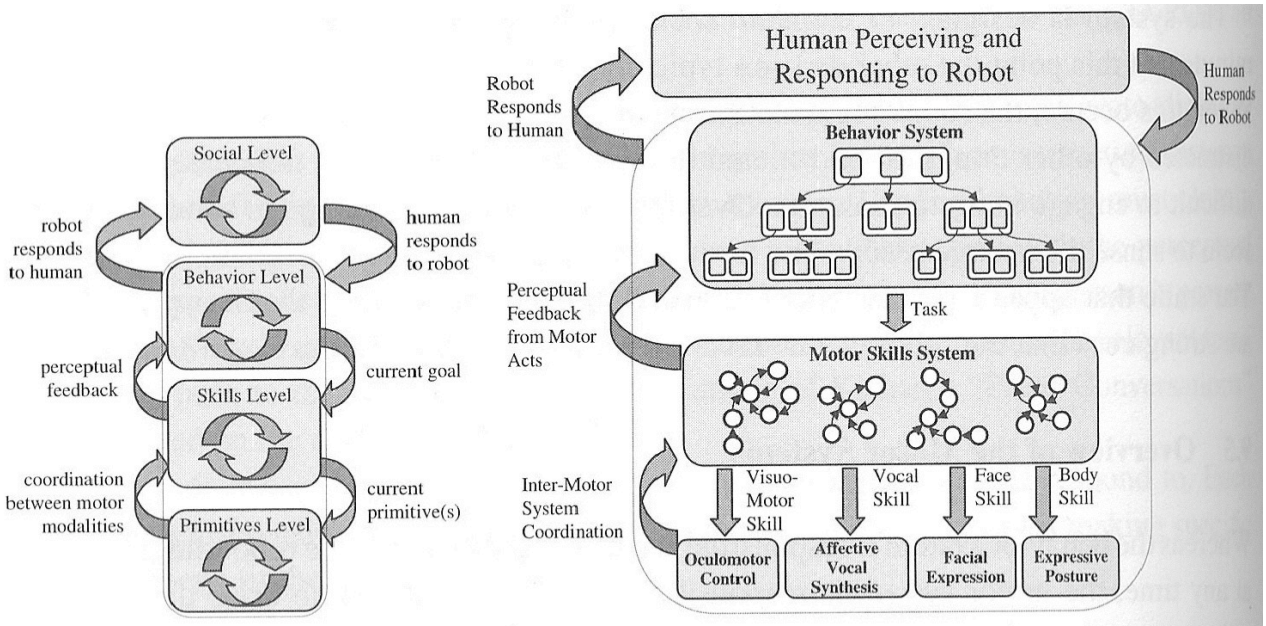


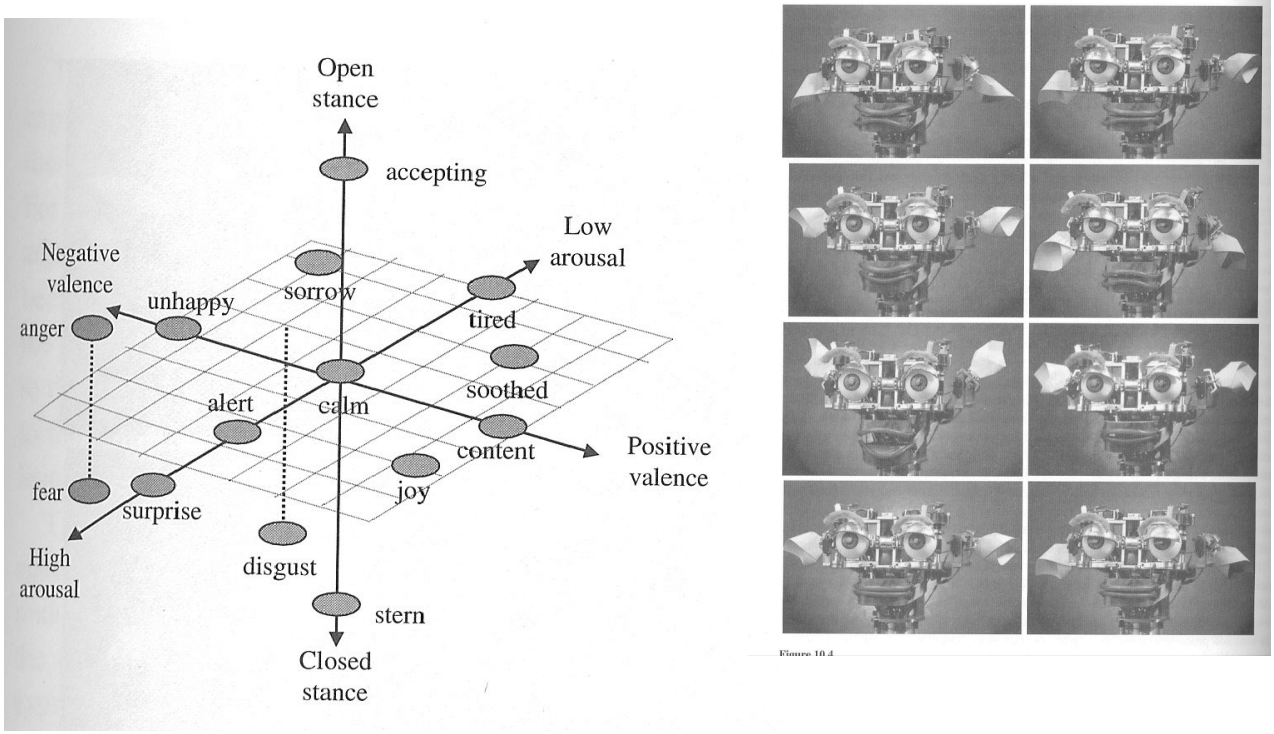
Figure 9.7

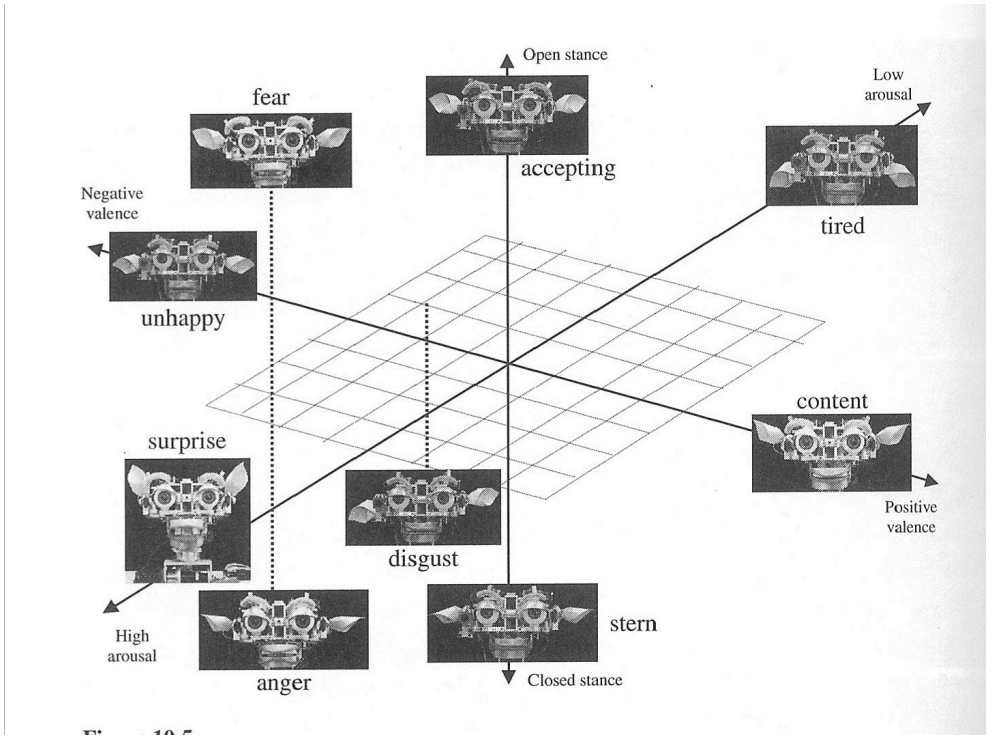
The figure shows the overall interaction inside Kismet. The interaction between a human and Kismet consists of several interactions between modules while each module also has an internal interaction. A human has an internal interaction of social level, the behavior system has the one of behavior level, the motor skill system has the one of skills level, and the hardware has the one of primitive level. Moreover, interaction between the primitive level, the skills level, and the behavior level becomes Kismet's responds to the human and his/her behaviors have effects on the overall interaction.

The interactions between the levels have interesting feature. The communications from behavior level to primitives level through skills level correspond to giving a command or a task. On the other hand, the communications of the opposite direction gives constraints on selecting the command or the task. The limitation of hardware which comes from the primitive level restricts the freedoms of the motor skill. And the action selected by the motor skill restricts the target of recognition which used in the behavior system. For example, if Kismet turns its head to the left direction, it cannot find an object in the right direction.

Also, the interaction between a human and Kismet also has the same feature. Kismet's vague actions like an infants does not have any sense in its alone. However, if humans behave socially in response to the actions, the actions make sense for him/her or other observers. Moreover, the human's behavior gives effects on selecting Kismet's behavior. In other words, their behaviors are connected each other heavily.

6.12 Emotion Expression





The picture and figures in the previous page show Kismet's emotional expressions. The emotion of Kismet is mapped on the 3D coordinate. One axis denotes the arousal of Kismet. The other denotes the positive or negative valence. The vertical one denotes the open or close stance of Kismet. 14 emotions are arranged on the map.

The pictures in the right side of Kismet show examples of Kismet's facial expressions and the figure including Kismet's face shows what facial expressions are used in the 3D coordinate.

6.13 The Sociality of Robots

The crucial point of social robots is to use appropriate social cues in the appropriate situations. Also, emotional expressions are significant to induce human's social response to the robot.

- Gaze expression to manifest curious stimulus or persons
- Facial expression or posture expression to express state of the curious about environmental input and to induce human's social response.

7. Dual Dynamics Model

I omit the contents of this topic from the handout of this year. If there is a remaining time, I will explain it with ppt. Please refer to the explanation.

8. Cognitive entrainment

I omit the contents of this topic from the handout of this year. If there is a remaining time, I will explain it with ppt. Please refer to the explanation.

9. Human Robot Interaction in Physical World

This section summarizes this handout. I explained the primary point to develop a communication system in the physical world.

Generating situated behavior is important to deal with dynamics in the real world. I introduced Subsumption Architecture, ANA, and several examples related to the behaviors.

Joint attention is a most important cognitive phenomenon to establish a communication in the real world. The study on infant's development and several psychological experiments indicated the significance of eye-contact and gaze behaviors. Moreover, social behaviors induce human's social response to robots. The social response also results in the establishment of joint attention.

Synchronizing behaviors is also important to achieve communication between people

and a system in the real world. Synergetics and Dual Dynamics Model dealt with the difficulty from a viewpoint of physics. By using Adiabatic approximation, many independent entities construct an ordered system. Dual Dynamics Model have shown that the physical phenomenon also occurs in human-machine interaction. Cognitive entrainment have shown the significance of the synchronizing behaviors when a human and a robot communication.