A Theoretical Consideration on Robotic Imitation of Human Action According to Demonstration plus Suggestion

Masao Yokota [1]

Abstract. The Mental Image Directed Semantic Theory (MIDST) has proposed an omnisensory mental image model and its description language $\boldsymbol{L_{md}}$ intended to facilitate intuitive human-system interaction such that happens between non-expert people and home robots. The most remarkable feature of $\boldsymbol{L_{md}}$ is its capability of formalizing both temporal and spatial event concepts on the level of human/robotic sensations. This paper presents a brief sketch of $\boldsymbol{L_{md}}$ and a theoretical consideration on robotic imitation of human action driven by human suggestion interpreted in $\boldsymbol{L_{md}}$, controlling the robotic attention mechanism efficiently.

1 INTRODUCTION

Robotic or artificial imitation is one kind of machine learning on human actions and there have been reported a considerable number of studies on imitation learning from human actions demonstrated without any verbal hint [e.g., 1-3]. In this case, it is extremely difficult for a robot to understand which part of human demonstration is significant or not because there are too many things to attend to as it is. That is, it is an important issue where the attention of the observer should be focused on when a demonstrator performs an action. Whereas there have been several proposals to control attention mechanisms efficiently in such top-down ways as guided by the prediction or strategy based on sensory data and knowledge of goals or tasks [e.g., 4, 5, 14], they are not realistic when a large number of actions must be imitated distinctively with various speeds, directions, trajectories, etc.

The author has been working on integrated multimedia understanding for intuitive human-robot interaction, that is, interaction between non-expert or ordinary people and home robots, where natural language is the leading information medium for their intuitive communication [6, 12]. For ordinary people, natural language is the most important because it can convey the exact intention of the sender to the receiver due to its syntax and semantics common to its users, which is not necessarily the case for another medium such as gesture or so. Therefore, the author believes that it is most desirable to realize robotic imitation aided by human verbal suggestion where robotic attention to human demonstration is efficiently controllable based on semantic understanding of the suggestion.

For such a purpose, it is essential to develop a systematically computable knowledge representation language (KRL) as well as representation-free technologies such as neural networks for processing unstructured sensory/motory data. This type of language is indispensable to *knowledge-based* processing such as *understanding* sensory events, *planning* appropriate actions and *knowledgeable* communication with ordinary people in natural language, and therefore it needs to have at least a good capability of representing spatiotemporal events that correspond to humans'/robots' sensations and actions in the real world.

Most of conventional methods have provided robotic systems with such quasi-natural language expressions as 'move(*Velocity*, *Distance*, *Direction*)', 'find(*Object*, *Shape*, *Color*)' and so on for human instruction or suggestion, uniquely related to computer programs to deploy sensors/ motors [e.g., 7, 8]. In association with robotic imitation intended here, however, these expression schemas are too linguistic or coarse to represent and compute sensory/motory events in an integrated way.

The Mental Image Directed Semantic Theory (MIDST) [9] has proposed a model of human attention-guided perception yielding omnisensory images that inevitably reflect certain movements of the focus of attention of the observer (FAO) scanning certain matters in the world. More analytically, these omnisensory images are associated with spatiotemporal changes (or constancies) in certain attributes of the matters scanned by FAO and modeled as temporally parameterized "loci in attribute spaces", so called, to be formulated in a formal language $\boldsymbol{L_{md}}$. This language has already been implemented on several types of computerized intelligent systems [e.g., 10, 12].

This paper presents a brief sketch of the formal language $\boldsymbol{L_{md}}$ and a theoretical consideration on robotic imitation of human demonstrated action aided by human suggestion interpreted as semantic expression in $\boldsymbol{L_{md}}$. The most remarkable feature of $\boldsymbol{L_{md}}$ is its capability of formalizing spatiotemporal matter concepts grounded in human/robotic sensation while the other similar KRLs are designed to describe the logical relations among conceptual primitives represented by lexical tokens [e.g., 11]. In $\boldsymbol{L_{md}}$ expression are hinted what and how should be attended to in human action as analogy of human FAO movement and thereby the robotic attention can be controlled in a top-down way.

2 A BRIEF SKETCH OF $\boldsymbol{L_{md}}$

An attribute space corresponds with a certain measuring instrument just like a barometer, thermometer or so and the loci represent the movements of its indicator. For example, the moving black triangular object shown in Figure 1 is assumed to be perceived as the loci in the three attribute spaces, namely, those of 'Location', 'Color' and 'Shape' in the observer's brain.

[1] Fukuoka Institute of Technology, Japan, email: yokota@fit.ac.jp

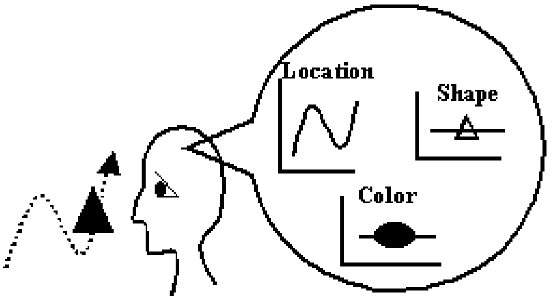


Figure1. Mental image model

Such a locus is to be articulated by “Atomic Locus” with an ***absolute*** time-interval $[t_i, t_f]$ $(t_i< t_f)$ as depicted in Figure 2 (up) and formulated as (1).

$$L(x,y,p,q,a,g,k) \qquad (1)$$

This formula is called ‘Atomic Locus Formula’ whose first two arguments are often referred to as ‘Event Causer (EC)’ and ‘Attribute Carrier (AC)’, respectively. A logical combination of atomic locus formulas defined as a well-formed formula (i.e., wff) in predicate logic is called simply ‘Locus Formula’. The intuitive interpretation of (1) is given as follows, where ‘matter’ refers to ‘object’ or ‘event’ largely.

> ***“Matter ‘x’ causes Attribute ‘a’ of Matter ‘y’ to keep (p=q) or change ($p \neq q$) its values temporally ($g=G_t$) or spatially ($g=G_s$) over a time-interval, where the values ‘p’ and ‘q’ are relative to the standard ‘k’.”***

When $g=G_t$ and $g=G_s$, the locus indicates monotonic change or constancy of the attribute in time domain and that in space domain, respectively. The former is called ‘temporal event’ and the latter, ‘spatial event’. For example, the motion of the ‘bus’ represented by S1 is a temporal event and the ranging or extension of the ‘road’ by S2 is a spatial event whose meanings or concepts are formulated as (2) and (3), respectively, where A_{12} denotes ‘Physical Location’. These two formulas are different only at ‘Event Type (i.e., *g*)’.

(S1) The bus runs from Tokyo to Osaka.

$$(\exists x,y,k)L(x,y,Tokyo,Osaka,A_{12},\mathbf{G_t},k)\wedge bus(y) \qquad (2)$$

(S2) The road runs from Tokyo to Osaka.

$$(\exists x,y,k)L(x,y,Tokyo,Osaka,A_{12},\mathbf{G_s},k)\wedge road(y) \qquad (3)$$

The author has hypothesized that the difference between temporal and spatial event concepts can be attributed to the relationship between the Attribute Carrier (AC) and the Focus of the Attention of the Observer (FAO) [9]. To be brief, it is assumed that the FAO is fixed on the whole AC in a temporal event but *runs* about on the AC in a spatial event. According to this assumption, as shown in Figure 3, the *bus* and the FAO move together in the case of S1 while the FAO solely moves along the *road* in the case of S2.

Any locus in a certain Attribute Space can be formalized as a combination of atomic locus formulas and, so called, tempo-logical connectives, among which the most frequently used are ‘Simultaneous AND (Π)’ and ‘Consecutive AND (•)’ as appear in the conceptual definition (4) of the English verb ‘fetch’ depicted in Figure 2 (down).

$$(\lambda x,y)fetch(x,y)\leftrightarrow(\lambda x,y)(\exists p_1,p_2,k)L(x,x,p_1,p_2,A_{12},G_t,k)\bullet ((L(x,x,p_2,p_1,A_{12},G_t,k)\Pi L(x,y,p_2,p_1,A_{12},G_t,k))\wedge x\neq y\wedge p_1\neq p_2 \qquad (4)$$

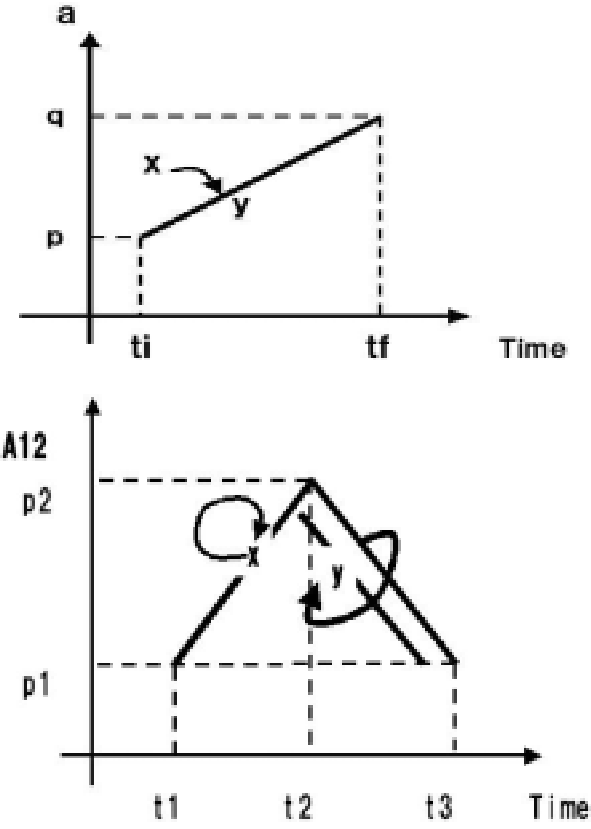


Figure 2. Atomic Locus (up) and Locus of ‘fetch’ (down)

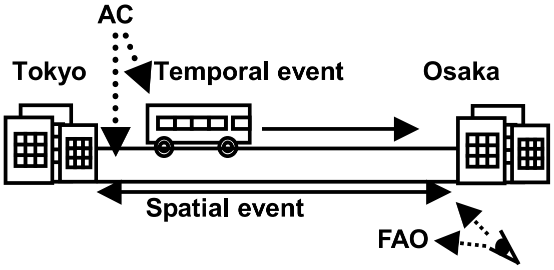


Figure 3. FAO movements and Event types

In order for explicit indication of time duration, ‘Empty Event (EE)’ denoted by ‘ε’ is introduced by the definition (5) with the attribute ‘Time Point (A34)’. According to this scheme, the duration $[t_a, t_b]$ of an arbitrary locus χ can be expressed as (6).

$$\varepsilon([t_1,t_2])\leftrightarrow(\exists x,y,g,k)\ L(x,y,t_1,t_2,A34,g,k) \qquad (5)$$

$$\chi\ \Pi\ \varepsilon([t_a, t_b]) \qquad (6)$$

All the same way, an object concept is also defined and expressed in $\boldsymbol{L_{md}}$ as a combination of potential events on its properties and its relations with others. For example, the conceptual descriptions of ‘rain’, ‘wind’ and ‘air’ can be given as (7)-(9), reading ‘Rain is water attracted from the sky by the earth, makes an object wetter, is pushed an umbrella to by a human,…,’ ‘Wind is air, affects the direction of rain,… ,’ and ‘Air has no shape, no taste, no vitality, …,’ respectively.

$$\begin{aligned}&(\lambda x)\mathrm{rain}(x)\leftrightarrow(\lambda x)(\exists x_1,x_2,\ldots)L(\_,x,x_1,x_1,A_{41},G_t,\_)\\&\Pi L(\mathrm{Earth},x,\mathrm{Sky},\mathrm{Earth},A_{12},G_t,\_)\Pi L(x,x_2,p,q,A_{25},G_t,\_)\\&\Pi L(x_3,x_4,x,x,A_{19},G_t,x_3)\wedge\mathrm{water}(x_1)\\&\wedge\mathrm{object}(x_2)\wedge\mathrm{human}(x_3)\wedge\mathrm{umbrella}(x_4)\wedge(p<q)\ldots\end{aligned}\tag{7}$$

$$\begin{aligned}&(\lambda x)\mathrm{wind}(x)\leftrightarrow(\lambda x)(\exists x_1,x_2,\ldots)L(\_,x,x_1,x_1,A_{41},G_t,\_)\\&\wedge\mathrm{air}(x_1)\wedge(L(x,x_2,p,q,A_{13},G_t,\_)\wedge\mathrm{rain}(x_2)\ldots\end{aligned}\tag{8}$$

$$\begin{aligned}&(\lambda x)\mathrm{air}(x)\leftrightarrow(\lambda x)(\ldots\wedge L^*(\_,x,/,/,A_{11},G_t,\_)\wedge\ldots\wedge\\&L^*(\_,x,/,/,A_{29},G_t,\_)\wedge\ldots\wedge L^*(\_,x,/,/,A_{39},G_t,\_)\wedge\ \ldots)\end{aligned}\tag{9}$$

Hereafter, for simplicity of ***L**_md_* expression, the special symbols '*', '_'and '/' are often employed to represent 'always', 'something (or some value)' and 'nothing (no value)' as defined by (10)-(12), respectively.

$$X^* \leftrightarrow (\forall[p,q])X\ \Pi\ \varepsilon([p,q]) \tag{10}$$

$$L(\ldots,\_,\ldots) \leftrightarrow (\exists\omega)L(\ldots,\omega,\ldots) \tag{11}$$

$$L(\ldots,/,\ldots) \leftrightarrow \sim(\exists p)\ L(\ldots,\omega,\ldots) \tag{12}$$

Table 1 shows about 50 attributes extracted exclusively from English and Japanese words of common use contained in certain thesauri [9]. Most of them (i.e., A01-A45) correspond to the sensory receptive fields in human brains. For example, those marked with '*' in this table can be associated to the sense 'sight'. Correspondingly, six categories of standards shown in Table 2 have been extracted that are necessary for representing relative values of each attribute in Table 1. ***These tables show that ordinary people live their casual lives, attending to tens of attributes of the matters in the world to cognize them in comparison with several kinds of standards.***

Table 1. List of attributes

Code	Attribute [Property†] (*words/phrases concerned*)
*A01	PLACE OF EXISTE NCE [N] (*happen, perish*)
*A02	LENGTH [S] (*long, shorten, close, away*)
*A03	HEIGHT [S] (*high, lower*)
*A04	WIDTH [S] (*widen, narrow*)
*A05	THICKNESS [S] (*thick, thin*)
*A06	DEPTH1 [S] (*deep, shallow*)
*A07	DEPTH2 [S] (*deep, concave*)
*A08	DIAMETER [S] (*across, in diameter*)
*A09	AREA [S] (*square meters, acre*)
*A10	VOLUME [S] (*litter, gallon*)
*A11	SHAPE [N] (*round, triangle*)
*A12	PHYSICAL LOCATION [N] (*move, stay*)
*A13	DIRECTION [N] (*turn, wind, left*)
*A14	ORIENTATION [N] (*orientate, command*)
*A15	TRAJECTORY [N] (*zigzag, circle*)
*A16	VELOCITY [S] (*fast, slow*)
*A17	MILEAGE [S] (*far, near*)
A18	STRENGTH OF EFFECT [S] (*strong, powerful*)
A19	DIRECTION OF EFFECT [N] (*pull, push*)
A20	DENSITY [S] (*dense, thin*)
A21	HARDNESS [S] (*hard, soft*)
A22	ELASTICITY [S] (*elastic, flexible*)
A23	TOUGHNESS [S] (*fragile, stiff*)
A24	TACTILE FEELING [S] (*rough, smooth*)
A25	HUMIDITY [S] (*wet, dry*)
A26	VISCOSITY [S] (*oily, watery*)
A27	WEIGHT [S] (*heavy, light*)
A28	TEMPERATURE [S] (*hot, cold*)
A29	TASTE [N] (*sour, sweet, bitter*)
A30	ODOUR [N] (*pungent, sweet*)
A31	SOUND [N] (*noisy, silent, loud*)
*A32	COLOR [N] (*red, white*)
A33	INTERNAL SENSATION [N] (*tired, hungry*)
A34	TIME POINT [S] (*o'clock, elapse*)
A35	DURATION [S] (*hour, minute, long, short*)
A36	NUMBER [S] (*ten, quantity, number*)
A37	ORDER [S] (*first, last*)
A38	FREQUENCY [S] (*sometimes, frequent*)
A39	VITALITY [S] (*alive, dead, vivid*)
A40	SEX [S] (*male, female*)
A41	QUALITY [N] (*make, destroy*)
A42	NAME [V] (*name, token*)
A43	CONCEPTUAL CATEGORY [V] (*mammal*)
*A44	TOPOLOGY [V] (*in, out, touch*)
*A45	ANGULARITY [S] (*sharp, dull, rectangle*)
B01	WORTH [N] (*improve, praise, deny, alright*)
B02	LOCATION OF INFORMATION [N] (*tell, hear*)
B03	EMOTION [N] (*like, hate*)
B04	BELIEF VALUE [S] (*believe, trust*)
	…………………………..

†S: scalar value, N: non-scalar value. *Attributes concerning the sense of sight.

Table 2. List of standards

Categories	Remarks
Rigid Standard	Objective standards such as denoted by measuring *units* (meter, gram, etc.).
Species Standard	The *attribute value ordinary* for a species. A *short train* is ordinarily longer than a *long pencil*.
Proportional Standard	'*Oblong*' means that the width is greater than the height at a physical object.
Individual Standard	*Much* money for one person can be too *little* for another.
Purposive Standard	One room large enough for a person's *sleeping* must be too small for his *jogging*.
Declarative Standard	The origin of an order such as 'next' must be declared explicitly just as 'next *to him*'.

3 INTELLIGENT SYSTEM IMAGES-M

3.1 System configuration

The intelligent system IMAGES-M [e.g., 10, 12] is assumed to be the main intelligence of the robot intended here. As shown in Figure 4, IMAGES-M is one kind of expert system equipped with five kinds of user interfaces for multimedia communication, that is, Sensory Data Processing Unit (SDPU), Speech Processing Unit (SPU), Picture Processing Unit (PPU), Text Processing Unit (TPU), and Action Data Processing Unit (ADPU) besides Inference Engine (IE) and Knowledge Base (KB). Each processing unit in collaboration with IE performs mutual conversion between each type of information medium and locus formulas.

IMAGES-M is a language-centered intelligent system in order to facilitate intuitive interaction between humans and robots. For comprehensible communication with humans, robots must understand natural language *semantically* and *pragmatically*. Here, as shown in Figure 5, semantic understanding means associating symbols to conceptual images of matters (i.e., objects or events), and pragmatic understanding means anchoring symbols to real matters by unifying conceptual images with perceptual images.

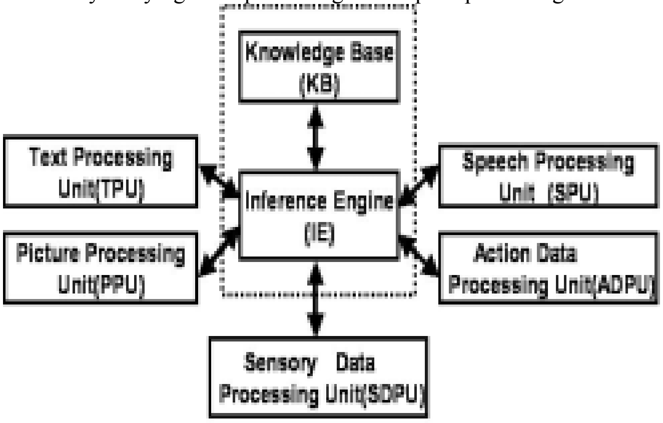

Knowledge Base (KB)
Text Processing Unit(TPU)
Inference Engine (IE)
Speech Processing Unit (SPU)
Picture Processing Unit(PPU)
Action Data Processing Unit(ADPU)
Sensory Data Processing Unit(SDPU)

Figure 4. Configuration of IMAGES-M

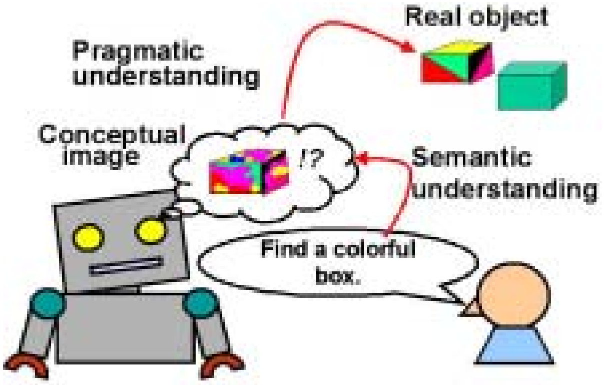


Figure 5. Semantic and pragmatic understanding

3.2 Semantic understanding

As shown in Figure 6, natural language expression (i.e, surface structure) and $\boldsymbol{L_{md}}$ expression (i.e., conceptual structure) are mutually translatable through surface dependency structure by utilizing syntactic rules and word meaning descriptions [9].

A word meaning description M_w is defined by (13) as a pair of 'Concept Part (C_p)' and 'Unification Part (U_p)'.

$$M_w \leftrightarrow [C_p{:}U_p] \qquad (13)$$

The C_p of a word W is a locus formula about properties and relations of the matters involved such as shapes, colors, functions, potentialities, etc while its U_p is a set of operations for unifying the C_ps of W's syntactic governors or dependents. For example, the meaning of the English verb 'carry' can be given by (14).

$$[(\exists x,y,p_1,p_2)\ L(x,x,p_1,p_2,A12,Gt,\_)\Pi$$
$$L(x,y,p_1,p_2,A12,Gt,\_)\wedge x\neq y\wedge p_1\neq p_2{:}ARG(Dep.1,x);$$
$$ARG(Dep.2,y);] \qquad (14)$$

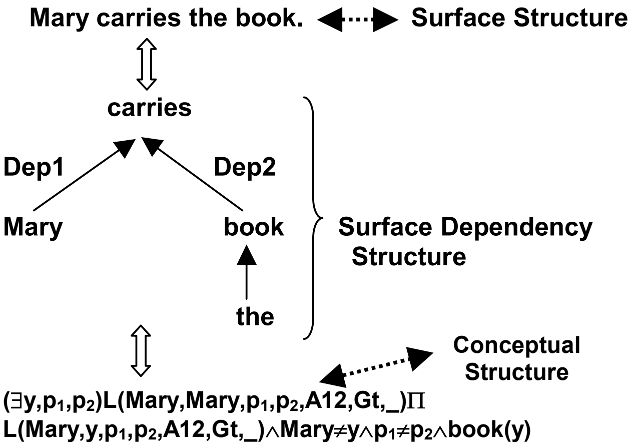

Mary carries the book.
Surface Structure
carries
Dep1
Dep2
Mary
book
the
Surface Dependency Structure
Conceptual Structure
$(\exists y,p_1,p_2)L(Mary,Mary,p_1,p_2,A12,Gt,\_)\Pi$
$L(Mary,y,p_1,p_2,A12,Gt,\_)\wedge Mary\neq y\wedge p_1\neq p_2\wedge book(y)$

Figure 6. Mutual conversion between natural language and $\boldsymbol{L_{md}}$

(Input)

With the long red stick Tom precedes Jim.

(Output)

Tom with the long red stick goes before Jim goes.
Jim goes after Tom goes with the long red stick.
Jim follows Tom with the long red stick.
Tom carries the long red stick before Jim goes.
…………………

Figure 7. Paraphrasing as semantic understanding by IMAGES-M

The U_p above consists of two operations to unify the first dependent (Dep.1) and the second dependent (Dep.2) of the current word with the variables *x* and *y*, respectively. Here, Dep.1 and Dep.2 are the 'subject' and the 'object' of 'carry', respectively. Therefore, the surface structure '*Mary carries a book*' is translated into the conceptual structure (15) via the surface dependency structure shown in Figure 6.

$$(\exists y,p_1,p_2)L(Mary,Mary,p_1,p_2,A12,Gt,\_)\Pi$$
$$L(Mary,y,p_1,p_2,A12,Gt,\_)\wedge Mary\neq y\wedge p_1\neq p_2\wedge book(y) \qquad (15)$$

For another example, the meaning description of the English preposition 'through' is also given by (16).

$$[(\exists x,y,p_1,z,p_3,g,p_4)(\underline{L(x,y,p_1,z,A12,g,\_)}\bullet$$
$$L(x,y,z,p_3,A12,g,\_))\Pi\ L(x,y,p_4,p_4,A13,g,\_)\wedge p_1\neq z\wedge z\neq p_3$$
$${:}ARG(Dep.1,z);\ IF(Gov=Verb)\rightarrow PAT(Gov,(1,1));$$
$$IF(Gov=Noun)\rightarrow ARG(Gov,y);] \qquad (16)$$

The U_p above is for unifying the C_ps of the very word, its governor (Gov, a verb or a noun) and its dependent (Dep.1, a noun). The second argument (1,1) of the command PAT indicates the underlined part of (13) and in general (*i,j*) refers to the partial formula covering from the *i*th to the *j*th atomic formula of the current C_p. This part is the pattern common to both the C_ps to be unified. This is called 'Unification Handle (U_h)' and when missing, the C_ps are to be combined simply with '$\wedge$'.

Therefore the sentences S3, S4 and S5 are interpreted as (17)-(19), respectively. The underlined parts of these formulas are the results of PAT operations. The expression (20) is the C_p of the adjective 'long' implying 'there is some value greater than some standard of 'Length (A02)' which is often simplified as (20').

(S3) The train runs through the tunnel.

$(\exists x,y,p_1,z,p_3,p_4)(\underline{L(x,y,p_1,z,A12,Gt,\_)}\bullet$
$L(x,y,z,p_3,A12,Gt,\_))\Pi\ L(x,y,p_4,p_4,A13,Gt,\_)$
$\wedge p_1\neq z \wedge z\neq p_3 \wedge train(y) \wedge tunnel(z)$ (17)

(S4) The path runs through the forest.

$(\exists x,y,p_1,z,p_3,p_4)(\underline{L(x,y,p_1,z,A12,Gs,\_)}\bullet$
$L(x,y,z,p_3,A12,Gs,\_))\Pi\ L(x,y,p_4,p_4,A13,Gs,\_)$
$\wedge p_1\neq z \wedge z\neq p_3 \wedge path(y) \wedge forest(z)$ (18)

(S5) The path through the forest is long.

$(\exists x,y,p_1,z,p_3,x_1,q,p_4,k_1)$
$(L(x,y,p_1,z,A12,Gs,\_)\bullet L(x,y,z,p_3,A12,Gs,\_))$
$\Pi\ L(x,y,p_4,p_4,A13,Gs,\_) \wedge L(x_1,y,q,q,A02,Gt,k_1)$
$\wedge p_1\neq z\wedge z\neq p_3\wedge q>k_1\wedge path(y)\wedge forest(z)$ (19)
$(\exists x_1,y_1,q,k_1)L(x_1,y_1,q,q,A02,Gt,k_1)\wedge q>k_1$ (20)
$(\exists x_1,y_1,k_1)L(x_1,y_1,Long,Long,A02,Gt,k_1)$ (20')

The process above is completely reversible except that multiple natural expressions as paraphrases can be generated by TPU in IMAGES-M as shown in Figure 7 because such event patterns as shown in Figure 2 are sharable among multiple word concepts. This is one of the most remarkable features of MIDST and is also possible between different languages as understanding-based translation [10, 12].

3.3 Pragmatic understanding

An event expressed in L_{md} is compared to a movie film recorded through a floating camera because it is necessarily grounded in FAO's movement over the event. For example, it is not the 'path' but the 'FAO' that 'sinks' in S6 or 'rises' in S7. Therefore, such expressions refer to the same scene pragmatically in spite of their appearances, whose semantic descriptions are given as (21) and (22), respectively, where 'A_{13}', '↑' and '↓' refer to the attribute 'Direction', and its values 'upward' and 'downward', respectively. This fact is generalized as '***Postulate of Reversibility of a Spatial Event*** (PRS)' belonging to people's intuitive knowledge about geography, and the conceptual descriptions (21) and (22) are called ***equivalent in the PRS***.

(S6) The path sinks to the brook.

$(\exists x,y,p,z)L(x,y,p,z,A_{12},G_s,\_)\Pi L(x,y,\downarrow,\downarrow,A_{13},G_s,\_)$
$\wedge path(y) \wedge brook(z) \wedge p\neq z$ (21)

(S7) The path rises from the brook.

$(\exists x,y,p,z)L(x,y,z,p,A_{12},G_s,\_)\Pi L(x,y,\uparrow,\uparrow,A_{13},G_s,k_2)$
$\wedge path(y) \wedge brook(z) \wedge p\neq z$ (22)

For another example of spatial event, Figure 8 (up) concerns human perception of the formation of multiple distinct objects, where FAO runs along an imaginary object so called 'Imaginary Space Region (ISR)'. This spatial event can be verbalized as S8 using the preposition 'between' and formulated as (22), corresponding also to such concepts as 'row', 'line-up', etc. Any type of topological relation between two objects is also to be formulated by employing an ISR. For example, S9 is translated into (23) or (23'), where '*In*', and '*Cont*' are the values 'inside' and 'contains' of the attribute 'Topology (A44)' represented by 3x3 matrices at the Sandard of '9-intersection model (*IM*)' [13], where '*In'* and *'Cont'* are the transposes each other.

(S8) □ is between Δ and ○.

$(\exists y,p)(L(\_,y,\Delta,\square,A_{12},G_s,\_)\bullet L(\_,y,\square,\circ,A_{12},G_s,\_))\Pi$
$L(\_,y,p,p,A_{13},G_s,\_) \wedge ISR(y)$ (22)

(S9) □ is in the room.

$(\exists x,y)L(\_,x,y,\square,A12,Gs,\_)\Pi L(\_,x,\mathit{In},\mathit{In},A44,Gt,\mathit{IM})$
$\wedge ISR(x)\wedge room(y)$ (23)
$(\exists x,y)L(\_,x,\square,y,A12,Gs,\_)\Pi L(\_,x,\mathit{Cont},\mathit{Cont},A44,Gt,\mathit{IM})$
$\wedge ISR(x)\wedge room(y)$ (23')

For more complicated examples, consider S10 and S11. The underlined parts are deemed to refer to some events neglected in time and in space, respectively. These events correspond with skipping of FAOs and are called 'Temporal Empty Event' and 'Spatial Empty Event', denoted by 'ε_t' and 'ε_s' as Empty Events with $g=G_t$ and $g=G_s$ at (5), respectively. Their concepts are described as (24) and (25), where 'A_{15}' and 'A_{17}' represent the attribute 'Trajectory' and 'Mileage', respectively. From the viewpoint of pragmatic understanding, the formula (25) can refer to such a spatial event depicted as the still picture in Figure 8 (down) while (24), a temporal event to be recorded as a movie.

(S10) The *bus* runs 10km straight east from A to B, and *after a while*, at C it meets the street with the sidewalk.

$(\exists x,y,z,p,q)(L(\_,x,A,B,A_{12},G_{t,}\_)\Pi$
$L(\_,x,0,10km,A_{17},G_{t,}\_)\Pi L(\_,x,Point,Line,A_{15},G_{t,}\_)\Pi$
$L(\_,x,East,East,A_{13},G_{t,}\_))\bullet\varepsilon_t\bullet(L(\_,x,p,C,A_{12},G_{t,}\_)$
$\Pi L(\_,y,q,C,A_{12},G_{s,}\_)\Pi L(\_,z,y,y,A_{12},G_{s,}\_))$
$\wedge bus(x)\wedge street(y)\wedge sidewalk(z)\wedge p\neq q$ (24)

(S11) The *road* runs 10km straight east from A to B, and *after a while*, at C it meets the street with the sidewalk.

$(\exists x,y,z,p,q)(L(\_,x,A,B,A_{12},G_{s,}\_)\Pi$
$L(\_,x,0,10km,A_{17},G_{s,}\_)\Pi L(\_,x,Point,Line,A_{15},G_{s,}\_)\Pi$
$L(\_,x,East,East,A_{13},G_{s,}\_))\bullet\varepsilon_s\bullet(L(\_,x,p,C,A_{12},G_{s,}\_)$
$\Pi L(\_,y,q,C,A_{12},G_{s,}\_)\Pi L(\_,z,y,y,A_{12},G_{s,}\_))$
$\wedge road(x)\wedge street(y)\wedge sidewalk(z)\wedge p\neq q$ (25)

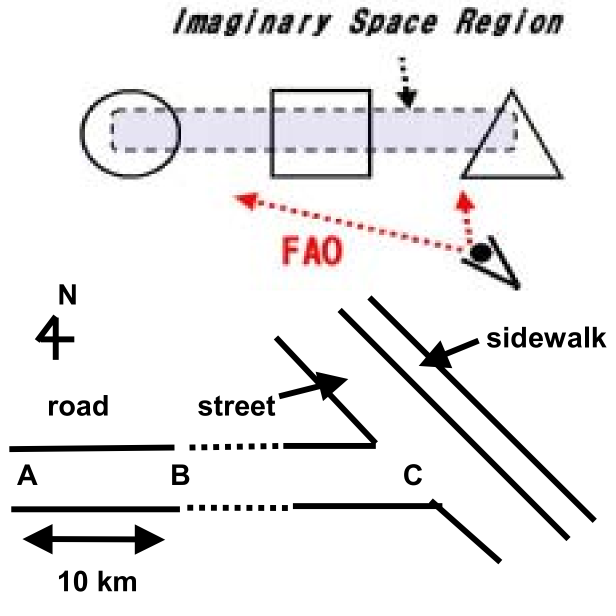
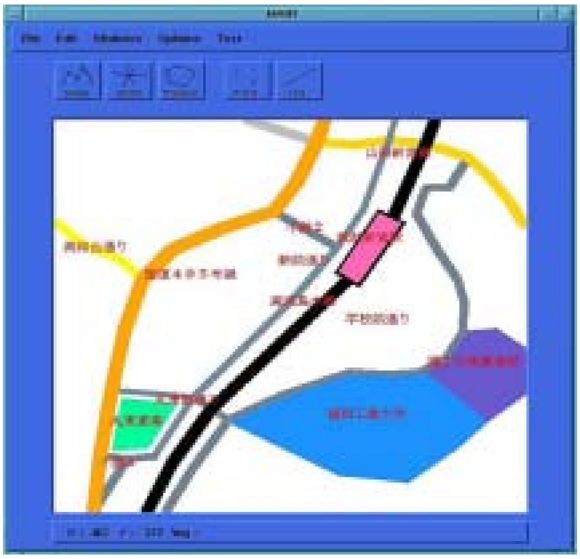


Figure 8. Complicated spatial events: 'row' (up) and 'example of road map' (down)



(a) A map generated from a locus formula by IMAGES-M

H: How does the national road run?
S: It extends between Pref. A and Pref. C via Pref. B.
H: Where does the bus go from the rail way station A?
S: It reaches the town D.
H: What is between the buildings A and B?
S: The railway D.
H: Where do the street A and the road B meet?
S: At the crossing C.
H: Where do the street A and the road B separate?
S: At the crossing C.

(b) Q-A on the map (a) by human (H) and IMAGES-M (S)

Figure 9. Cross-media operations as pragmatic understanding

Figures 9 (b) shows an example of question-answering on the real map (a) between a human and IMAGES-M [6, 10, 12], where the map is a pictorial interpretation of a locus formula by PPU. The system understood the query texts pragmatically by anchoring them to the map as a model of the real world, utilizing effectively several kinds of intuitive postulates such as PRS, as a matter of course, where distinction between temporal and spatial events is crucially important.

4 IMITATION GUIDED BY SUGGESTION

4.1 Definition

As shown in Figures 10 and 11, robotic imitation intended here is defined as a human-robot interaction where a human presents a robot a pair of demonstration and suggestion that is the expression of his/her intention and it behaviouralizes its conception, namely, the result of semantic and pragmatic understanding of the suggestion.

The processes shown in Figures 10 and 11 can be formalized as follows, where the pair of $\mathbf{P}_i$ and $\mathbf{Def}_i$ is called 'Conception' for the i-th imitation and denoted by $\mathbf{C}_i$.

$\mathbf{Int}_i \Rightarrow \mathbf{T}_i , \mathbf{D}_i$

$\mathbf{T}_i , \mathbf{K}_L \Rightarrow \mathbf{S}_i$

$\mathbf{D}_i , \mathbf{K}_D \Rightarrow \mathbf{Per}_i$

$\mathbf{S}_i , \mathbf{Per}_i, \mathbf{K}_D \Rightarrow \mathbf{P}_i , \mathbf{Def}_i\ (= \mathbf{C}_i)$

$\mathbf{P}_i , \mathbf{Def}_i, \mathbf{K}_D \Rightarrow \mathbf{I}_i$

, where

$\mathbf{Int}_i$: The i-th intention by the human,
$\mathbf{T}_i$: The i-th suggestion by the human,
$\mathbf{S}_i$: Result of semantic understanding of the i-th suggestion,
$\mathbf{K}_L$: Linguistic knowledge in the robot,
$\mathbf{D}_i$: The i-th demonstration by the human,
$\mathbf{K}_D$: Domain-specific knowledge in the robot at the i-th session,
$\mathbf{Per}_i$: Perception of the i-th demonstration,
$\mathbf{P}_i$: Result of pragmatic understanding of the i-th suggestion,
$\mathbf{Def}_i$: Default specification for the i-th imitation,
$\mathbf{I}_i$: The i-th imitation by the robot,
⇒ : Conversion process (e.g., inference, translation).

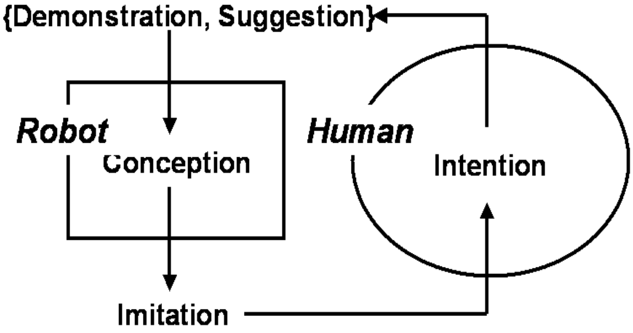


Figure 10. Imitation as human-robot interaction

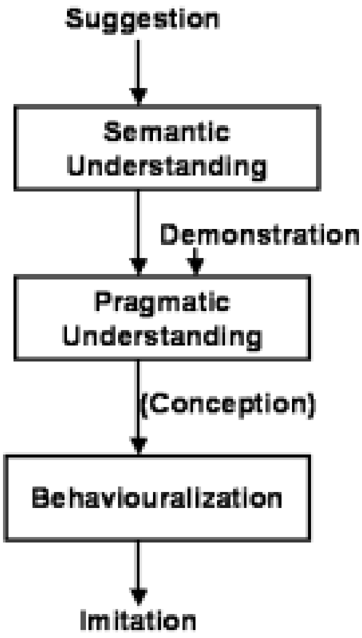


Figure 11. Imitation guided by suggestion

4.2 Theoretical simulation

As shown in Figure 10, it is assumed that there is a feedback loop between a human and a robot in order for the human to improve his/her previous suggestion or demonstration and for the robot to correct its previous imitation. For example, consider the scenario presented below and depicted in Figure12.

Scenario :

Robby is an intelligent humanoid robot and Tom is his user. Robby is called by Tom and enters Tom's room. This is Robby's first visit there. <u>Robby sees Tom leftward and the brown pillar forward</u> (, but doesn't see the green box or the yellow table). After a while, Tom tells Robby "Imitate me to my demonstration and suggestion."……

Here is described a theoretical simulation of the robotic imitation driven by the top-down control of the attention mechanism, which is almost that of problem finding/solving in the filed of AI [6, 12].

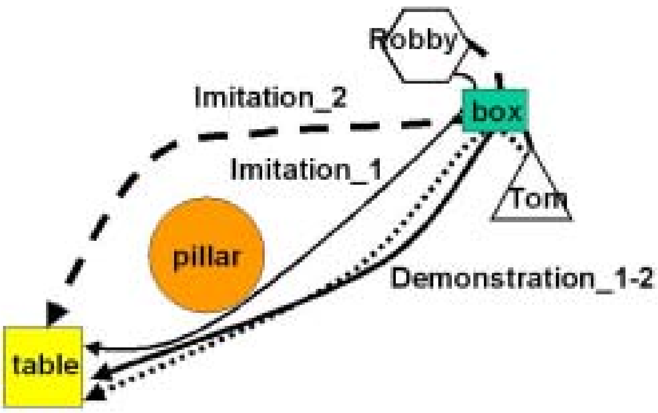


Figure 12. Tom's demonstrations and Robby's imitations

The sequence of the events assumed to happen is as follows.

[Robby's Perception of the initial situation, $\boldsymbol{Sit_0}$]

$$\begin{array}{l}\boldsymbol{Sit_0}\leftrightarrow L(\_,O_{21},\mathrm{Brown},\mathrm{Brown},A_{32},G_{t},\_)\Pi\\ L(\_,O_{22},\mathrm{Robby},\mathrm{Tom},A_{12},G_{s},\_)\Pi\\ L(\_,O_{22},\mathrm{Lw}_{21},\mathrm{Lw}_{21},A_{13},G_{s},\mathrm{Robby})\Pi\\ L(\_,O_{23},\mathrm{Robby},O_{21},A_{12},G_{s},\_)\Pi\\ L(\_,O_{23},\mathrm{Fw}_{21},\mathrm{Fw}_{21},A_{13},G_{s},\mathrm{Robby})\\ \wedge\mathrm{pillar}(O_{21})\wedge\mathrm{ISR}(O_{22})\wedge\mathrm{ISR}(O_{23})\end{array}$$

Robby's perception of the situation (i.e., the underlined part of the scenario) is still rough due to its economical working mode that is to be specified by each Standard (or precision). The attributes A_{32} *and* A_{13} *are 'Color' and 'Direction', respectively. The values* Fw_{21} *and* Lw_{21} *stand for 'forward' and 'leftward' viewed from Robby as designated at the Standard, respectively.*

[Tom's Intention_1, $\boldsymbol{Int_1}$]

$$\begin{array}{l}\boldsymbol{Int_1}\leftrightarrow\underline{L(\mathrm{Robby},\mathrm{Robby},O_{11},O_{13},A_{12},G_{t},\_)\Pi}\\ \underline{L(\mathrm{Robby},O_{11},\mathrm{Robby},\mathrm{Robby},A_{12},G_{t},\_)\Pi}\\ \underline{(L(\_,O_{14},\mathrm{Tom},O_{11},A_{12},G_{s},\_)\bullet L(\_,O_{14},O_{11},\mathrm{Robby},A_{12},G_{s},\_))\Pi}\\ \underline{L(\_,O_{14},D_{11},D_{11},A_{13},G_{s},\_)}\Pi L(\mathrm{Robby},\mathrm{Robby},V_{11},V_{11},A_{16},G_{t},\_)\Pi L\\ (\_,O_{15},\mathrm{Robby},O_{12},A_{12},G_{s},\_)\Pi L(\mathrm{Robby},O_{15},\mathrm{Dis},\mathrm{Dis},A_{44},G_{t},\_)\\ \wedge\mathrm{box}(O_{11})\wedge\mathrm{pillar}(O_{12})\wedge\mathrm{table}(O_{13})\wedge\mathrm{ISR}(O_{14})\wedge\mathrm{ISR}(O_{15})\end{array}$$

This formula implies that Tom wants Robby to carry the box between them to the table at a certain 'Velocity(A_{16})*',* V_{11} *without touching the pillar on the way, where '*O_{11}*' and '*O_{13}*' as the values of* A_{12} *represent their locations at each time point, and '*D_{11}*' is the direction to the box and Robby viewed from Tom.*

Tom is conscious that every attribute value to specify Robby's action is essentially vague but he believes that it should be imitated within certain tolerance associated with each Standard. The values **Dis** *and* **Meet** *stand for 'disjoint' and 'meet (or touch)' in Topology*(A_{44})*, respectively.*

<SESSION_1>

[Tom's Suggestion_1, $\boldsymbol{T_1}$ and Demonstration_1, $\boldsymbol{D_1}$]

$\boldsymbol{Int_1\Rightarrow T_1, D_1}$

$\boldsymbol{T_1}\leftrightarrow$ "Go to the table with the box between us like this."

$\boldsymbol{D_1}\leftrightarrow$ Figure 12

Tom decides to verbalize only the underlined part of Intention_1, $\boldsymbol{Int_1}$ *saliently with the belief that the rest can be included in his demonstration. Tom converts (or translates)* $\boldsymbol{Int_1}$ *into* $\boldsymbol{T_1}$ *and* $\boldsymbol{D_1}$*.*

[Robby's Semantic_Understanding_1, $\boldsymbol{S_1}$]

$\boldsymbol{T_1, K_L\Rightarrow S_1}$

$$\begin{array}{l}\boldsymbol{S_1}\leftrightarrow(\exists\, x_1,x_2,x,y,z,p)L(x_2,x_2,y,x,A_{12},G_{t},\_)\Pi\\ \quad L(x_2,y,x_2,x_2,A_{12},G_{t},\_)\Pi\,(L(\_,z,x_2,y,A_{12},G_{s},\_)\bullet\\ \quad L(\_,z,y,x_1,A_{12},G_{s},\_))\Pi L(\_,z,p,p,A_{13},G_{s},\_)\\ \quad\wedge x_2\neq x\wedge x_2\neq y\wedge\mathrm{box}(y)\wedge\mathrm{table}(x)\wedge\mathrm{ISR}(z)\\ \quad\wedge\mathrm{person\_1}(x_1)\wedge\mathrm{person\_2}(x_2)\end{array}$$

Robby interprets $\boldsymbol{T_1}$ *into* $\boldsymbol{S_1}$*. The variable '*x*' or '*y*' is not yet anchored to the 'real table' or the 'real box' in the real environment because Robby has not perceived them yet. The predicates '*person_1*' and '*person_2*' refer to the first person (I) and the second person (You) and are to be pragmatically understood as 'Tom' and 'Robby', respectively.*

[Robby's Pragmatic_Understanding_1, $\boldsymbol{P_1}$ and Default_1, $\boldsymbol{Def_1}$]

$\boldsymbol{D_1\Rightarrow Per_1}$

$\boldsymbol{S_1, Per_1, K_D \Rightarrow P_1, Def_1}$

$$\begin{array}{l}\boldsymbol{P_1}\leftrightarrow L(\mathrm{Robby},\mathrm{Robby},O_{24},O_{25},A_{12},G_{t},\_)\Pi\\ \quad L(\mathrm{Robby},O_{24},\mathrm{Robby},\mathrm{Robby},A_{12},G_{t},\_)\Pi\\ \quad(L(\_,O_{26},\mathrm{Robby},O_{25},A_{12},G_{s},\_)\bullet L(\_,O_{26},O_{25},\mathrm{Tom},A_{12},G_{s},\_))\Pi\\ \quad L(\_,O_{26},\mathrm{Lw}_{21},\mathrm{Lw}_{21},A_{13},G_{s},\_)\wedge\mathrm{box}(O_{24})\wedge\mathrm{table}(O_{25})\wedge\mathrm{ISR}(O_{26})\\ \boldsymbol{Def_1}\leftrightarrow L(\mathrm{Robby},\mathrm{Robby},1\mathrm{m/sec},1\mathrm{m/sec},A_{16},G_{t},\_)\wedge\ldots\end{array}$$

The 'Location (A_{12})*' is attended to according to* $\boldsymbol{S_1}$*.* $\boldsymbol{Per_1}$ *makes Robby aware that the words 'box' and 'table' should be anchored to the 'green object* O_{24}*' and the 'yellow object* O_{25}*' behind the pillar in the real environment, respectively. Robby conceives that he should approach to the table at his certain Standard.* $\boldsymbol{Def_1}$ *is inferred from* $\boldsymbol{Per_1}$ *and* $\boldsymbol{K_D}$ *as the default specification for the attributes not explicit in* $\boldsymbol{T_1}$*.*

[Robby's Imitation_1, $\boldsymbol{I_1}$]

$\boldsymbol{P_1, Def_1, K_D \Rightarrow I_1}$

$\boldsymbol{I_1}\leftrightarrow$Figure12

Robby imitates $\boldsymbol{D_1}$ *according to* $\boldsymbol{P_1}$*,* $\boldsymbol{Def_1}$ *and* $\boldsymbol{K_D}$*.*

----- Resetting the situation to the initial situation $\boldsymbol{Sit_0}$-----

<SESSION_2>

[Tom's Suggestion_2, $\boldsymbol{T_2}$ and Demonstration_2, $\boldsymbol{D_2}$]

$\boldsymbol{I_1\Rightarrow PI_1}$

$\boldsymbol{Int_1, {\sim}PI_1\Rightarrow Int_2}$

$\boldsymbol{Int_2\Rightarrow T_2, D_2}$

$\boldsymbol{T_2}\leftrightarrow$"Don't touch the pillar."

$\boldsymbol{D_2}\leftrightarrow$ Figure 12

Tom perceives $\boldsymbol{I_1}$ *as* $\boldsymbol{PI_1}$*. He denies* $\boldsymbol{PI_1}$ *and creates* $\boldsymbol{Int_2}$ *followed by* $\boldsymbol{T_2}$ and $\boldsymbol{D_2}$*.*

[Robby's Semantic_Understanding_2, S_2]

$T_2, K_L \Rightarrow S_2$

$S_2 \leftrightarrow (\exists x)L(\_,y,Robby,O_{21},A_{12},G_s,\_)$
$\Pi{\sim}L(Robby,x,Dis,Meet,A_{44},G_t,\_) \wedge ISR(x) \wedge pillar(O_{21})$

Robby gets aware that his imitation has been denied at the change of attribute 'Topology (A44)*' from '***Dis**joint*' to '***Meet***'.*

[Robby's Pragmatic_Understanding_2, P_2 and Default_2, Def_2]

$D_2 \Rightarrow Per_2$

$S_2, Per_2, K_D \Rightarrow P_2, Def_2$

$P_2 \leftrightarrow P_1 \wedge \underline{L(\_,O_{27},Robby,O_{21},A_{12},G_s,\_)\Pi}$
$\underline{L(Robby,O_{27},Dis,Dis,A_{44},G_t,\_) \wedge pillar(O_{21}) \wedge ISR(O_{27})}$

$Def_2 \leftrightarrow L(Robby,Robby, 1m/sec, 1m/sec,A_{16},G_t,\_) \wedge \ldots$

According to S_2*, the 'Location* (A12)*' of Robby and the pillar and their 'Topology* (A44)*' are especially attended to, and the underlined part is conceived in addition to* P_1*. No special attention is paid to the other attributes unmentioned yet.*

[Robby's Imitation_2, I_2]

$P_2, Def_2, K_D \Rightarrow I_2$

$I_2 \leftrightarrow$ Figure 12

-----Resetting the situation to the initial situation Sit_0-----

<SESSION_3>

[Tom's Suggestion_3, T_3 and Demonstration_3, D_3]

$I_2 \Rightarrow PI_2$

$Int_2, {\sim}PI_2 \Rightarrow Int_3\ (\leftrightarrow Null)$

$Int_3 \Rightarrow T_3, D_3$

$T_3 \leftrightarrow$"Alright."

$D_3 \leftrightarrow Null$

Tom fails to deny PI_2 *and comes to have no other intention (*$Int_3 \leftrightarrow Null$*). That is, Tom is satisfied by* I_2 *and only tells Robby* "Alright."

[Robby's Semantic_Understanding_3, S_3]

$T_3, K_L \Rightarrow S_3$

$S_3 \leftrightarrow (\exists x,y,k)L(x,y,1,1,B_{01},G_t,k) \wedge person(x)$

Tom gets aware that something 'y' *has evaluated by some person* 'x' *as perfect* '1' *at 'Worth (*B01*)' with a certain Standard* 'k'.

[Robby's Pragmatic_Understanding_3, P_3 and Default_3, Def_3]

$S_3, Per_3, K_D \Rightarrow P_3, Def_3$

$P_3 \leftrightarrow L(Tom,I_2,1,1,B_{01},G_t,Tom) \wedge person(Tom)$

$Def_3 \leftrightarrow L(Robby, I_3,/,/,A_{01},G_t,\_)$

Finally, Robby pragmatically conceives that Tom is satisfied by I_2 *at Tom's Standard and believes that the next imitation,* I_3 *is not needed to take 'Place of Existence* (A01)*'.*

[Robby's Imitation_3, I_3]

$P_3, Def_3, K_D \Rightarrow I_3$

$I_3 \leftrightarrow$ Null

Finally, no more imitation is performed.

-----End of all the sessions-----

5 TOP-DOWN CONTROL BASED ON L_{md}

5.1 Attention mechanism

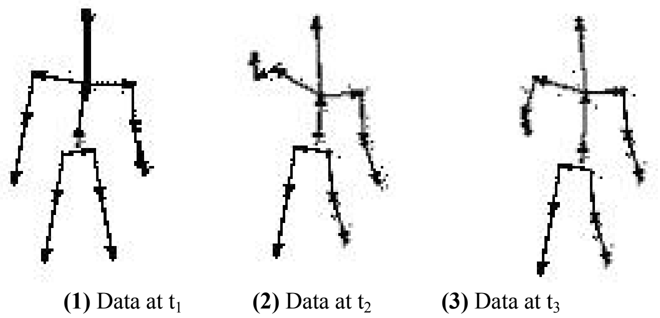
As mentioned above, the semantic understanding of human verbal suggestion makes a robot abstractly (i.e., conceptually) aware which matters and attributes involved in human demonstration should be attended to, and its pragmatic understanding provides the robot with concrete idea of real matters with real attribute values significant for imitation. More exactly, semantic understanding in L_{md} of human suggestion enables the robot to control its attention mechanism in such a top-down way that focuses the robot's attention on the significant attributes of the significant matters involved in human demonstration. Successively, in order for pragmatic understanding in L_{md} of human suggestion, the robot is to select the appropriate sensors corresponding with the suggested attributes and make them run on the suggested matters so as to pattern after the movements of human FAO implied by the locus formulas yielded in semantic understanding. ***That is to say in short,*** L_{md} ***expression suggests a robot what and how should be attended to in human demonstration and its environment.***

For example, consider such a suggestion as S12 presented to a robot by a human. In this case, unless the robot is aware of the existence of a certain box between the stool and the desk, such semantic understanding of the underlined part as (26) and such a semantic definition of the word 'box' as (27) are very helpful for it. The attributes A_{12} (Location), A_{13} (Direction), A_{32} (Color), A_{11} (Shape) and the spatial event on A_{12} in these L_{md} expressions indicate that the robot has only to activate its vision system in order to search for the box from the stool to the desk during the pragmatic understanding. That is, the robot can attempt to understand pragmatically the words of objects and events in an integrated top-down way.

(S12) Avoid <u>the green box between the stool and the desk</u>.

$$(\exists x_1,x_2,x_3,x_4,p)(L(\_,x_4,x_1,x_2,A_{12},G_s,\_)\bullet L((\_,x_4,x_2,x_3,A_{12},G_s,\_))\Pi$$
$$L(\_,x_4,p,p,A_{13},G_s,\_)\Pi L(\_,x_2,Green,Green,A_{32},G_t,\_)$$
$$\wedge stool(x_1)\wedge box(x_2)\wedge desk(x_3)\wedge ISR(x_4) \qquad (26)$$

$$(\lambda x)box(x)\leftrightarrow(\lambda x)L(\_,x,Hexahedron,Hexahedron,A_{11},G_t,\_)$$
$$\wedge container(x) \qquad (27)$$



(1) Data at t_1 **(2)** Data at t_2 **(3)** Data at t_3

Figure 13. Graphical interpretations of real motion data

Tom moved the right arm.
Tom raised the right arm.
Tom bent the right arm.
……………

(a) Text for motion data from t_1 to t_2.

……………
Tom lowered the right arm.
Tom stretched the right arm and simultaneously lowered the right arm.
……………

(b) Text for motion data from t_2 to t_3.

Figure 14. Texts generated from real motion data

This top-down control of attention mechanism enables IMAGES-M can take in real human motion data through the motion capturing system in SDPU. For example, Figure 13 shows graphical interpretations of the real motion data taken in at the time point t_1, t_2 and t_3. These real data were translated via $\boldsymbol{L_{md}}$ into such texts as shown in Figure 14 by TPU. In this case, IMAGES-M's attention was guided by the suggestion S13 below.

(S13) Move your right arm like this.

5.2 Utilization of domain-specific knowledge

The linguistic knowledge $\mathbf{K_L}$ is employed exclusively for semantic understanding, consisting of syntactic and semantic rules and dictionaries. On the other hand, the domain-specific knowledge $\mathbf{K_D}$ is employed for pragmatic understanding and behaviouralization, containing all kinds of knowledge pieces acquired so far concerning the robot, the human and their environment. For example, the human body can be described in a computable form using locus formulas. That is, the structure of the human body is one kind of spatial event where the body parts such as head, trunk, and limbs extend spatially and connect with each other. The expressions (28) and (29) are examples of these descriptions in $\boldsymbol{L_{md}}$, reading that an arm extends from a hand to a shoulder and that a wrist connects a hand and a forearm, respectively.

$$(\lambda x)\mathrm{arm}(x)\leftrightarrow(\lambda x)(\exists y_1,y_2)L(\_,x,y_1,y_2,A_{12},G_{s,}\_)$$
$$\wedge\mathrm{shoulder}(y_1)\wedge\mathrm{hand}(y_2) \qquad (28)$$

$$(\lambda x)\mathrm{wrist}(x)\leftrightarrow(\lambda x)(\exists y_1,y_2,y_3,y_4)(L(\_,y_1,y_2,x,A_{12},G_{s,}\_)\bullet$$
$$L(\_,y_1,x,y_3,A_{12},G_{s,}\_))\wedge\text{body-part}(y_1)\wedge\mathrm{forearm}(y_2)$$
$$\wedge\mathrm{hand}(y_3) \qquad (29)$$

These descriptions are necessary for the robot to understand human action and text well enough to obtain an appropriate conception, eliminating such an anomalous one as is represented by S14 in a top-down way.

(S14) The left arm moved away from the left shoulder and
the left hand.

Each of such human's/robot's motions (M_k) as 'walk' and 'bow' is given as an ordered set of its standardized characteristic snapshots (S_k) called 'Standard Motion' and defined by (30). In turn, a family (F_X) of S_ks is called 'Family of Standard Motions' and defined by (31), where the suffix 'X' refers to 'human (X=H)' or 'robot (X=R)'. The families F_H and F_R are contained in $\mathbf{K_D}$ and their members are employed for the default motions, namely, motions not specified in human suggestion or demonstration, during pragmatic understanding.

$$S_k=\{M_{kS}, \ldots, M_{kE}\} \qquad (30)$$

$$F_X=\{S_1, S_2, \ldots, M_N\} \qquad (31)$$

For example, the $\boldsymbol{L_{md}}$ expression of human walking in default is given by (32), reading that a human moves by his/her legs making his/her shape change monotonically from $\mathrm{Walk_S}$ to $\mathrm{Walk_E}$.

$$(\exists x,y,p_1,p_2,q_1,q_2)\ L(\_,y,x,x,A_{01},G_{t,}\_)\Pi$$
$$L(y,x,q_1,q_2,A_{12},G_{t,}\_)\Pi\ L(x,x,\mathrm{Walk_S},\mathrm{Walk_E},A_{11},G_t,F_H)$$
$$\wedge q_1\neq q_2\wedge\mathrm{human}(x)\wedge\mathrm{legs}(y) \qquad (32)$$

For another example, the $\boldsymbol{L_{md}}$ expression (33) is for the robotic motion of head shaking in default, reading that a robot affects its head in the Orientation (A_{14}), making its shape change monotonically from $\mathrm{Shake\_head_S}$ to $\mathrm{Shake\_head_E}$. The shape values are given in a computable form general enough to reconstruct any human/robot motion in 3D graphics or so. Figure 15 shows an example of its interpretation in 3D graphics by PPU in IMAGES-M, which is also an example of cross-media translation from the text 'The robot shakes its head' into the animation.

$$(\exists x,y,p_1,p_2)L(\_,y,x,x,A_{01},G_{t,}\_)\Pi L(x,y,p_1,p_2,A_{14},G_{t,}\_)\Pi$$
$$L(x,x,\mathrm{Shake\_head_S},\mathrm{Shake\_head_E},A_{11},G_t,F_R)$$
$$\wedge\mathrm{robot}(x)\wedge\mathrm{head}(y) \qquad (33)$$

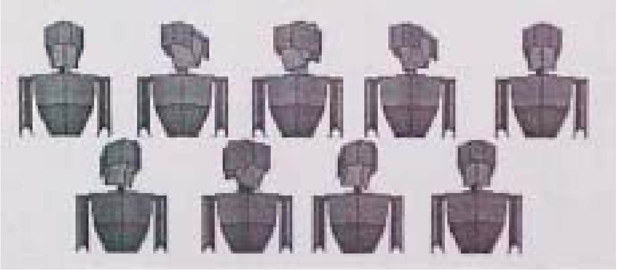


Figure 15. 3D animation of 'The robot shakes its head.'

5.3 Behaviouralization

The process for behaviouralization is to translate a conception (i.e., $\mathbf{C_i}$) into an imitation (i.e., $\mathbf{I_i}$) as a appropriate sequence of control codes for certain sensors or actuators in the robot to be decoded into a real behaviour by SDPU or ADPU in IMAGES-M. For this purpose, there are needed two kinds of core procedures so called 'Locus formula paraphrasing' and 'Behaviour chain alignment' as detailed below.

5.3.1 Locus formula paraphrasing

The attributes listed in Table 1 are essentially for human sensors or actuators and therefore the locus formula as $\mathbf{C_i}$ should be translated into its equivalent concerning the attributes specific to the robot's. For example, an atomic locus of the robot's 'Shape (A_{11})' specified by the human should be paraphrased into a set of atomic loci of the 'Angularity (A_{45})' of each joint in the robot. For another example, 'Velocity (A_{16})' for the human into a set of change rates in 'Angularity (A_{45})' over 'Duration (A_{35})' (i.e., A_{45}/A_{35}) of the robot's joints involved. These knowledge pieces are called 'Attribute Paraphrasing Rules (APRs)' [10] and contained in $\mathbf{K_D}$.

5.3.2 Behaviour chain alignment

Ideally, the atomic loci in the conception $\mathbf{C_i}$ (original or paraphrased) should be realized as the imitation $\mathbf{I_i}$ in a perfect correspondence with an appropriate chain of sensor or actuator deployments. Actually, however, such a chain as a direct translation of $\mathbf{C_i}$ must often be aligned to be feasible for the robot due to the situational, structural or functional differences between the human and the robot. For example of situational difference, in the simulation above, the robot must interpolate the travel from its initial location to the green box and the action to pick up the box. On the other hand, for example of structural or functional difference, consider the case of imitation by a non-humanoid robot. Figure 16 shows the action by a dog-shaped robot (SONY) to the suggestion 'Walk and wave your left hand.' The robot pragmatically understood the suggestion as '*I* walk and wave *my* left *foreleg*' based on the knowledge piece that only forelegs can be waved' and behaviouralized its conception as 'I walk *BEFORE* sitting down *BEFORE* waving my left foreleg' but not as 'I walk, *SIMULTANEOUSLY* waving my left foreleg', in order not to fall down.

The procedure here [6, 12] is based on the conventional AI, where a problem is defined as the difference or gap between a 'Current State' and a 'Goal State' and a task as its cancellation. Here, the term 'Event' is preferred to the term 'State' and 'State' is defined as static 'Event' which corresponds to a level locus. On this line, the robot needs to interpolate some transit event X_T between the two events, 'Current Event (X_C)' and 'Goal Event (X_G)' as (34).

$$X_C \bullet X_T \bullet X_G \qquad (34)$$

According to this formalization, a problem X_P can be defined as $X_T \bullet X_G$ and a task can be defined as its realization and any problem is to be detected by the unit of atomic locus. For example, employing such a postulate as (35) implying 'Continuity in attribute values', the event X in (36) is to be inferred as (37).

$$L(x,y,p_1,p_2,a,g,k) \bullet L(z,y,p_3,p_4,a,g,k) .\supset. p_3=p_2 \qquad (35)$$

$$L(x,y,q_1,q_2,a,g,k) \bullet X \bullet L(z,y,q_3,q_4,a,g,k) \qquad (36)$$

$$L(z',y,q_2,q_3,a,g,k) \qquad (37)$$

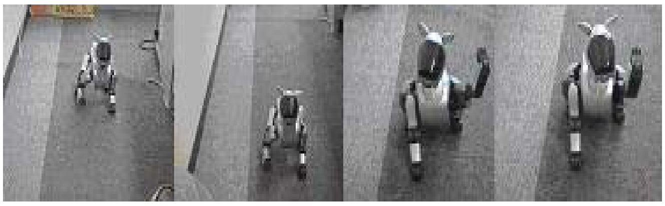


Figure 16. Robot's action to 'Walk and wave your left hand'

6 DISCUSSION AND CONCLUSION

The key contribution of this paper is the proposal of a novel idea of robotic imitation driven by semantic representation of human suggestion, where are hinted in the formal language $\boldsymbol{L_{md}}$ what and how should be attended to in human action as analogy of human FAO movement and thereby the robotic attention can be controlled in a top-down way. Without such a control, a robot is to simultaneously attend to tens of attributes of every matter involved in human action as shown in Table 1. This is not realistic, considering the difficulties in autonomous robotic vision understanding today. The author has a good perspective for the proposed theory of robotic imitation based on his previous work utilizing $\boldsymbol{L_{md}}$ for robot manipulation by text [6, 12]. This is one kind of cross-media operation via intermediate $\boldsymbol{L_{md}}$ representation [e.g., 6, 10, 12]. At my best knowledge, there is no other theory or system that can perform cross-media operations in such a seamless way as ours. This is due to the descriptive power of $\boldsymbol{L_{md}}$ enabling systematic organization and computation of spatiotemporal knowledge including sensation and action. Our future work will include establishment of learning facilities for automatic acquisition of word concepts from sensory data and multimodal interaction between humans and robots under real environments in order to realize the robotic imitation proposed here.

ACKNOWLEDGEMENTS

This work was partially funded by the Grants from Computer Science Laboratory, Fukuoka Institute of Technology and Ministry of Education, Culture, Sports, Science and Technology, Japanese Government, numbered 14580436 and 17500132.

REFERENCES

[1] A.Billard, 'Learning motor skills by imitation: biologically inspired robotic model', ***Cybernetics and Systems***, **32**, 155–193, (2000).

[2] A.Alissandrakis, C.L.Nehaniv, and K.Dautenhahn, 'Imitating with ALICE: Learning to imitate corresponding actions across dissimilar embodiments', ***IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans***, **32-4**, 482–496, (2003).

[3] J.Nakanishi, J.Morimoto, G.Endo, G.Cheng, S.Schaal, and M.Kawato, 'Learning from demonstration and adaptation of biped locomotion', ***Robotics and Autonomous Systems***, **47**(2-3), 79–81, (2004).

[4] J.M.Wolfe, 'Visual search in continuous, naturalistic stimuli', ***Vision Research***, **34**, 1187–1195, (1994).

[5] Y.Demiris and B.Khadhouri, 'Hierarchical attentive multiple models for execution and recognition of actions', ***Robotics and Autonomous Systems***, 54, 361–369, (2006).

[6] M.Yokota, 'Towards a universal language for distributed iIntelligent robot networking', ***Proc. of 2006 IEEE International Conference on Systems, Man and Cybernetics***, Taipei, Taiwan, (Oct., 2006).

[7] S.Coradeschi and A.Saffiotti, 'An introduction to the anchoring problem', ***Robotics and Autonomous Systems***, **43**, 85–96, (2003).

[8] E.Drumwright, V.Ng-Thow-Hing, and M.J.Mataric´, 'Toward a vocabulary of primitive task programs for humanoid robots', ***Proc. of International Conference on Development and Learning*** (*ICDL*), Bloomington,IN, (May, 2006).

[9] M.Yokota, 'An approach to integrated spatial language understanding based on Mental Image Directed Semantic Theory', ***Proc. of 5th Workshop on Language and Space***, Bremen, Germany, (Oct., 2005).

[10] M.Yokota and G.Capi, 'Cross-media operations between text and picture based on Mental Image Directed Semantic theory', ***WSEAS Trans. on INFORMATION SCIENCE and APPLICATIONS***, Issue 10, 2, 1541–1550, (2005).

[11] J.F. Sowa, ***Knowledge Representation: Logical, Philosophical, and Computational Foundations***, Brooks Cole Publishing Co., Pacific Grove, CA, (2000).

[12] M.Yokota, 'Integrated multimedia understanding for ubiquitous intelligence based on Mental Image Directed Semantic Theory', ***Handbook on Mobile and Ubiquitous Computing Innovations and Perspectives***, American Scientific Publishers, (in press).

[13] M.Egenhofer, 'Point-set topological spatial relations. Geographical Information Systems', 5,2 161-174 (1991).

[14] M.Nicolescu, M.J.Mataric´, 'Task Learning Through Imitation and Human-Robot Interaction, in Models and Mechanisms of Imitation and Social Learning in Robots, Humans and Animals: Behavioural, Social and Communicative Dimensions, Kerstin Dautenhahn and Chrystopher Nehaniv Eds., 407-424, (2006).