The Fourteenth International Symposium on Artificial Life and Robotics 2009 (AROB 14th '09),
B-Con Plaza, Beppu, Oita, Japan, February 5 - 7, 2009

A method for top-down control of robotic attention based on Mental-image Description Language, L_{md}

Masao Yokota, Kenji Sasaki, Reiichi Kaida, Tetsushi Oka and Kaoru Sugita

Fukuoka Institute of Technology, 3-30-1 Wajiro-higashi, Higashi-ku, Fukuoka-shi, 811-0295, Japan
(Tel : 81-92-606-5897; Fax : 81-92-606-8923)
(yokota@fit.ac.jp)

Abstract: Mental Image Directed Semantic Theory (MIDST) has defined the semantic content (i.e., concept) of a spatiotemporal expression as a certain generalized mental image of its referents in the physical world and proposed a method to model mental images as "loci in attribute spaces" formalized in the formal language L_{md}. The most remarkable feature of L_{md} is its capability of formalizing spatiotemporal events based on a hypothesis of human attention mechanism. This paper presents a systematic method for top-down control of robotic attention by L_{md} representation with some computer simulation results.

Keywords: Natural language, Multimedia understanding, Robotic sensation and action.

I. INTRODUCTION

The authors have been working on integrated multimedia understanding for intuitive human-robot interaction, that is, interaction between non-expert or ordinary people and home robots as shown in Fig.1 [1-4]. In such a situation, natural language is the leading information medium for their communication as well as for the communication between ordinary people 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.

For such an intuitive human-robot interaction intended here, 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 human/robotic 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., 5, 6]. These expression schemas, however, are too linguistic or coarse to represent and compute sensory/motory events in such an integrated way as intended here.

Mental Image Directed Semantic Theory (MIDST) [1] 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, L_{md} (Mental-image Description Language). This language is employed for predicate logic and has already been implemented on several types of computerized intelligent systems [1-4].

This paper presents a systematic method for top-down control of robotic attention by L_{md} representation with some computer simulation results.

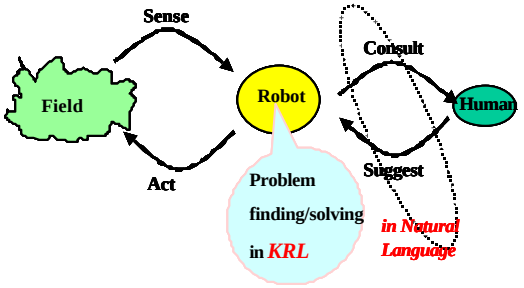

Fig.1. Intuitive human-robot interaction

II. IMAGE, ATTENTION AND L_{md}

MIDST models omnisensory mental images as "Loci in Attribute Spaces". 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 gray triangular object shown in Fig.2-Left 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. A general locus is to be articulated by "Atomic Locus" as depicted in Fig.2-Right and formulated as (1).

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

The intuitive interpretation of (1) is given as follows.

"Matter 'x' causes Attribute 'a' of Matter 'y' to keep (p=q) or change (p ≠ q) its values temporally (g=Gt) or spatially (g=Gs) over a time-interval, where the values 'p' and 'q' are relative to the standard 'k'."

When $g=G_t$, the locus indicates monotonic change or constancy of the attribute in time domain and when $g=G_s$, 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 the attribute 'Physical Location'. These two formulas are different only at the term 'Event Type'.

(S1) The bus runs from Tokyo to Osaka.

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

(S2) The road runs from Tokyo to Osaka.

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

The formal language $\boldsymbol{L_{md}}$ has employed 'tempo-logical connectives (TLCs)' representing both logical and temporal relations between loci. Articulated loci are combined with tempo-logical conjunctions, where 'SAND ($\wedge_0$)' and 'CAND ($\wedge_1$)' are most frequently utilized, standing for 'Simultaneous AND' and 'Consecutive AND', conventionally symbolized as 'Π' and '•', respectively. For example, the expression (4) is the definition of the English verb concept 'fetch' depicted as Fig.3-Left. This implies such a temporal event that 'x' goes for 'y' and then comes back with it.

$$(\lambda x,y)fetch(x,y)$$
$$\leftrightarrow(\lambda x,y)(\exists p1,p2,k)L(x,x,p1,p2,A12,Gt,k)\bullet$$
$$((\underline{L(x,x,p2,p1,A12,Gt,k)}\Pi L(x,y,p2,p1,A12,Gt,k))$$
$$\wedge x\neq y\wedge p1\neq p2 \quad (4)$$

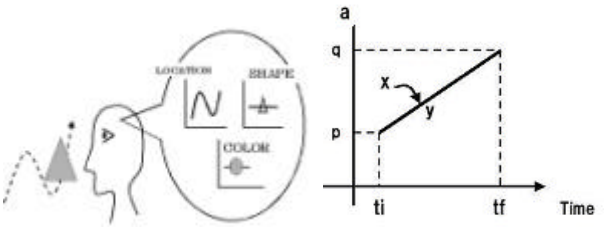

Fig.2. Mental image model (Left) and Atomic Locus in Attribute Space (Right).

Fig.3. Image of 'fetch'(Left) and Event types (Right).

It has been often argued that human active sensing processes may affect perception and in turn conceptualization and recognition of the physical world. 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). To be brief, the FAO is fixed on the whole AC in a temporal event but *runs* about on the AC in a spatial event. Consequently, as shown in Fig.3-Left, 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. That is, ***all loci in attribute spaces correspond one to one with movements or, more generally, temporal events of the FAO. This implies that L_{md} expression can suggest a robot what and how should be attended to in its environment.*** And this is why S3 and S4 can refer to the same scene in spite of their appearances, where what 'sinks' or 'rises' is the FAO and whose conceptual descriptions are given as (5) and (6), respectively, where 'A13', '↑' and '↓' refer to the attribute 'Direction' and its values 'upward' and 'downward', respectively. Such a fact is generalized as '***Postulate of Reversibility of a Spatial Event*** (PRS) ' that can be one of the principal inference rules belonging to people's common-sense knowledge about geography. These pairs of conceptual descriptions are called ***equivalent in the PRS***, and the paired sentences are treated as ***paraphrases*** each other.

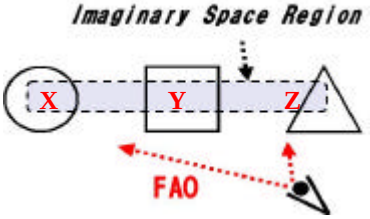

Fig.4. Spatial event 'row' and FAO movement.

(S3) The path *sinks to* the brook.

$$(\exists x,y,p,z,k_1,k_2)L(x,y,p,z,A_{12},G_s,k_1)\Pi$$
$$L(x,y,\downarrow,\downarrow,A_{13},G_s,k_2)\wedge path(y)$$
$$\wedge brook(z)\wedge p\neq z \quad (5)$$

(S4) The path *rises from* the brook.

$$(\exists x,y,p,z,k_1,k_2)L(x,y,z,p,A_{12},G_s,k_1)\Pi$$
$$L(x,y,\uparrow,\uparrow,A_{13},G_s,k_2)\wedge path(y)$$
$$\wedge brook(z)\wedge p\neq z \quad (6)$$

For another example of spatial event, Fig.4 concerns the perception of the formation of multiple isolated objects, where FAO runs along an imaginary object so called 'Imaginary Space Region (ISR)'. This spatial event can be verbalized as S5 using the preposition 'between' and formulated as (7) or (8), corresponding also to such concepts as 'row', 'line-up', etc.

(S5) Y is between X and Z.

$$(\exists x,y,p,q,k_1,k_2)(L(x,y,X,Y,A_{12},G_s,k_1)\Pi$$
$$L(x,y,p,p,A_{13},G_s,k_2))\bullet(L(x,y,Y,Z,A12,G_s,k_1)\Pi$$
$$L(x,y,q,q,A_{13},G_s,k_2))\wedge ISR(y)\wedge p=q \quad (7)$$

$$(\exists x,y,p,k_1,k_2)(L(x,y,Z,Y,A_{12},G_s,k_1)\bullet$$
$$L(x,y,Y,X,A_{12},G_s,k_1))\Pi L(x,y,p,p,A_{13},G_s,k_2)\wedge ISR(y) \quad (8)$$

At our best knowledge, there is no other theory or method [e.g., 7, 8] that can provide spatiotemporal expressions with semantic interpretation in such a systematic way where both temporal and spatial events are simply and adequately formulated by controlling the term of Event Type of the atomic locus formula reflecting FAO movement. 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 imply that ordinary people live their casual life, attending to tens of attributes of the matters in the world to cognize them in comparison with several kinds of standards. That is, without any verbal hint, it is extremely difficult for a robot to understand which part of its environment is significant or not for people because there are too many things to attend to as it is.***

Table 1. List of attributes

Code	Attribute [Property†] (words/phrases
*A01	PLACE OF EXISTENCE [N] (*happen*,
*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*,
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*,
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*,
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*'.

III. ATTENTION CONTROL BY L_{md}

The description of an event in $\boldsymbol{L_{md}}$ is compared to a movie film recorded through a floating camera because it is necessarily grounded in FAO's movement over the event. ***That is to say in short, L_{md} expression suggests a robot what and how should be attended to in its environment.*** Therefore, the robotic attention can be controlled in a top-down way based on $\boldsymbol{L_{md}}$ expression.

For example, consider such a suggestion to a robot as S6 whose semantic interpretation is given by (26), where 'avoid' is defined as 'keep Topology (A_{44}) Disjoint (=***Dis***)'. In this case, unless the robot is aware of the existence of a certain box between the stool and the desk, such semantic understanding as the underlined part of (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 $\boldsymbol{L_{md}}$ 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.

(S6) Avoid the green box between the stool and the desk.

$(\exists x_1,x_2,x_3,x_4,x_5,x_6,y_1,y_2,p,k_1,k_2,k_3,k_4)$
$L(x_6,x_5,\boldsymbol{Dis},\boldsymbol{Dis},A_{44},G_t,k_4)\Pi L(x_6,x_5,x_2,x_6,A_{12},G_s,k_1)\Pi$
$(\underline{L(y_1,x_4,x_1,x_2,A_{12},G_s,k_1)}\bullet\underline{L(y_1,x_4,x_2,x_3,A_{12},G_s,k_1)})\Pi$
$L(y_1,x_4,p,p,A_{13},G_s,k_2)\Pi L(y_2,x_2,\mathrm{Green},\mathrm{Green},A_{32},G_t,k_3)$
$\wedge \mathrm{stool}(x_1)\wedge \mathrm{box}(x_2)\wedge \mathrm{desk}(x_3)\wedge \mathrm{ISR}(x_4)\wedge \mathrm{ISR}(x_5)$
$\wedge \mathrm{robot}(x_6)$ (26)

$(\lambda x)\mathrm{box}(x)\leftrightarrow(\lambda x)(\exists y,k)L(y,x,\mathbf{Hexahedron},\mathbf{Hexahedron},A_{11},G_t,k)\wedge \mathrm{container}(x)$ (27)

Figure 5 shows the simulated action of a virtual robot to the command S7. The robot's pragmatic understanding of this command is given as (28), where, 'Robot_0' refers to the virtual robot itself, 'D_c' is the direction from '$\mathrm{Rectangle}_1$' to '$\mathrm{Triangle}_1$' calculated from their locations, and P_c and P_g are the current and the goal locations of 'Robot_0', respectively.

(S7) Go to between the rectangle and the triangle, avoiding the pentagon.

$L(\mathrm{Robot}_0,\mathrm{Robot}_0,P_c,P_g,A_{12},G_t,\_)$
$\Pi L(\mathrm{Robot}_0,\mathrm{ISR}_2,\mathrm{Disjoint},\mathrm{Disjoint},A_{44},G_t,\_)$
$\Pi(L(\_,\mathrm{ISR}_1,\mathrm{Rectangle}_1,P_g,A_{12},G_s,\_)\bullet$
$L(\_,\mathrm{ISR}_1,P_g,\mathrm{Triangle}_1,A_{12},G_s,\_))$
$\Pi L(\mathrm{Robot}_0,\mathrm{ISR}_1,D_c,D_c,A_{13},G_s,\_)$
$\Pi L(\mathrm{Robot}_0,\mathrm{ISR}_2,\mathrm{Pentagon}_1,\mathrm{Robot}_0,A_{12},G_s,\_)$ (28)

The process flow for this simulation is roughly as follows [1-4].

[STEP1] Syntactic interpretation: production of a surface dependency structure (SDS) from S7.

[STEP2] Semantic understanding: production of a generalized (conceptual) interpretation U_s based on the SDSs and the semantic definitions of the words included in S7.

[STEP3] Pragmatic understanding: production of such a concrete interpretation as (28) by grounding the variables of U_s onto the matters in the environment.

[STEP4] Behavioralization: production of the action to S7 so as to satisfy the conditions indicated in (28) in the top-down way controlled by the attributes involved.

The text understanding process above is completely reversible except that multiple paraphrases can be generated by tempological reasoning as shown in Fig.6 because event patterns are sharable among multiple word concepts, where text-generation is also controlled in a top-down way in use of attributes involved.

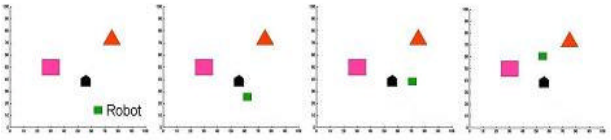

Fig. 5. Simulation in Matlab of the command 'Go to between the rectangle and the triangle, avoiding the pentagon'.

(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.
The stick moves simultaneously when Tom goes.
……

Fig.2. Text paraphrasing by tempological reasoning.

IV. CONCLUSION

Yokota,M. has analyzed a considerable number of spatiotemporal event terms over various kinds of English words such as prepositions, verbs, adverbs, etc. categorized as 'Dimensions', 'Form' and 'Motion' in the class 'SPACE' of the Roget's thesaurus [9], and found that almost all the concepts of spatiotemporal event terms can be defined in exclusive use of six kinds of attributes for FAOs, namely, 'Physical location (A12)', 'Direction (A13)', 'Trajectory (A15)', 'Velocity (A16)', 'Mileage (A17)' and 'Topology (A44)'. This fact implies that $\boldsymbol{L_{md}}$ expression can control robotic attention mechanism very efficiently in a top-down way in the physical world.

REFERENCES

[1] Yokota M (2005), 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.

[2] Yokota M & Capi G (2005), Cross-media Operations between Text and Picture Based on Mental Image Directed Semantic Theory. WSEAS Trans. on Information Science and Applications, 10-2:1541-1550

[3] Yokota,M., Sugita,K. & Oka,T. (2008), Natural language understanding based on mental image description language $\boldsymbol{L_{md}}$ and its application to language-centered robot manipulation. Artificial Life and Robotics, 13:84-88

[4] Yokota,M. (2008), In Lazinica,A. (Ed.), Humanoid Robots, I-Tech Education and Publishing, 333-362

[5] Coradeschi, S. & Saffiotti, A.: "An introduction to the anchoring problem", Robotics and Autonomous Systems, 43, pp.85–96, 2003.

[6] Drumwright,E. Ng-Thow-Hing,V. & Mataric,M.J. "Toward a vocabulary of primitive task programs for humanoid robots", *Proceedings of International Conference on Development and Learning (ICDL06),* Bloomington IN, May 2006.

[7] Sowa JF (2000), Knowledge Representation: Logical, Philosophical, and Computational Foundations, Brooks Cole Publishing Co., Pacific Grove, CA

[8] Miller GA & Johnson-Laird PN (1976), Language and Perception, Harvard University Press

[9] Roget P (1975), Thesaurus of English Words and Phrases, J.M.Dent & Sons Ltd., London