Integrated Multimedia Understanding for Ubiquitous Intelligence Based on Mental Image Directed Semantic Theory

Masao Yokota.
Fukuoka Institute of Technology
1 Research 3-30-1 Wajiro-higashi, Higashi-ku, Fukuoka-shi, JAPAN 881-0279
yokota@fit.ac.jp
(Dated: December 2006)

Contents

1. INTRODUCTION

At present, the realization of wireless sensor and actor network (WSAN) is one of the challenging topics in the concerned research fields, and a considerable number of important issues have been proposed especially from the viewpoint of networking [1-3]. From the viewpoint of Artificial Intelligence and Cognitive Robotics [18], a WSAN can be considered as an intelligent robot system with distributed sensors and actuators that can gather information of high density and perform appropriate actions upon its environment over wide areas. Furthermore, as an ideal ubiquitous computing environment, a WSAN must be intelligent enough to work autonomously and to interact with ordinary people when their aids are needed. In order to realize such an environment, we have proposed the concept of distributed intelligent robot network (DIRN [5]).

As shown in FIG.1, a DIRN is one kind of wireless sensor and actor network (WSAN), consisting of one brain node and numerous sensor and actor nodes with human-friendly interfaces. It is assumed, for example, that sensors and actuators can collaborate autonomously to perform appropriate actions just like reflexive actions in humans and that the brain node works exclusively for complicated computation based on profound knowledge in

order to control the other kinds of nodes, to communicate with people, etc.

In order to realize well-coordinated DIRNs, it is very important to develop a systematically computable knowledge representation language (KRL) [5, 6, 7] as well as efficient networking technologies [3]. This type of language is indispensable to *knowledge-based* processing such as *understanding* sensory events, *planning* appropriate actions and *knowledgeable* communication even with humans, 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.

Conventionally, such quasi-natural language expressions as 'move(10meters)', 'find(box, red)' and so on, uniquely related to computer programs, were employed for deploying sensors/ motors in robotic systems [e.g., 8, 27]. These kinds of expressions, however, were very specific to devices and apt to have miscellaneous syntactic variants among them such as 'move(*Distance*, *Speed*)', 'move(*Speed*, *Distance*, *Direction*)', etc. for motors and 'find(*Object*, *Color*)', 'find(*Object*, *Shape*, *Color*)', etc. for sensors. This is very inconvenient for communications especially between devices unknown to each other and therefore it is very important to develop such a language as is universal among all kinds of equipments.

Yokota, M. has proposed a semantic theory for natural languages so called 'Mental Image Directed Semantic Theory (MIDST)' [9, 10]. In the MIDST, word concepts are associated with omnisensory mental images of the external or physical world and are formalized in an intermediate language L_{md} [9]. This language is employed for many-sorted predicate logic with five types of terms. The most remarkable feature of L_{md} is its capability of formalizing both temporal and spatial event concepts on the level of human sensations while the other similar knowledge representation languages are designed to describe the logical relations among conceptual primitives represented by natural-language words [11-13] or formally defined tokens [14-16].

The language L_{md} was originally proposed for formalizing the natural semantics, that is, the semantics specific to humans, but it is general enough for the artificial semantics, that is, the semantics specific to each artificial device such as robot. This language has already been implemented on several types of computerized intelligent systems [5, 17, 18, 19] and there is a feedback loop between them for their mutual refinement, unlike other similar ones [20, 21].

This paper presents the concept of DIRN and a sketch of the formal language L_{md} and focuses on the semantic processing of multimedia information represented in L_{md}, simulating the interactions between robots and their environments including humans.

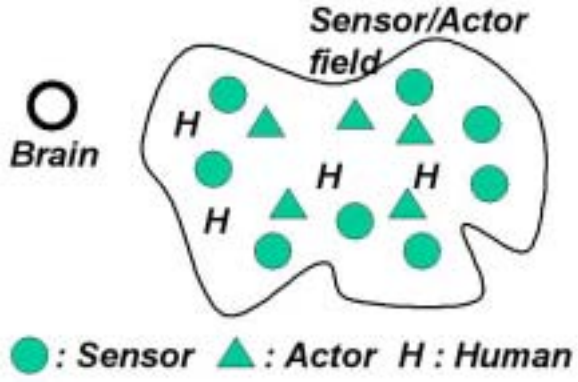


FIG.1 Physical architecture of DIRN

2. MENTAL IMAGE DESCRIPTION LANGUAGE L_{md}

2.1. Omnisensory image model

In the MIDST, word meanings are treated in association with mental images, not limited to visual but omnisensory, modeled 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 (up) 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" with the duration $[t_i, t_f]$ as depicted in FIG.2 (down) and formulated as (1).

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

This is a formula in many-sorted predicate logic, where "L" is a predicate constant with five types of terms: "Matter" (at 'x' and 'y'), "Attribute Value" (at 'p' and 'q'), "Attribute" (at 'a'), "Event Type" (at 'g') and "Standard" (at 'k'). Conventionally, Matter variables are headed by 'x',

'y' and 'z'. This formula is called 'Atomic Locus Formula' whose first two arguments are sometimes referred to as 'Event Causer (EC)' and 'Attribute Carrier (AC)', respectively while ECs are often optional in natural concepts such as intransitive verbs. For simplicity, the syntax of L_{md} allows Matter terms (e.g., 'Tokyo' and 'Osaka' in (2) and (3)) to appear at Attribute Values or Standard in order to represent their values at the time. Moreover, when it is not so significant to discern ECs or Standards, anonymous variables, usually symbolized as '_', can be employed in their places (See (39)).

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=Gt* and *g=Gs*, 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 'A12' denotes the attribute 'Physical Location'. These two formulas are different only at the term 'Event Type'.

(S1) The bus runs from Tokyo to Osaka.

$$(\exists x,y,k)L(x,y,Tokyo,Osaka,A12,Gt,k)\wedge bus(y) \qquad (2)$$

(S2) The road runs from Tokyo to Osaka.

$$(\exists x,y,k)L(x,y,Tokyo,Osaka,A12,Gs,k)\wedge road(y) \qquad (3)$$

A considerable number of works [e.g., 23-26] have shown that human active sensing processes may affect perception and in turn conceptualization and recognition of the physical world. 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). 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, 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, this hypothesis can be rephrased that ***all loci in attribute spaces correspond one to one with movements or, more generally, temporal events of the FAO***.

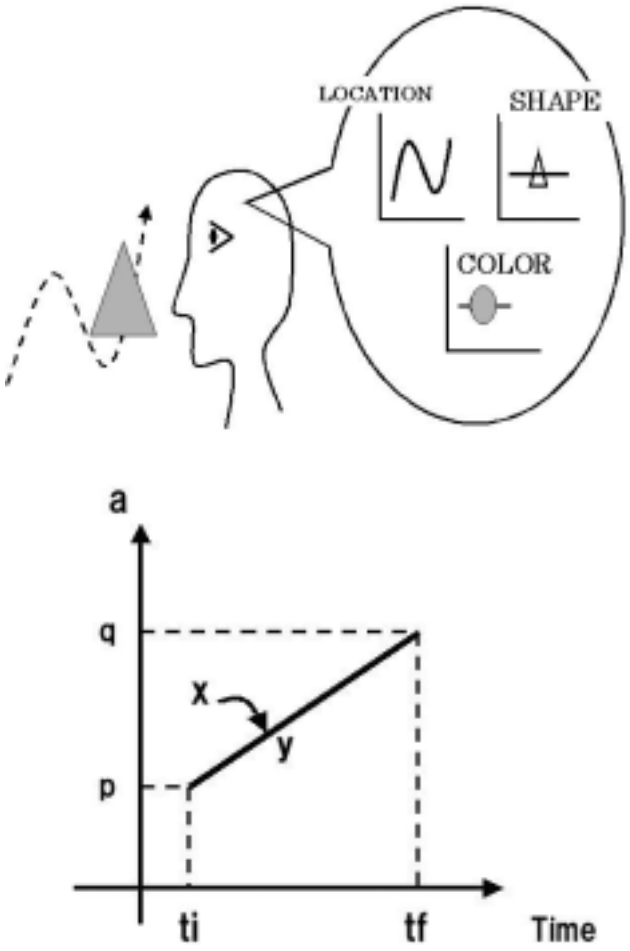


FIG.2 Mental image model (up) and Atomic Locus in Attribute Space (down)

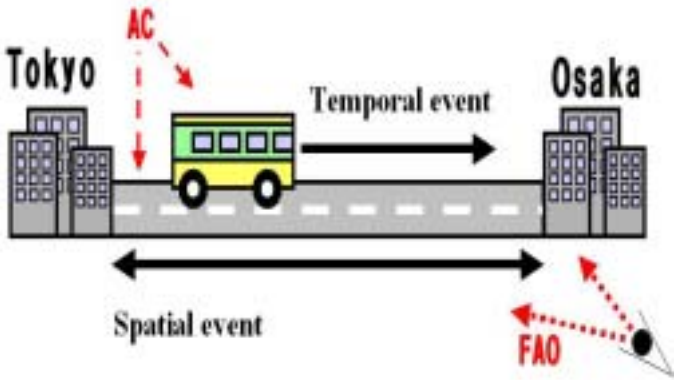


FIG.3 FAO movements and Event types

2.2. Tempo-logical connectives

The duration of an atomic locus, suppressed in the atomic locus formula, corresponds to the time-interval over which the FAO is put on the corresponding phenomenon outside. The MIDST has employed 'tempo-logical connectives (TLCs)' representing both logical and temporal relations between loci at a time.

A tempo-logical connective K_i is defined by (4), where τ_i, χ and K refer to one of the temporal relations indexed by an integer 'i', a locus, and an ordinary binary logical connective such as the conjunction '$\wedge$', respectively. This is more natural and economical than explicit indication of time intervals, considering that people do not consult chronometers all the time in their daily lives. The definition of τ_i ($-6 \leq i \leq 6$) is given in Table I from which the theorem (5) can be deduced. This table shows 13 types of temporal relations between two events χ_1 and χ_2 whose durations are $[t_{11},t_{12}]$ and $[t_{21},t_{22}]$, respectively. This is in accordance with Allen's notation [14], which, to be strict, is exclusively for 'temporal conjunctions ($=\wedge_i$)' such as introduced below.

$$\chi_1 \, K_i \, \chi_2 \leftrightarrow (\chi_1 \, K \, \chi_2) \wedge \tau_i(\chi_1, \chi_2) \quad (4)$$

$$\tau_{-i}(\chi_2, \chi_1) \equiv \tau_i(\chi_1, \chi_2) \quad (\forall i \in \{0,\pm1,\pm2,\pm3,\pm4,\pm5,\pm6\}) \quad (5)$$

The TLCs used most frequently are 'SAND ($\wedge_0$)' and 'CAND ($\wedge_1$)', standing for 'Simultaneous AND' and 'Consecutive AND' and conventionally symbolized as 'Π' and '$\bullet$', respectively. For example, the concepts of the English verbs 'carry' and 'return' are to be defined as (6) and (7), respectively. These formulas can be depicted as FIG.4 (up) and (down), where the optional ECs can be omitted as shown in FIG.5 (up) and (down), respectively.

$$(\lambda x,y)\mathrm{carry}(x,y) \leftrightarrow (\lambda x,y)(\exists p,q,k)\mathrm{L}(x,x,p,q,\mathrm{A12},\mathrm{Gt},k)\Pi \mathrm{L}(x,y,p,q,\mathrm{A12},\mathrm{Gt},k) \wedge x \neq y \wedge p \neq q \quad (6)$$

$$(\lambda x)\mathrm{return}(x) \leftrightarrow (\lambda x)(\exists p,q,k)\mathrm{L}(x,x,p,q,\mathrm{A12},\mathrm{Gt},k)\bullet \mathrm{L}(x,x,p,q,\mathrm{A12},\mathrm{Gt},k) \wedge x \neq y \wedge p \neq q \quad (7)$$

The expression (8) is the definition of the English verb concept 'fetch' depicted as FIG.6 (up). This implies such a temporal event that 'x' goes for 'y' and then comes back with it.

$$(\lambda x,y)\mathrm{fetch}(x,y) \leftrightarrow (\lambda x,y)(\exists p_1,p_2,k)\mathrm{L}(x,x,p_1,p_2,\mathrm{A12},\mathrm{Gt},k)\bullet ((\underline{\mathrm{L}(x,x,p_2,p_1,\mathrm{A12},\mathrm{Gt},k)\Pi \mathrm{L}(x,y,p_2,p_1,\mathrm{A12},\mathrm{Gt},k)}) \wedge x \neq y \wedge p_1 \neq p_2 \quad (8)$$

In the same way, the English verb concept 'hand' or 'receive' depicted as FIG.6 (down) is defined uniformly as (9) or its abbreviation (9').

$$(\lambda x,y,z)\mathrm{hand}(x,y,z) \leftrightarrow (\lambda x,y,z)\mathrm{receive}(z,,y,x) \leftrightarrow (\lambda x,y,z)(\exists k)\mathrm{L}(x,y,x,z,\mathrm{A12},\mathrm{Gt},k)\Pi \mathrm{L}(z,y,x,z,\mathrm{A12},\mathrm{Gt},k) \wedge x \neq y \wedge y \neq z \wedge z \neq x \quad (9)$$

$$.\equiv.(\lambda x,y,z)(\exists k)\mathrm{L}(\{x,z\},y,x,z,\mathrm{A12},\mathrm{Gt},k) \wedge x \neq y \wedge y \neq z \wedge z \neq x \quad (9')$$

Such locus formulas as correspond with natural event concepts are called 'Event Patterns' and about 40 kinds of event patterns have been found concerning the attribute 'Physical Location (A12)', for example, *start*, *stop*, *meet*, *separate*, *carry*, *return*, etc [10].

TABLE I: List of temporal relations.

$\tau_i(\chi_1, \chi_2)$	Allen's notation	Definition
$\tau_0(\chi_1, \chi_2)$	equals(χ_1,χ_2)	$t_{11}=t_{21}$
$\tau_0(\chi_2, \chi_1)$	equals(χ_2, χ_1)	$\wedge t_{12}=t_{22}$
$\tau_1(\chi_1, \chi_2)$	meets(χ_1, χ_2)	$t_{12}=t_{21}$
$\tau_{-1}(\chi_2, \chi_1)$	met-by(χ_2, χ_1)	
$\tau_2(\chi_1, \chi_2)$	starts(χ_1, χ_2)	$t_{11}=t_{21}$
$\tau_{-2}(\chi_2, \chi_1)$	started-by(χ_2, χ_1)	$\wedge t_{12}<t_{22}$
$\tau_3(\chi_1, \chi_2)$	during(χ_1, χ_2)	$t_{11}>t_{21}$
$\tau_{-3}(\chi_2, \chi_1)$	contains(χ_2, χ_1)	$\wedge t_{12}<t_{22}$
$\tau_4(\chi_1, \chi_2)$	finishes(χ_1, χ_2)	$t_{11}>t_{21}$
$\tau_{-4}(\chi_2, \chi_1)$	finished-by(χ_2, χ_1)	$\wedge t_{12}=t_{22}$
$\tau_5(\chi_1, \chi_2)$	before(χ_1, χ_2)	$t_{12}<t_{21}$
$\tau_{-5}(\chi_2, \chi_1)$	after(χ_2, χ_1)	
$\tau_6(\chi_1, \chi_2)$	overlaps(χ_1, χ_2)	$t_{11}<t_{21} \wedge t_{21}<t_{12}$
$\tau_{-6}(\chi_2, \chi_1)$	overlapped-by(χ_2, χ_1)	$\wedge t_{12}<t_{22}$

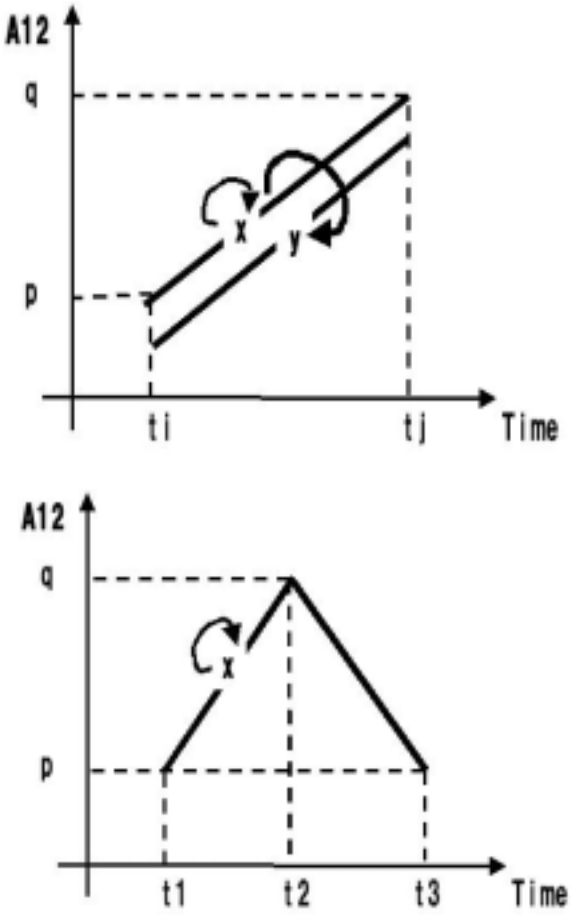
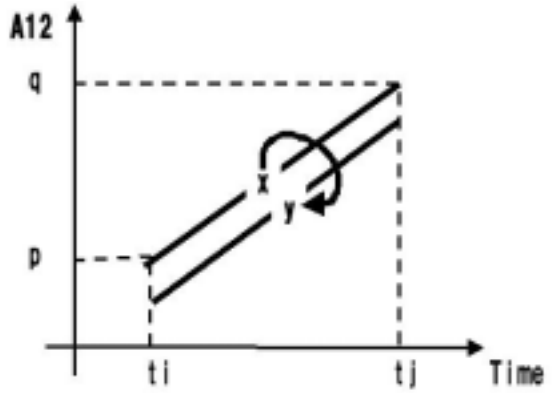


FIG.4. Depictions of loci: 'carry' (up) and 'return' (down)



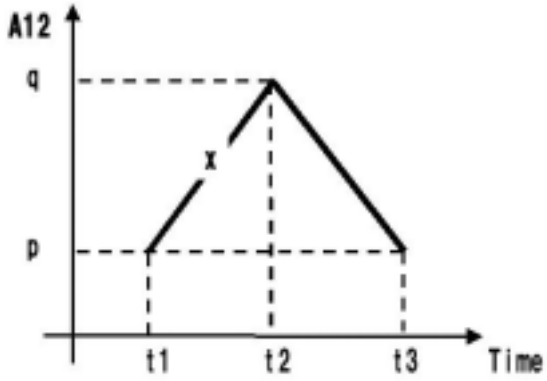
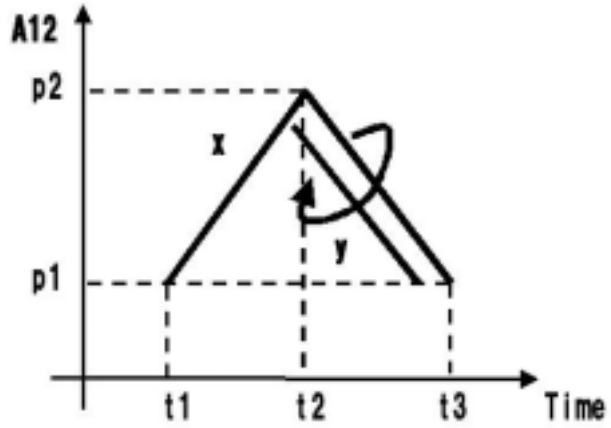


FIG.5 Simplified depictions of loci: 'carry' (up) and 'return' (down)



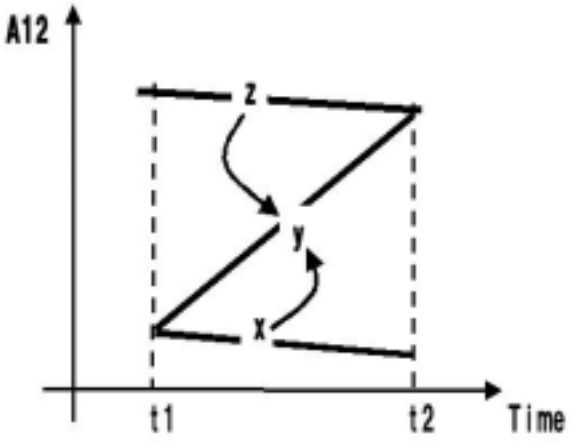


FIG.6 Loci of 'fetch' (up) and 'hand/receive' (down)

Furthermore, a very important concept called 'Empty Event (EE)' and denoted by 'ε' is introduced. An EE stands for nothing but for time elapsing and is explicitly defined as (10) 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 (11).

$$\varepsilon([t_1,t_2]) \leftrightarrow (\exists x,y,g,k)\ L(x,y,t_1,t_2,A34,g,k) \qquad (10)$$

$$\chi\ \Pi\ \varepsilon([t_a, t_b]) \qquad (11)$$

Any pair of loci temporally related in certain attribute spaces can be formulated as (12)-(16) in exclusive use of SANDs, CANDs and EEs. For example, the loci shown in FIG.7 (up) and (down) correspond to the formulas (13) and (16), respectively.

$$\chi_1 \wedge_2 \chi_2 \ .\equiv.\ (\chi_1 \bullet \varepsilon)\Pi\chi_2 \qquad (12)$$

$$\chi_1 \wedge_3 \chi_2 \ .\equiv.\ (\varepsilon_1 \bullet \chi_1 \bullet \varepsilon_2)\Pi\chi_2 \qquad (13)$$

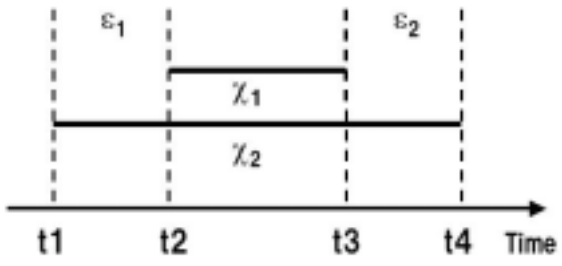
$$\chi_1 \wedge_4 \chi_2 \ .\equiv.\ (\varepsilon \bullet \chi_1)\Pi\chi_2 \qquad (14)$$

$$\chi_1 \wedge_5 \chi_2 \ .\equiv.\ \chi_1 \bullet \varepsilon \bullet \chi_2 \qquad (15)$$

$$\chi_1 \wedge_6 \chi_2 \ .\equiv.\ (\chi_1 \bullet \varepsilon_3)\Pi(\varepsilon_1 \bullet \chi_2)\Pi(\varepsilon_1 \bullet \varepsilon_2 \bullet \varepsilon_3) \qquad (16)$$

Employing these TLCs, tempo-logical relationships between miscellaneous event concepts can be formulated without explicit indication of time intervals. For example, an event 'fetch(x,y)' is necessarily *finished by* an event 'carry(x,y)' as indicated by the underline at (8). This fact can be formulated as (17), where '$\supset_{-4}$' is the '*implication* ($\supset$)' furnished with the temporal relation '*finished-by* (τ_{-4})'. This kind of formula is not an axiom but a theorem deducible from the definitions of event concepts in our formal system.

$$\text{fetch}(x,y) \supset_{-4} \text{carry}(x,y) \qquad (17)$$



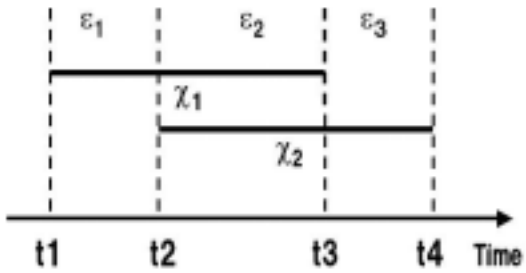


FIG.7 Tempo-logical relations: during(χ_1,χ_2) (up) and overlaps(χ_1,χ_2) (down)

2.3. Attributes and Standards

The attribute spaces for humans correspond to the sensory receptive fields in their brains. At present, about 50 attributes concerning the physical world have been extracted exclusively from English and Japanese words as shown in Table II. They are associated with all of the 5 senses (i.e. sight, hearing, smell, taste and feeling) in our everyday life while those for information media other than languages correspond to limited senses. For example, those for pictorial media, marked with '*' in Table II, associate limitedly with the sense 'sight' as a matter of course. The attributes of this sense occupy the greater part of all, which implies that the sight is essential for humans to conceptualize the external world by. And this kind of classification of attributes plays a very important role in our cross-media operating system [18].

Correspondingly, six categories of standards shown in Table III have been extracted that are assumed necessary for representing values of each attribute in Table II. In general, the attribute values represented by words are relative to certain standards as explained briefly in Table III. For example, (18) and (19) are different formulations of a locus due to the different standards 'k_1' and 'k_2' for scaling as shown in FIG.8-a and b, respectively. That is, whether the point (t_2, q) is significant or not, more generally, how to articulate a locus depends on the precisions or the granularities of these standards, which can be formulated as (20) and (21), so called, '***Postulate of Arbitrariness in Locus Articulation***'. This postulate affects the process of conceptualization on a word based on its referents in the world [32] and is applied in a DIRN as 'Data Interpretation Function (F_d)' that translates a set of sensory data into a locus formula at the precision of a Standard (See 3.2).

TABLE II: Examples of attributes

Code	Attribute[Property†]
*A01	PLACE OF EXISTENCE [N]
*A02	LENGTH [S]
*A03	HEIGHT [S]
*A04	WIDTH [S]
*A05	THICKNESS [S]
*A06	DEPTH1 [S]
*A07	DEPTH2 [S]
*A08	DIAMETER [S]
*A09	AREA [S]
*A10	VOLUME [S]
*A11	SHAPE [N]
*A12	PHYSICAL LOCATION [N]
*A13	DIRECTION [N]
*A14	ORIENTATION [N]
*A15	TRAJECTORY [N]
*A16	VELOCITY [S]
*A17	MILEAGE [S]
A18	STRENGTH OF EFFECT [S]
A19	DIRECTION OF EFFECT [N]
A20	DENSITY [S]
A21	HARDNESS [S]
A22	ELASTICITY [S]
A23	TOUGHNESS [S]
A24	FEELING [S]
A25	HUMIDITY [S]
A26	VISCOSITY [S]
A27	WEIGHT [S]
A28	TEMPERATURE [S]
A29	TASTE [N]
A30	ODOUR [N]
A31	SOUND [N]
*A32	COLOR [N]
A33	INTERNAL SENSATION [N]
A34	TIME POINT [S]
A35	DURATION [S]
A36	NUMBER [S]
A37	ORDER [S]
A38	FREQUENCY [S]
A39	VITALITY [S]
A40	SEX [S]
A41	QUALITY [N]
A42	NAME [V]
A43	CONCEPTUAL CATEGORY [V]
*A44	TOPOLOGY [V]
*A45	ANGULARITY [S]
	……

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

TABLE VI: 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*'.

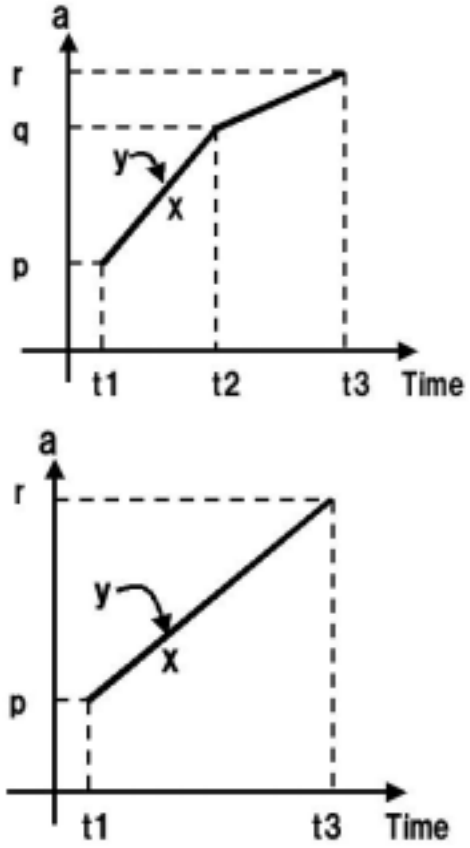


FIG. 8 Arbitrariness in locus articulation due to standards: Standard *k1* (up) is finer than *k2* (down)

$$(L(y,x,p,q,a,g,k_1)\ \Pi\ \varepsilon([t_1,t_2]))\bullet(L(y,x,q,r,a,g,k_1)\ \Pi\varepsilon([t_2,t_3])) \quad (18)$$

$$L(y,x,p,r,a,g,k_2)\Pi\ \varepsilon([t_1,t_3]) \quad (19)$$

$$(\forall p,q,r,k)(L(y,x,p,q,a,g,k)\bullet L(y,x,q,r,a,g,k).\supset.\ (\exists k')L(y,x,p,r,a,g,k')\wedge k'\neq k) \quad (20)$$

$$(\forall p,r,k)(L(y,x,p,r,a,g,k).\supset.\ (\exists q,k')L(y,x,p,q,a,g,k')\bullet L(y,x,q,r,a,g,k')\wedge k'\neq k) \quad (21)$$

2.4. Attribute Values and Atomic Loci

In our formal system, a constant term of Attribute Value is to be assigned a point set, possibly, with a fuzzy boundary due to its semantic vagueness. This is the case especially when such a term is associated with a certain word concept and the boundary of its point set assigned is to be controlled by the Standard specific to the term. For example, a word of color such as "*red*" is semantically vague possibly depending on a certain Individual Standard and can be assigned a certain region in the attribute space, conventionally called 'Color Solid', with the three dimensions of 'Chrome', 'Hue' and 'Value'.

In general, the relation between a word of Attribute Value 'v_i' and its corresponding point set '$\mathcal{S}(p_i,a_i,k_i)$' can be formalized as (22) by the function '*Assign*', where 'p_i', 'a_i' and 'k_i' are 'Attribute Value', 'Attribute' and 'Standard' specific to 'v_i', respectively. For example, the word "*long* (=v_i)" can be assigned a point set '$\{p_i \mid p_i > k_i\}$', where k_i is some Standard for being long.

$$Assign(v_i)=\mathcal{S}(p_i,a_i,k_i) \quad (22)$$

According to this assumption of Attribute Value, the formal interpretation of an atomic locus formula such as underlined in (23) can be given as follows.

$$\underline{L(x,y,p_1,p_2,a,g,k)}\Pi\varepsilon(\ [t_1,t_2]) \quad (23)$$

Firstly, the real sensations and the terms for them are related as (24), where 'Vsense(y,a,t_j)' is the region in the attribute space 'a' *onto* which the real sensation of 'y' is projected at the time t_j.

$$Vsense(y,a,t_1)\subseteq\mathcal{S}(p_1,a,k),\ Vsense(y,a,t_2)\subseteq\mathcal{S}(p_2,a,k) \quad (24)$$

Secondly, the scan path of the FAO from V_1 to V_2, 'Fpath(V_1,V_2)', is related to the bundle of the shortest paths from S_1 to S_2, 'Pbundle(S_1,S_2)', as shown in FIG.9 and formulated by (25) and (26), where 'Spath(q_1,q_2)' is the shortest path from q_1 to q_2.

$$Fpath(Vsense(y,a,t_1),\ Vsense(y,a,t_2))\subseteq Pbundle(\mathcal{S}(p_1,a,k),\ \mathcal{S}(p_2,a,k)) \quad (25)$$

$$Pbundle(S_1,S_2)=\{Spath(q_1,q_2)\mid\forall q_1\in S_1,\ \forall q_2\in S_2\} \quad (26)$$

As easily imagined, 'Fpath' corresponds with 'Atomic Locus' without the causation (i.e., 'x→y') and the paths are to follow the structure specific to the attribute space. For example, the topology between two regions can be represented as a node of such a graph as shown in FIG.10 [28], where the Fpath from the node "*inside*" to the node "*meet*" must pass the nodes "*coveredBy*" and "*overlap*" on the way of such monotonic translation as shown in FIG.11 without any other deformation.

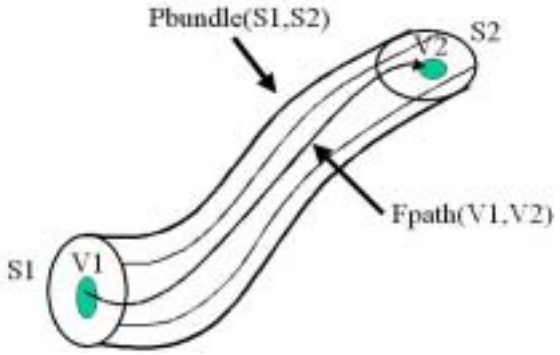


FIG.9 Fpath as a member of Pbundle.

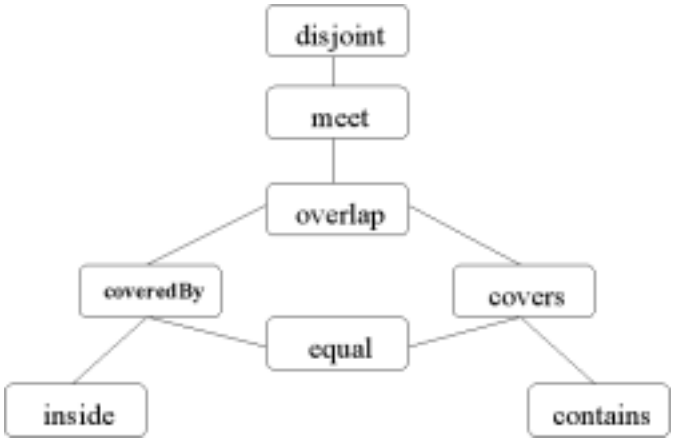


FIG.10 The Closest-Topological-Relationship-Graph [28].

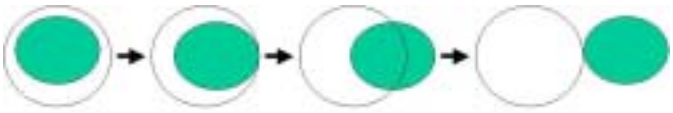


FIG.11 Monotonic change from 'inside' to 'meet' by translation.

3. SPECIFICATION OF DIRN'S WORLD

'The world for a DIRN' (***W***) refers to 'the set of matters observable for the DIRN' and is defined by (27) as the union of the set of its nodes (***D***) and the set of the objects in its environment. The set ***D*** is the union of the sets of a brain node ({B}), sensor nodes (***Se***) and actor nodes (***Ac***) as represented by (28) while the set ***O*** includes possibly humans and the other DIRNs.

$$W=D\cup O \tag{27}$$

$$D=\{B\}\cup Se\cup Ac \tag{28}$$

'A constituent C_k of the world for a DIRN' (i.e., $C_k\in W$) can be specified by the loci in the attribute spaces distinguishable by the sets of Attributes and Standards unique to the DIRN.

3.1. Specification of objects

An object in the environment of a DIRN (i.e., $C_k\in O$) can be characterized by the loci of its structure and so on. For example, the characteristics of a tree 'C1' in the environment can be represented by such a locus formula as (29), reading its height (A03) is between 4m and 5m, its location (A12) is in the park 'C2',…. For another example, a road 'C3' that runs from a town 'C4' to a town 'C5' via a town 'C6' can be defined by (30).

$$\begin{aligned}&tree(C1)\leftrightarrow(\exists x,p,k,\ldots)L(x,C1,p,p,A03,Gt,Me)\\&\quad\wedge(4m\le p\le 5m)\wedge L(x,C1,C2,C2,A12,Gt,k)\\&\quad\wedge park(C2)\wedge\ldots\end{aligned} \tag{29}$$

$$\begin{aligned}&road(C3)\leftrightarrow(\exists x,k,\ldots)L(x,C3,C4,C6,A12,Gs,k)\bullet\\&\quad L(x,C3,C6,C5,A12,Gs,k)\wedge town(C4)\wedge town(C5)\\&\quad\wedge town(C6)\end{aligned} \tag{30}$$

3.2. Specification of a sensor node

A sensor node (i.e., $C_k\in Se$) can be specified by the loci of its structure and its collectable sensory data. In general, a sensor can be distinguished by the definition (31) from another kind of constituent, reading that a sensor 'x' is what takes in some data 'y' from some constituent. A data set is to be translated into a locus formula by 'Data Interpretation Function (F_d)' (See 8.2) as defined by (32).

$$\begin{aligned}&(\lambda x)sensor(x)\leftrightarrow(\lambda x)(\exists y,z,g_1,k_1)\ L(x,y,z,x,A12,g_1,k_1)\\&\quad\wedge data(y)\end{aligned} \tag{31}$$

$$\begin{aligned}&F_d(y)=(\exists z,z_1,\ldots,z_n,a,g,k,p_0,\ldots,p_n)L(z_1,z,p_0,p_1,a,g,k)\bullet\ldots\bullet\\&\quad L(z_n,z,p_{n-1},p_n,a,g,k)\end{aligned} \tag{32}$$

The left hand of (32) is given as such a locus formula as characterized by the attribute 'a' and the standard 'k' unique to the sensor. For example, a thermometer with the measurable range [-10°C, +100°C] can be characterized by (33) with the attribute 'temperature (A28)' and the rigid standard of 'Celsius (Ce)'.

$$\begin{aligned}&(\exists z,z_1,\ldots,z_n,p_1,\ldots,p_n)L(z_1,z,p_0,p_1,A28,Gt,Ce)\bullet\ldots\bullet\\&L(z_n,z,p_{n-1},p_n,A28,Gt,Ce)\wedge(-10^{o}C\le p_i\le+100^{o}C\\&\wedge 0\le i\le n)\end{aligned} \tag{33}$$

3.3. Specification of an actor node

An actor (i.e., $C_k\in Ac$) can be specified by the loci of its structure, performable actions and, if any sensors with it, collectable sensory data. For example, a tanker 'C8' with the coverage [0km,100km] can be characterized by (34) with the attribute 'mileage (A17)' at the rigid standard of 'Meter (Me)'.

$$\begin{aligned}&(\exists x,p)L(C8,x,0,p,A17,Gt,Me)\wedge(0km\le p\le 100km)\\&\wedge liquid(x)\end{aligned} \tag{34}$$

3.4. Specification of the brain node

The brain node (i.e., *B*) can be specified by its commonsense knowledge and world knowledge including such specifications of the other constituents as mentioned above. For example, (35) is an example of commonsense knowledge piece, reading that ***a matter has never different values of an attribute at a time***.

$$L(x,y,p_1,q_1,a,g,k)\Pi L(z,y,p_2,q_2,a,g,k).\supset.p_1=p_2\wedge q_1=q_2 \tag{35}$$

The intelligence of the brain node must be conscious of all about the other constituents but can be unconscious of its own structure (e.g., hardware configuration) and computational performance specification (e.g., CPU speed) because they are what only meta-systems such as OS and meta-brain node have to concern. In our case, the brain node is a personal computer with the intelligent system IMAGES-M [4] installed under the OS WINDOWS/XP.

4. INTERACTION BETWEEN DIRN AND ITS WORLD

4.1 Intelligent system IMAGES-M

As shown in FIG.12, a DIRN is to gather information pieces from its world, interpret them into locus formulas, find/solve problems, and act appropriately upon its field. All these intelligent performances are executed by IMAGES-M possibly with aids of humans. The intelligent system IMAGES-M [4], still under development, is intended to facilitate integrated multimedia information understanding, including miscellaneous cross-media operations. This system has employed locus formula as intermediate knowledge representation, through which it can integrally understand and generate sensor data, speech, visual image, text, and action data. IMAGES-M is to work as the main intelligence of the brain node of a DIRN while the intelligence of each sensor or actuator is a small-scaled IMAGES-M adapted for its specialized function.

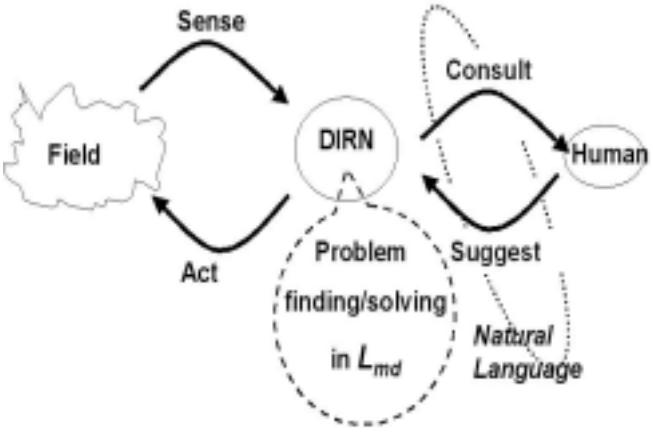


FIG.12 Interaction between DIRN and its world.

IMAGES-M, as shown in FIG.13, 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), Image Processing Unit (IPU), 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.

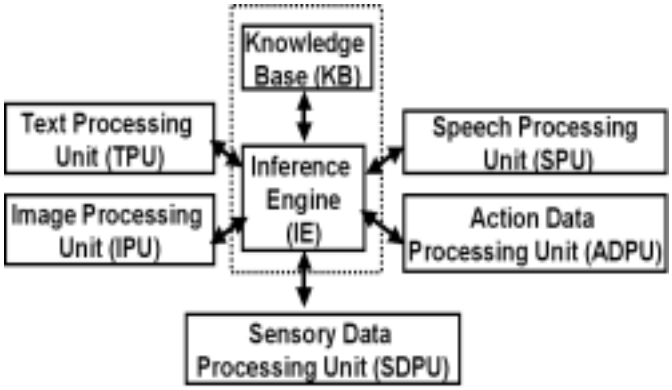


FIG.13. Configuration of IMAGES-M

4.2. Fundamental computations on L_{md}

The fundamental computations on L_{md} by IMAGES-M are to detect semantic anomalies, ambiguities and identities in data or expressions. These are performed as inferential operations on locus formulas at IE.

Detection of semantic anomalies is very important to avoid succession of meaningless computations or actions. For an extreme example, consider such a report from certain sensors as (36) represented in L_{md}, where '…' and 'A29' stand for descriptive omission and the attribute 'Taste'. This locus formula can be translated into the English sentence S3 by TPU, but it is semantically anomalous because a 'desk' has ordinarily no taste.

$$(\exists x,y,k)L(y,x,Sweet,Sweet,A29,Gt,k)\wedge desk(x) \qquad (36)$$

(S3) The desk is sweet.

These kinds of semantic anomalies can be detected in the following processes.

Firstly, assume the postulate (37) as the commonsense or default knowledge of "desk", stored in KB, where 'A39' is

the attribute 'vitality'. The special symbol '*' represents 'always' as defined by (38), where '$\varepsilon([t_1,t_2])$' is a simplified atomic locus formula standing for time elapsing with an interval $[t_1,t_2]$. Furthermore, '_' and '/' are anonymous variables employed for descriptive simplicity and defined by (39) and (39'), respectively.

$$(\lambda x)\mathrm{desk}(x)\leftrightarrow(\lambda x)(\ldots L^*(\_,x,/,/,A29,Gt,\_)\wedge\ldots\wedge L^*(\_,x,/,/,A39,Gt,\_)\wedge\ldots) \quad (37)$$

$$X^*\leftrightarrow(\forall[t_1,t_2])X\Pi\,\varepsilon([t_1,t_2]) \quad (38)$$

$$X(\_)\leftrightarrow(\exists u)\ X(u) \quad (39)$$

$$X(/)\leftrightarrow\neg(\exists u)\ X(u) \quad (39')$$

Secondly, the postulates expressed by (40) and (41) in KB are utilized. The formula (40) means that ***if one of two loci exists every time interval, then they can coexist***. The formula (41) states that ***a matter has never different values of an attribute at a time***.

$$X\wedge Y^*\ .\supset.\ X\Pi Y \quad (40)$$

$$L(x,y,p,q,a,g,k)\Pi L(z,y,r,s,a,g,k).\supset.p=r\wedge q=s \quad (41)$$

Lastly, IE detects the semantic anomaly of "sweet desk" by using (37)-(41). That is, the formula (42) below is finally deduced from (37)-(41), which violates the postulate (37), that is, " $Sweet \neq /$ ".

$$L(\_,x,Sweet,Sweet,A29,Gt,\_)\Pi L(z,x,/,/,A29,Gt,\_) \quad (42)$$

These processes above are also employed for dissolving syntactic ambiguities in people's utterances such as S4. IE rejects 'sweet desk' and eventually adopts 'sweet coffee' as a plausible interpretation.

(S4) Bring me the coffee on the desk, which is very sweet.

If multiple plausible interpretations of a text or another type of information are represented in different locus formulas, it is semantically ambiguous. In such a case, IMAGES-M will ask for further information in order for disambiguation.

Furthermore, if two different representations are interpreted into the same locus formula, they are paraphrases of each other. Such detection of semantic identities is very useful for deleting redundant information, for cross-media translation, etc. [4].

5. PROBLEM FINDING AND SOLVING BY DIRN

5.1. Definition of problem and task for DIRN

The problems for a DIRN can be classified roughly into two categories as follows.

(CP) Creation Problem:
e.g.) house building, food cooking, etc.

(MP) Maintenance Problem:
e.g.) fire extinguishing, room cleaning, etc.

In general, an MP is relatively simple one that the DIRN can find and solve autonomously while a CP is relatively difficult one that is given to the DIRN, possibly, by humans and to be solved in cooperation with them. A DIRN must determine its task to solve a problem in the world.

The conventional AI defines a problem 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 DIRN needs to interpolate some transit event X_T between the two events, namely, 'Current Event (X_C)' and 'Goal Event (X_G)' as (43).

$$X_C\bullet X_T\bullet X_G \quad (43)$$

According to this formalization, a problem X_P can be defined as $X_T\bullet X_G$ and a task for the DIRN can be defined as its realization.

The events in the world are described as loci in certain attribute spaces and a problem is to be detected by the unit of atomic locus. For example, employing such a postulate as (44) implying 'Continuity in attribute values', the event X in (45) is to be inferred as (46).

$$L(x,y,p_1,p_2,a,g,k)\bullet L(z,y,p_3,p_4,a,g,k).\supset.p_3=p_2 \quad (44)$$

$$L(x,y,p_1,p_2,a,g,k)\bullet X\bullet L(z,y,p_3,p_4,a,g,k) \quad (45)$$

$$L(z',y,p_2,p_3,a,g,k) \quad (46)$$

5.2. CP finding and solving

Consider a verbal command such as S3 uttered by a human. Its interpretation is given by (47) as the goal event

X_G. If the current event X_C is given by (48), then (49) with the transit event X_T underlined can be inferred as the problem corresponding to S5.

(S5) Keep the temperature of 'room C9' at 20.

$$L(z,C9,20,20,A28,Gt,k)\wedge room(C9)\wedge(z\in \mathcal{O}) \quad (47)$$

$$L(x,C9,p,p,A28,Gt,k) \wedge room(C9) \quad (48)$$

$$\underline{L(z1,C9,p,20,A28,Gt,k)}\bullet\ L(z,C9,20,20,A28,Gt,k) \wedge room(C9) \wedge (z_1\in \mathcal{O}) \quad (49)$$

For this problem, the DIRN is to execute a job deploying a certain thermometer and actors 'z_1' and 'z'. The selection of the actor 'z_1' is performed as follows:

If 20-p <0 then z_1 is a cooler, otherwise
if 20-p >0 then z_1 is a heater, otherwise
20-p =0 and no actor is deployed as z_1.

The selection of 'z' is a job in case of MP described in the next section.

5.3. MP finding and solving

In general, the goal event X_G for an MP is that for another CP such as S3 given possibly by humans and solved by the DIRN in advance. That is, the job in this case is to autonomously restore the goal event X_G created in advance to the current event X_C as shown in (50), where the transit event X_T is the reversal of such X-$_T$ that has been already detected as 'abnormal' by the DIRN.

For example, if X_G is given by (47) in advance, then X_T is also represented as the underlined part of (49) while X-$_T$ as (51). Therefore the job here is quite the same that was described in the previous section.

$$X_G \bullet X\text{-}_T \bullet\ X_C \bullet X_T \bullet\ X_G \quad (50)$$

$$L(z_1,C9,20,p,A28,Gt,k) \wedge room(C9) \wedge (z_1\in \mathcal{O}) \quad (51)$$

6. NATURAL LANGUADE UNDERSTANDING

6.1 Word meaning description

Natural language is the most important information medium 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 picture. IMAGES-M can translate systematically natural language, either spoken or written, and $\boldsymbol{L_{md}}$ expression into each other by utilizing syntactic rules and word meaning descriptions of natural language [9].

A word meaning description M_w is given by (52) as a pair of 'Concept Part (C_p)' and 'Unification Part (U_p)'.

$$M_w \leftrightarrow [C_p; U_p] \quad (52)$$

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 (53).

$$[(\exists x,y,p_1,p_2,k)\ L(x,x,p_1,p_2,A12,Gt,k)\Pi$$
$$L(x,y,p_1,p_2,A12,Gt,k)\wedge x\neq y\wedge p_1\neq p_2\text{:}ARG(Dep.1,x);$$
$$ARG(Dep.2,y);] \quad (53)$$

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 (54) via the surface dependency structure shown in FIG.14.

$$(\exists y,p_1,p_2,k)L(Mary,Mary,p_1,p_2,A12,Gt,k)\Pi$$
$$L(Mary,y,p_1,p_2,A12,Gt,k)\wedge Mary\neq y\wedge p_1\neq p_2\wedge book(y) \quad (54)$$

For another example, the meaning description of the English preposition 'through' is also given by (55).

$$[(\exists x,y,p_1,z,p_3,g,k,p_4,k_0)(\underline{L(x,y,p_1,z,A12,g,k)}\bullet L(x,y,z,p_3,A12,g,k))\Pi\ L(x,y,p_4,p_4,A13,g,k_0)\wedge p_1\neq z\wedge z\neq p_3\text{:}ARG(Dep.1,z);$$
$$IF(Gov=Verb)\rightarrow PAT(Gov,(1,1));$$
$$IF(Gov=Noun)\rightarrow ARG(Gov,y);] \quad (55)$$

6.2. Mutual Conversion between text and locus formula

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 (55) and in general (i,j) refers to the partial formula covering from the ith to the jth 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 S6, S7 and S8 are interpreted as (56), (57) and (58), respectively. The underlined parts of these formulas are the results of PAT operations. The expression (59) 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 (59').

(S6) The train runs through the tunnel.

$$(\exists x,y,p_1,z,p_3,k,p_4,k_0)(\underline{L(x,y,p_1,z,A12,Gt,k)}\bullet L(x,y,z,p_3,A12,Gt,k))\Pi\ L(x,y,p_4,p_4,A13,Gt,k_0) \wedge p_1\neq z \wedge z\neq p_3 \wedge \text{train}(y) \wedge \text{tunnel}(z) \quad (56)$$

(S7) The path runs through the forest.

$$(\exists x,y,p_1,z,p_3,k,p_4,k_0)(\underline{L(x,y,p_1,z,A12,Gs,k)}\bullet L(x,y,z,p_3,A12,Gs,k))\Pi\ L(x,y,p_4,p_4,A13,Gs,k_0) \wedge p_1\neq z \wedge z\neq p_3 \wedge \text{path}(y) \wedge \text{forest}(z) \quad (57)$$

(S8) The path through the forest is long.

$$(\exists x,y,p_1,z,p_3,x_1,k,q,k_1,p_4,k_0)(L(x,y,p_1,z,A12,Gs,k)\bullet L(x,y,z,p_3,A12,Gs,k))\Pi\ L(x,y,p_4,p_4,A13,Gs,k_0)\ \wedge L(x_1,y,q,q,A02,Gt,k_1) \wedge p_1\neq z\wedge z\neq p_3\wedge q>k_1\wedge \text{path}(y)\wedge \text{forest}(z) \quad (58)$$

$$(\exists x_1,y_1,q,k_1)L(x_1,y_1,q,q,A02,Gt,k_1)\wedge q>k_1 \quad (59)$$

$$(\exists x_1,y_1,k_1)L(x_1,y_1,Long,Long,A02,Gt,k_1) \quad (59')$$

The process above is completely reversible except that multiple natural expressions as paraphrases can be generated as shown in FIG.15 because event patterns are sharable among multiple word concepts. This is one of the most remarkable features of IMAGES-M and is also possible between different languages as understanding-based translation [22].

(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.}
.....

FIG.15 Language to language translation

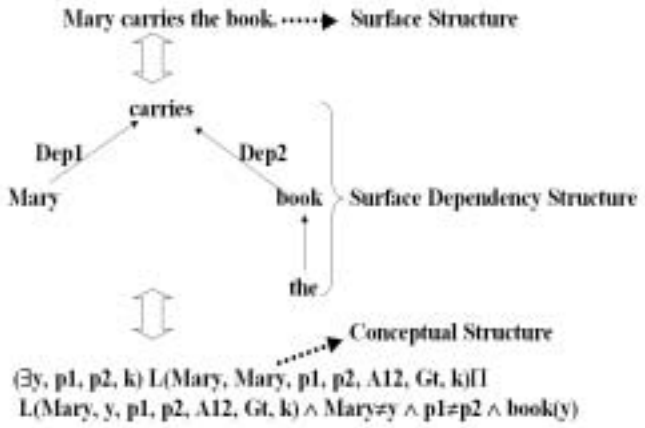


FIG.14 Mutual conversion between natural language and L_{md}

7. CROSS-MEDIA TRANSLATION

7.1. Functional requirements

The core technology for integrated multimedia information understanding in IMAGES-M is that for cross-media translation via intermediate representation in L_{md}. The author has considered that systematic cross-media translation must have such functions as follows.

(F1) To translate source representations into target ones as for contents describable by both source and target media. For example, positional relations between/among physical objects such as 'in', 'around' etc. are describable by both linguistic and pictorial media.

(F2) To filter out such contents that are describable by

source medium but not by target one. For example, linguistic representations of 'taste' and 'smell' such as 'sweet candy' and 'pungent gas' are not describable by usual pictorial media although they would be seemingly describable by cartoons, etc.

(F3) To supplement default contents, that is, such contents that need to be described in target representations but not explicitly described in source representations. For example, the shape of a physical object is necessarily described in pictorial representations but not in linguistic ones.

(F4) To replace default contents by definite ones given in the following contexts. For example, in such a context as "There is a box to the left of the pot. The box is red. …", the color of the box in a pictorial representation must be changed from default one to red.

For example, the text consisting such two sentences as 'There is a hard cubic object' and 'The object is large and gray' can be translated into a still picture in such a way as shown in FIG.16.

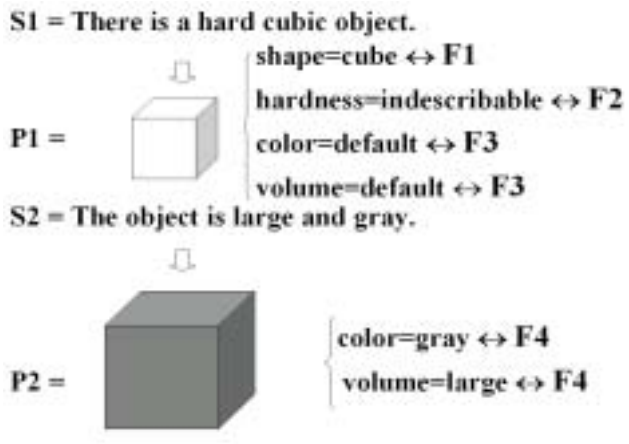


FIG.16 Systematic cross-media translation

7.2. Formalization

The MIDST assumes that any content conveyed by an information medium is to be associated with the loci in certain attribute spaces, and in turn that the world describable by each medium can be characterized by the maximal set of such attributes. This relation is conceptually formalized by the expression (60), where *Wm, Am,* and *F* mean 'the world describable by the information medium *m*', 'an attribute of the world', and 'a certain function for determining the maximal set of attributes of *Wm*', respectively.

$$F(Wm)=\{Am_1, Am_2, \ldots, Am_n\} \quad (60)$$

Considering this relation, cross-media translation is one kind of mapping from the world describable by the source medium (m_s) to that by the target medium (m_t) and can be defined by the expression (61).

$$Y(Sm_t)=\psi(X(Sm_s)) \quad (61)$$

where

Sm_s: the maximal set of attributes of the world describable by the source medium m_s ,

Sm_t: the maximal set of attributes of the world describable by the target medium m_t,

$X(Sm_s)$:a locus formula about the attributes belonging to Sm_s,

$Y(Sm_t)$: a locus formula about the attributes belonging to Sm_t ,

ψ : the function for transforming X into Y, so called, 'Locus formula paraphrasing function'.

The function ψ is designed to realize all the functions F1-F4 by inference processing at the level of locus formula representation.

7.3. Locus formula paraphrasing function ψ

In order to realize the function F1, a certain set of '***Attribute Paraphrasing Rules (APRs)***', so called, are defined ***at every pair of source and target media*** (See Table VI at 8.1).

The function F2 is realized by detecting locus formulas about ***the attributes without any corresponding APRs*** from the content of each input representation and replacing them by ***empty events***.

For F3, ***default reasoning*** is employed. That is, such an inference rule as defined by the expression (62) is introduced, which states if ***X is deducible and it is consistent to assume Y then conclude Z***.

This rule is applied typically to such instantiations of X, Y and Z as specified by the expression (63) which means that the indefinite attribute value 'p' with the indefinite standard 'k' of the indefinite matter 'y' is substitutable by

the constant attribute value '*P*' with the constant standard 'K' of the definite matter '$O_{\#}$' of the same kind of '*M*'.

$$X \circ Y \rightarrow Z \qquad (62)$$

$$\{X/(L(x,y,p,p,A,G,k)\wedge M(y))\wedge(L(z,O_{\#}P,P,A,G,K)\wedge M(O_{\#})),$$
$$Y/\ p=P \wedge k=K,\quad Z\ /\ L(x,y,P,P,A,G,K)\wedge M(y)\ \} \qquad (63)$$

Lastly, the function F4 is realized quite easily by memorizing the history of applications of default reasoning.

8. MISCELLANEOUS CROSS-MEDIA OPERATIONS

8.1. Mixed-media dialogue by text and picture

As easily imagined, IMAGES-M and humans can perform mixed-media dialogue employing text and picture as shown in FIG.17, where a drawing tool is utilized as graphical interface.

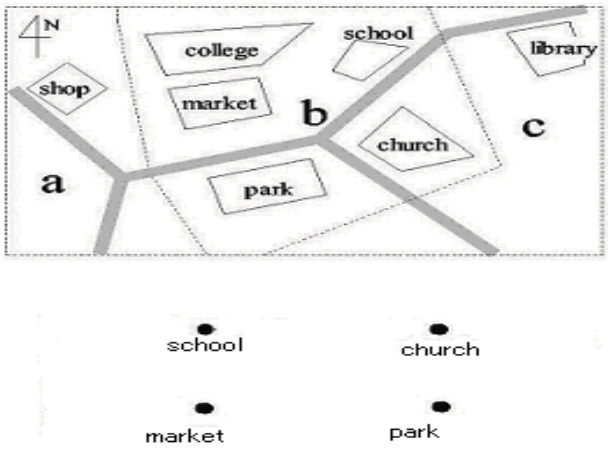


FIG.17 Q-A by mixture of text and picture between Humans (H) and IMAGES-M (S)

It is one of the most essential tasks for the system to determine how many pictures a locus formula should be interpreted into. Consider such somewhat complicated sentences as S9 and S10. The underlined parts are considered to refer to some events neglected in time and in space, respectively. These events are called 'Temporal Empty Event' and 'Spatial Empty Event', denoted by 'ε_t' and 'ε_s' as EEs with *g=Gt* and *g=Gs* at (10), respectively. The concepts of S13 and S14 are given by (64) and (65), where 'A15' and '_' represent the attribute 'Trajectory' and abbreviation of the variables bound by existential quantifiers, respectively.

In general, an atomic locus formula with *g=Gt* is to be depicted as a pair of pictures the formula and that with *g=Gs*, as one still picture. Therefore (65) is depicted as the still picture in FIG.18 while (64) as a series of still pictures, namely, a motion picture. FIG. 19 is an example of map generation from text via locus formula representation.

(S9) The *bus* runs 10km straight east from A to B, and *after a while*, at C it meets the street with the sidewalk.

$$(\exists x,y,z,p,q)(L(\_,x,A,B,A12,Gt,\_)\Pi L(\_,x,0,10km,A17,Gt,\_)$$
$$\Pi L(\_,x,Point,Line,A15,Gt,\_)\Pi L(\_,x,East,East,A13,Gt,\_))$$
$$\bullet\varepsilon_t\bullet(L(\_,x,p,C,A12,Gt,\_)\Pi L(\_,y,q,C,A12,Gs,\_)\Pi$$
$$L(\_,z,y,y,A12,Gs,\_))\wedge bus(x)\wedge street(y)\wedge$$
$$sidewalk(z)\wedge p\neq q \qquad (64)$$

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

$$(\exists x,y,z,p,q)(L(\_,x,A,B,A12,Gs,\_)\Pi L(\_,x,0,10km,A17,Gs,\_)$$
$$\Pi L(\_,x,Point,Line,A15,Gs,\_)\Pi L(\_,x,East,East,A13,Gs,\_))$$
$$\bullet\varepsilon_s\bullet(L(\_,x,p,C,A12,Gs,\_)\Pi L(\_,y,q,C,A12,Gs,\_)\Pi$$
$$L(\_,z,y,y,A12,Gs,\_))\wedge road(x)\wedge street(y)\wedge$$
$$sidewalk(z)\wedge p\neq q \qquad (65)$$

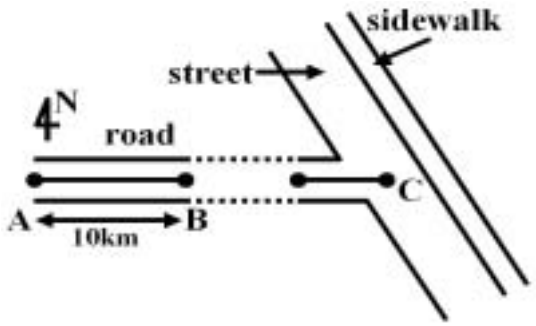


FIG.18 Pictorial interpretation of (65)

FIG.19 A map generated from a text via a locus formula

The APRs must be set up as relationships between the attributes concerning the paired media to be translated each other. For our experiment on text-picture translation [18], there were employed five kinds of APRs such as shown in Table VI, where *p,s,c,*…and *p',s',c',*…are linguistic expressions and their corresponding pictorial expressions of attribute values, respectively. Further details are as follows:

(1) APR-02 is used especially for a sentence such as "The box is 3 meters to the left of the chair." The symbols *p, d* and *l* correspond to 'the position of the chair', 'left' and '3 meters', respectively, yielding the pictorial expression of 'the position of the box', namely, " *p'+l'd'* ".

(2) APR-03 is used especially for a sentence such as "The pot is big." The symbols *s* and *v* correspond to 'the shape of the pot (default value)' and 'the volume of the pot ('big')', respectively. In pictorial expression, the shape and the volume of an object is inseparable and therefore they are represented only by the value of the attribute 'shape', namely, " *v's'* ".

(3) APR-05 is used especially for a sentence such as "The cat is under the desk." The symbols p_a , p_b and *m* correspond to 'the position of the desk', 'the position of the cat' and 'under' respectively, yielding a pair of pictorial expressions of the positions of the two objects.

TABLE VI: APRs for text-picture translation

APRs	Correspondences of attributes (Text : Picture)	Value conversion schema (Text ⇔ Picture)
APR-01	A12 : A12	p⇔p'
APR-02	{A12, A13, A17} : A12	{ p, d, l}⇔p'+l'd'
APR-03	{A11, A10} : A11	{s, v}⇔v's'
APR-04	A32 : A32	c⇔c'
APR-05	{A12, A44} : A12	{p_a,m}⇔{p_a', p_b'}

8.2. Linguistic interpretation of human motion data

The human body can be described in a computable form using locus formulas. That is, the structure of the human body is one of spatial event where the body parts such as head, trunk, and limbs extend spatially and connect with each other. The expressions (66) and (67) are examples of these descriptions using locus formulas which reads roughly that an arm extends from the hand to the shoulder and that a wrist connects the hand and the forearm, respectively.

$$(\lambda x)\mathrm{arm}(x)\leftrightarrow(\lambda x)\{\exists y_1,y_2,k)\ L(x,x,y_1,y_2,A12,Gs,k)\wedge \mathrm{shoulder}(y_1)\wedge \mathrm{hand}(y_2) \quad (66)$$

$$(\lambda x)\mathrm{wrist}(x)\leftrightarrow(\lambda x)\{\exists y_1,y_2,y_3,y_4,k)\ (L(y_1,y_1,y_2,x,A12,Gs,k)\Pi L(y_1,y_1,x,y_3,A12,Gs,k))\wedge \mathrm{body\text{-}part}(y_1)\wedge \mathrm{forearm}(y_2)\wedge \mathrm{hand}(y_3) \quad (67)$$

The structural description in the computable form is indispensable to mutual translation between human motion data and linguistic expressions. For example, it enables the system to recognize the anomaly of such a sentence as S11 in such a process described at the section IVB of this paper.

(S11) The left arm moved away from the left shoulder and the left hand.

Various kinds of human motions have been conceptualized as specific verbs in natural languages such as 'nod' and 'crouch'. For example, the conceptual description of 'nodding' is given by (68) which reads roughly that a person lets the head fall forward. The conceptual description of a verb gives the framework of the meaning representation of the sentence where the very verb appears. This kind of meaning representation is called 'Text Meaning Representation (TMR)' as mentioned below.

$$(\lambda x)\text{nodding}(x)\leftrightarrow(\lambda x)(\exists y_1,y_2,k_1,k_2,k_3)\\ L(y_1,\{y_1,y_2\},x,x,A01,Gt,k_1)\\ \Pi L(y_1,y_2,Down,Down,A13,Gt,k_2)\\ \Pi L(y_1,y_2,Forward,Forward,A13,Gt,k_3)\wedge\text{person}(y_1)\\ \wedge\text{head}(y_2)\wedge\text{motion}(x) \quad (68)$$

As for our experiment, colored markers were put on the upper half part of human body, namely, head, neck, shoulders, elbows, hands, and navel and their position data (i.e. 3D coordinates) were taken in through a motion capturing system at a sampling rate. FIG. 20 shows the structure of the wire frame model of the upper half of the human body. This model was implemented by using locus formula representation just like (66) and (67). Real motion data were graphically interpreted according to the model as shown in FIG.21.

The datum unit can be formalized by a quadruple (S, B, P, T), where S, B, P and T mean 'name of the subject', 'name of the body part, 'position of the body part' and 'time point of data sampling', respectively.

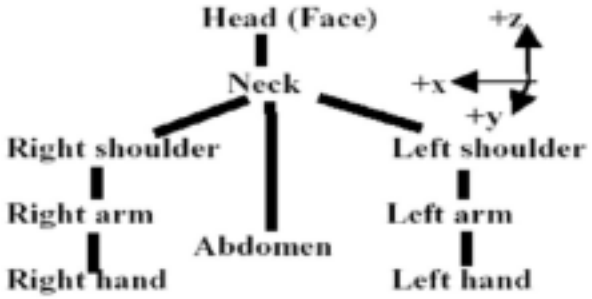


FIG.20 Wire frame model of upper half of human body

The Data Interpretation Function (F_d) digests a large number of motion data of the subject's head over a time interval into a locus formula such as (69), where '*Tom*' is the default name of the subject and *P*s are characteristic points of the movement of the head such as turning points. This type of expression is called 'Motion Meaning Representation (MMR)', where the Standard constant Mc means one of certain rigid standards specific to the motion capturing system.

$$L(Tom,Head,P_1,P_2,A12,Gt,Mc)\\ \bullet L(Tom,Head,P_2,P_3,A12,Gt,Mc)\\ \bullet\ldots\bullet L(Tom,Head,P_{n-2},P_{n-1},A12,Gt,Mc)\\ \bullet L(Tom,Head,P_{n-1},P_n,A12,Gt,Mc) \quad (69)$$

Human motion data gained through a motion capturing system associate limitedly with the sense 'sight' and its related attributes are A12 (Physical location) and A34 (Time point).

In translation between motion data and texts, these two attributes and the others are to be paraphrased with each other according to 'Attribute paraphrasing rules (APRs)' such as (70)-(72), where the left and right hands of the symbol '⇔' refer to the attributes concerning to MMRs and TMRs, respectively. And the attributes 'A01' and 'A13' refer to 'Place of existence' and 'Direction', respectively.

$$(\exists\mathbf{p},\mathbf{q})L(y_1,y_2,\mathbf{p},\mathbf{q},A12,Gt,Mc)\wedge\mathbf{q}\neq\mathbf{p}\wedge\mathbf{p}=(p_x,p_y,p_z)\\ \wedge\mathbf{q}=(q_x,q_y,q_z)\Leftrightarrow(\exists x,k)L(y_1,\{y_1,y_2\},x,x,A01,Gt,k)\\ \wedge\text{motion}(x) \quad (70)$$

$$(q_z-p_z<0,A12)\Leftrightarrow(Down,A13) \quad (71)$$

$$(q_y-p_y>0,A12)\Leftrightarrow(Forward,A13) \quad (72)$$

Based on APRs (70)-(72), the MMR (69) is unified with (68), namely, the conceptual description of the verb 'nod', which yields the TMR (73).

$$(\exists x,k_1,k_2,k_3)\ L(Tom,\{Tom,Head\},x,x,A01,Gt,k_1)\\ \Pi L(Tom,Head,Down,Down,A13,Gt,k_2)\\ \Pi L(Tom,Head,Forward,Forward,A13,Gt,k_3)\\ \wedge\text{person}(Tom)\wedge\text{head}(Head)\wedge\text{motion}(x) \quad (73)$$

The sentence 'Tom nodded.' is to be generated from this TMR using the sentence pattern of 'nod' which is generalized as 'y_1 **nod**' indicating the correspondence between the subject of the verb and the term 'y_1' in its

conceptual description (68).

FIG. 21 (1)-(3) are graphical interpretations of the real motion data at the time points, t_1, t_2 and t_3, respectively. The sets of real motion data over time intervals $[t_1, t_2]$ and $[t_2, t_3]$ were translated into the texts in FIG.22 (a) and (b), respectively.

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

FIG.21. Graphical interpretations of real motion data

Tom moved the right hand.
Tom moved the right arm.
Tom moved the right elbow.
.....
Tom put the right hand up.
Tom raised the right arm.
Tom bent the right arm.
Tom put the right hand up and simultaneously bent the right arm.
.....

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

.....
Tom put the right hand down.
Tom lowered the right arm.
Tom stretched the right arm and simultaneously lowered the right hand.
.....

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

FIG.22 Texts generated from real motion data

8.3. Robot manipulation by natural language

The intelligent system IMAGES-M can deploy SONY AIBOs, dog-shaped robots, as actors and gather information about the physical world through their microphones, cameras and tactile sensors. Communications between IMAGES-M and humans are performed though the keyboard, mouse, microphone and multicolor TV monitor of the personal computer.

Consider such a verbal command as S12 uttered to the robot, SONY AIBO, named 'John'.

(S12) John, walk forward and wave your left hand.

Firstly, late in the process of cross-media translation from text to AIBO's action, this command is to be interpreted into (74) with the attribute 'shape (A11)' and the values '$Walkf_1$' and so on at the standard of 'AIBO', reading that John makes himself walk forward and wave his left hand. Each action in AIBOs is defined as an ordered set of shapes (i.e., time-sequenced snapshots of the action) corresponding uniquely with the positions of their actuators determined by the rotations of the joints. For example, the actions 'walking forward (**W**alkf)' and 'waving left hand (**W**avelh)' are defined as (75) and (76), respectively.

$$L(John,John,Walkf_1,Walkf_m,A11,Gt,AIBO) \wedge L(John,John,Wavelh_1,Wavelh_n,A11,Gt,AIBO) \quad (74)$$

$$\mathbf{W}alkf=\{Walkf_1,Walkf_2,\ldots,Walkf_m\} \quad (75)$$

$$\mathbf{W}avelh=\{Wavelh_1,Wavelh_2,\ldots,Wavelh_n\} \quad (76)$$

Secondly, an AIBO cannot perform the two events (i.e., actions) simultaneously and therefore the transit event X_T between them is to be inferred as the underlined part of (77) which is the goal event X_G here.

$$L(John,John,Walkf_1,Walkf_m,A11,Gt,AIBO) \bullet \underline{L(John,John,Walkf_m,Wavelh_1,A11,Gt,AIBO)} \bullet L(John,John,Wavelh_1,Wavelh_n,A11,Gt,AIBO) \quad (77)$$

Thirdly, (78) is to be inferred, where the transit event, underlined, is interpolated between the current event X_C and the goal event X_G (=(77)).

$$L(John,John,p_1,p_2,A11,Gt,AIBO) \bullet \underline{L(John,John,p_2,Walkf_1,A11,Gt,AIBO)} \bullet X_G \quad (78)$$

Finally, (78) is interpreted into a series of joint rotations in the AIBO as shown in FIG.23.

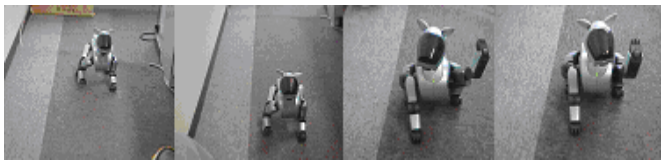


FIG.23 AIBO (Sony) behaving in accordance to the command 'Walk forward and wave your left hand'

9. DISCUSSION AND CONCLUSION

AI planning ("action planning") deals with the development of representation languages for planning problems and with the development of algorithms for plan construction [6, 7, 29]. The author formalized the performances of a DIRN as predicate logic in the formal language L_{md} and applied it to robot manipulation by text and so on as simulation of DIRN-world interaction.

This is one kind of cross-media operation via locus formulas as already reported. At my best knowledge [e.g., 30, 31], there is no other theory or system that can perform cross-media operations in such a seamless way as ours, which leads to the conclusion that employment of locus formulas has made both spatial and temporal event concepts remarkably computable in an integrated way and has proved to be very adequate to systematize cross-media operations. This is due to their medium-freeness and good correspondence with the performances of miscellaneous devices, which in turn implies that locus formula representation may make it easier for the devices to share a task than any other representation, even if, based on some precise ontology or mathematical definition [33, 34].

In this simulation, a problem for a DIRN to solve is defined as a goal event (X_G) and a transit event (X_T) between the current event (X_C) and the goal event. The task sharing and assignment among the nodes or agents are executed based on the information of a problem described as locus formulas in L_{md}. The most useful keys to task assignment are the attributes involved and about 50 kinds of attributes have been found in association with natural languages and human sensory organs as shown in Table II.

Furthermore, most of computations on L_{md} are simply for unifying (or identifying) atomic locus formulas and for evaluating arithmetic expressions such as '$p=q$', and therefore we believe that our formalism can reduce the computational complexities of the others [14-16] when applied to the same kinds of problems described here.

The simulation results lead to the conclusion that L_{md} can be a universal language appropriate for WSANs including DIRNs. Our future work will include establishment of learning facilities for automatic acquisition of word concepts from sensory data and human-robot interaction by natural language under real environments.

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] I.F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, *Wireless sensor networks: a survey. Computer Networks, 38-4, pp.393-422*, (2002)

[2] M. Haenggi, *Mobile Sensor-Actuator Networks: Opportunities and Challenges. Proc. of 7th IEEE Int. Workshop, Frankfurt,Germany, pp.283-290*, (2002)

[3] I.F. Akyildiz and I.H. Kasimoglu, *Wireless Sensor and Actor Networks: Research Challenges. Ad Hoc Networks, 2, pp.351-367*, (2004)

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

[5] M. Yokota, *Towards a Universal Knowledge Representation Language for Ubiquitous Intelligence Based on Mental Image Directed Semantic Theory. J.Ma et al.(Eds.) Ubiquitous Intelligence and Computing 2006 (UIC 2006), LNCS 4159, pp.1124-1133*, (2006)

[6] D.E. Wilkins and K.L. Myers, *A common knowledge representation for plan generation and reactive execution. Journal of Logic and Computation, 5-6, pp.731-761*, (1995)

[7] F. Kabanza, *Synchronizing multiagent plans using temporal logic specifications. Proceedings of the First International Conf. on Multi-Agent Systems (ICMAS-95), pp.217-224*, (1995)

[8] J. Bos and T. Oka, *A Spoken Language Interface to a Mobile Robot. Proceedings of the Eleventh International Symposium on Artificial Life and Robotics (AROB-11)*, (Oita, Japan, 2006)

[9] M. Yokota, *An approach to natural language understanding based on a mental image model. Proc. of the second International Workshop on Natural Language Understanding and Cognitive Science, pp.22-31*, (2005)

[10] M. Yokota, et al, *Mental-image directed semantic theory and its application to natural language understanding systems. Proc. of NLPRS'91, pp.280-287*, (1991)

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

[12] G.P. Zarri, *NKRL, a Knowledge Representation Tool for Encoding the 'Meaning' of Complex Narrative Texts. Natural Language Engineering - Special Issue on Knowledge Representation for Natural Language Processing in Implemented Systems, 3, pp.231-253*, (1997)

[13] B. Dorr and J. Bonnie, *Large-Scale Dictionary Construction for Foreign Language Tutoring and Interlingual Machine Translation. Machine Translation, 12-4, pp.271-322*, (1997)

[14] J.F. Allen, *Towards a general theory of action and time. Artificial Intelligence, 23-2, pp.123-154 (1984)*

[15] Y. Shoham, Time for actions: on the relationship between time, knowledge, and action. Proc. of IJCAI89, pp.954-959, (Detroit, MI 1989)

[16] P. Haddawy, *A logic of time, chance, and action for repesenting plans. Artificial Intelligence, 80-2, pp.243-308*, (1996)

[17] S. Oda, M. Oda, and M. Yokota, *Conceptual Analysis Description of Words for Color and Lightness for Grounding them on Sensory Data. Trans. of JSAI,Vol.16-5-E, pp.436-444*, (2001)

[18] M. Yokota and G. Capi, *Integrated Multimedia Understanding for Ubiquitous Intelligence Based on Mental Image Directed Semantic Theory. IFIP EUC'05 UISW2005, pp.538-546*, (Nagasaki, 2005)

[19] M. Amano, et al , *Linguistic interpretation of human motion based on Mental Image Directed Semantic Theory. Proc. of IEEE AINA-2005, pp.139-144*, (Taipei, 2005)

[20] R. Langacker, *Concept, Image and Symbol*, (Mouton de Gruyter, Berlin/New York, 1991)

[21] G.A. Miller and P.N. Johnson-Laird, *Language and Perception*, (Harvard University Press, 1976)

[22] M. Yokota, H. Yoshitake, and T. Tamati, *Japanese-English translation of weather reports by ISOBAR. Trans of IECE Japan, E67-6, pp.315-322*, (1984)

[23] E. Leisi, *Der Woltinhalt —Seine Struktur in Deutchen und Englischen—*, (Quelle & Meyer, Heidelberg, 1961)

[24] E. De Bono, *The mechanism of mind*, (C.Tuffle Co. Inc., Tokyo, 1969)

[25] D. Noton and L. Stark, *Scanpaths in eye movements during pattern perception, Science, 171, 3968, pp. 308-311*, (1971)

[26] I.A. Rybak, V.I. Gusakova, A.V. Golovan, L.N. Podladchikova, and N.A. Shevtsova, *A model of attention-guided visual perception and recognition, Vision Research, 38, pp.2387-2400*, (1998)

[27] S. Coradeschi and A. Saffiotti, *An Introduction to the Anchoring Problem, Robotics and Autonomous Systems 43, pp.85-96*, (2003)

[28] M. Egenhofer and K. Khaled, *Reasoning about Gradual Changes of Topological Relations, Proc. of International Conference GIS—From Space to Territory: Theories and Methods of Spatio-Temporal Reasoning, pp.196-219*, (Pisa, Italy, 1992)

[29] R. Lundh, L. Karlson, and A. Saffiotti, *Plan-Based Configuration of a Group of Robots. Proc. of the seventeenth European Conference on Artificial Intelligence (ECAI)*, (Riva del Garda, Italy, 2006)

[30] G. Adorni, M. Di Manzo, and F. Giunchiglia, *Natural Language Driven Image Generation. Proc. of* COLING 84, *pp.495-500*, (1984)

[31] J.P. Eakins, M.E. Graham, *Content-based Image Retrieval, A report to the JISC Technology Applications Programme*, (Institute for Image Data Research, University of Northumbria at Newcastle, 1999)

[32] M. Yokota, *An Approach to Integrated Spatial Language Understanding Based on Mental Image Directed Semantic Theory. Proc. of the 5th Workshop on Language and Space*, (Bremen, Germany, Oct. 2005)

[33] N. Guarino and C.A. Welty, *An Overview of OntoClean, Handbook on Ontologies, pp.151-172*, (2004)

[34] P. Gardenfors, *Representing actions and functional properties in conceptual sapces, Ziemke,T. and Zlatev,J.(Eds.) Language and Mind*, (2004)