Proceedings of the 5th WSEAS Int. Conf. on MULTIMEDIA, INTERNET AND VIDEO TECHNOLOGIES, Corfu, Greece, August 17-19, 2005 (pp204-211)

A Multimedia Description Language for Cross-media Operations Based on Mental Image Directed Semantic Theory

MASAO YOKOTA GENCI CAPI
Department of System Management, Faculty of Information Engineering
Fukuoka Institute of Technology
3-30-1 Wajiro-higashi, Higashi-ku, Fukuoka-shi
JAPAN
http://www.fit.ac.jp

Abstract: -The Mental Image Directed Semantic Theory (MIDST) has proposed an omnisensual mental image model and its description language $\boldsymbol{L_{md}}$. This language can provide multimedia expressions with intermediate semantic descriptions in predicate logic. This paper presents a brief sketch of $\boldsymbol{L_{md}}$ and its application to cross-media operations between linguistic and pictorial expressions of space and time.

Key-Words: - Cross-media operations, Multimedia description language, Omnisensual mental image model

1 Introduction

The need for more human-friendly intelligent systems has been brought by rapid increase of aged societies, floods of multimedia information over the internet, development of robots for practical use, and so on.

For example, it is very difficult for people to exploit necessary information from the immense multimedia contents over the internet. And it is still more difficult to search for desirable contents by queries in different media, for example, text queries for pictorial contents. In this case, intelligent systems facilitating cross-media references are very helpful.

In order to realize these kinds of intelligent systems, we think it is needed to develop such a computable knowledge representation language for multimedia contents that should have at least a capability of representing spatio-temporal events that people perceive in the real world. In this research area, it is most conventional that conceptual contents conveyed by information media such as languages and pictures are represented in computable forms independent of each other and translated via 'transfer' processes so called which are often very specific to task domains [8], [9], [10].

Yokota, M. et al ([2]) have proposed a semantic theory for natural languages so called 'Mental Image Directed Semantic Theory (MIDST)'. In the MIDST, word concepts are associated with omnisensual mental images of the external or physical world and are formalized in an intermediate language L_{md}, based on first-order predicate logic, while the other knowledge description schema such as [3], [4] are too linguistic (or English-like) to formalize omnisensual mental images.

The $\boldsymbol{L_{md}}$ is employed for many-sorted predicate logic and has been implemented on several types of computerized intelligent systems [1], [5]. There is a feedback loop between them for their mutual refinement unlike other similar theories [6], [7].

This paper presents a brief sketch of $\boldsymbol{L_{md}}$ and its application to cross-media operations between linguistic and pictorial expressions of space and time.

2 Brief sketch of $L_{\mathbf{md}}$

The MIDST treats word meanings in association with mental images, not limited to visual but omnisensual, modeled as "Loci in Attribute Spaces". An attribute space corresponds with a certain measuring instrument just like a barometer, a map measurer or so and the loci represent the movements of its indicator.

As a simple example, the black triangular object in motion, as shown in Fig.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.

A general locus is to be articulated by "Atomic Locus" with the duration $[t_i, t_f]$ as depicted in Fig.2 and formalized as (1).

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

For example, the motion of the 'bus' referred to 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 formalized as (2) and (3), respectively, where the attribute is 'Physical Location' denoted by 'A12'. For simplicity, Matter terms (e.g., 'Tokyo' and 'Osaka' in S1 and S2) are often placed at Attribute Values or Standard to represent their values at the time.

(S1) The bus runs from Tokyo to Osaka.

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

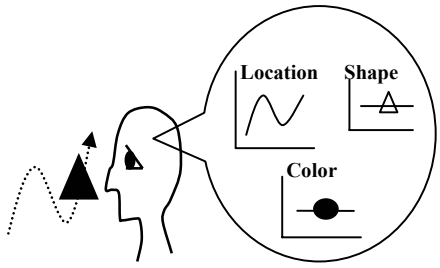

Fig.1. Mental image model.

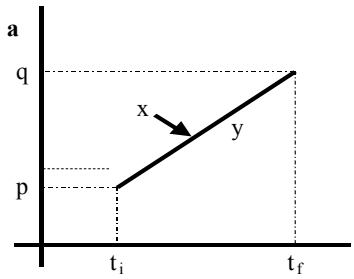

Fig.2. Atomic locus.

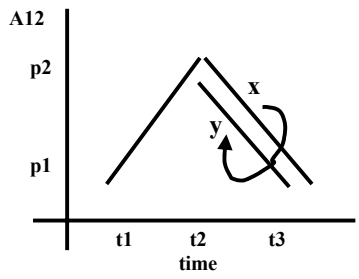

Fig.3. Conceptual image of 'fetch'.

(S2) The road runs from Tokyo to Osaka.

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

The expression (4) is the conceptual description of the English verb 'fetch' depicted as Fig.3, implying such a temporal event that 'x' goes for 'y' and then comes back with it, where 'Π'and '•' are the tempological connectives, 'SAND' and 'CAND', standing for 'Simultaneous AND' and 'Consecutive AND', respectively.

$$(\exists x,y,p1,p2,k)\ L(x,x,p1,p2,A12,Gt,k)\bullet \\ ((L(x,x,p2,p1,A12,Gt,k)\Pi L(x,y,p2,p1,A12,Gt,k)) \\ \wedge x\neq y\ \wedge p1\neq p2 \quad (4)$$

Such an expression as (4) is called 'Event Pattern' and about 40 kinds of event patterns have been found concerning the attribute 'Physical Location (A12)', for example, *start*, *stop*, *meet*, *separate*, *return*, etc[2].

Furthermore, a very important concept called 'Empty Event (EE)' and denoted by 'ε' is introduced. An EE stands for nothing but for time collapsing and is explicitly defined as (5) with the attribute 'Time Point (A34)'. According to this scheme, the duration $[p, q]$ of an arbitrary locus X can be expressed as (6).

$$\varepsilon \Leftrightarrow (\exists x,y,p,q,g,k)\ L(x,y,p,q,A34,g,k) \quad (5)$$

$$X\ \Pi\ \varepsilon(p,q) \quad (6)$$

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.4, 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***. Therefore, S3 and S4 refer to the same scene in spite of their appearances as shown in Fig.5 where, as easily imagined, what 'sinks' or 'rises' is the FAO, and whose conceptual descriptions are given as (7) and (8), respectively.

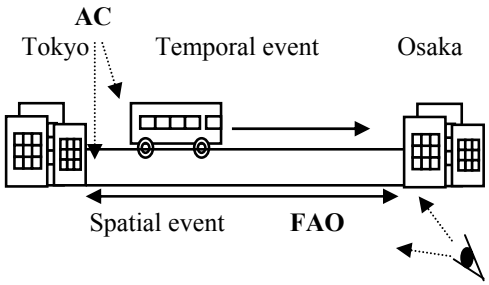

Fig.4. Event types and FAO movements.

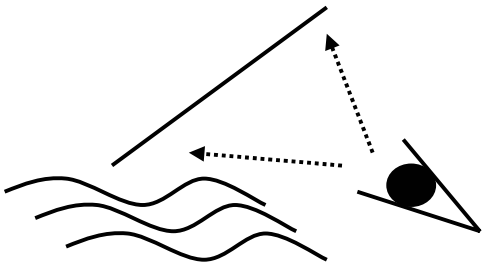

Fig.5. Slope as a spatial event.

(S3) The path *sinks to* the brook.
(∃x,y,p,z,k1,k2)L(x,y,p,z,A12,Gs,k1)ΠL(x,y,↓,↓,A13,Gs,k2)∧path(y)∧brook(z)∧p≠z (7)

(S4) The path *rises from* the brook.
(∃x,y,p,z,k1,k2)L(x,y,z,p,A12,Gs,k1)ΠL(x,y,↑,↑,A13,Gs,k2)∧path(y)∧brook(z)∧p≠z (8)

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. This postulation is also valid for such a pair of S5 and S6 interpreted as (9) and (10), respectively, where 'A13', '↑' and '↓' refer to the attribute 'Direction' and its values 'upward' and 'downward', respectively. These pairs of conceptual descriptions are called ***equivalent in the PRS***, and the paired sentences are treated as ***paraphrases*** each other.

(S5) Route A and Route B meet at the city.
(∃x,p,y,q,k)L(x,Route_A,p,y,A12,Gs,k)ΠL(x,Route_B,q,y,A12,Gs,k)∧city(y)∧p≠q (9)

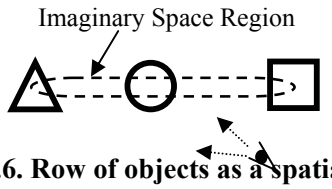

Fig.6. Row of objects as a spatial event.

(S6) Route A and Route B separate at the city.
(∃x,p,y,q,k)L(x,Route_A,y,p,A12,Gs,k)ΠL(x,Route_B,y,q,A12,Gs,k)∧city(y)∧p≠q (10)

For another example of spatial event, Fig.6 concerns the perception of the formation of multiple objects, where FAO runs along an imaginary object so called 'Imaginary Space Region (ISR)'.

This spatial event can be verbalized as S7 using the preposition 'between' and formalized as (11), corresponding also to such concepts as 'row', 'line-up', etc. Employing ISRs and the 9 intersection model [11], all the topological relations between two objects can be formalized in such expressions as (12) or (12') for S8, and (13) for S9, where '*In*', '*Cont*' and '*Dis*' are the values 'inside', 'contains' and 'disjoint' of the attribute 'Topology (A44)' with the standard '9 intersection model (*9IM*)', respectively.

(S7) ○ is between Δ and □.
(∃x,y,p,q,k)(L(x,y,Δ,○,A12,Gs,k)ΠL(x,y,p,p,A13,Gs,k))• (L(x,y,○,□,A12,Gs,k)ΠL(x,y,q,q,A13,Gs,k))∧ISR(y)∧p=q (11)

(S8) Tom is in the room.
(∃x,y,z,k)L(x,y,Tom,z,A12,Gs,k)ΠL(x,y,In,In,A44,Gt,9IM) ∧ISR(y) ∧room(z) (12)

(∃x,y,z,k)L(x,y,z,Tom,A12,Gs,k)ΠL(x,y,Cont,Cont,A44,Gt,9IM)∧ISR(y)∧room(z) (12')

(S9) Tom exits the room.
(∃x,y,z,k)L(x,y,Tom,z,A12,Gs,k)ΠL(x,y,In,Dis,A44,Gt,9IM) ∧ISR(y) ∧room(z) (13)

3 Word meaning descriptions

A word meaning description M_w is given by (14) as a pair of 'Concept Part (C_p)' and 'Unification Part (U_p)'.

$$M_w \Leftrightarrow [C_p : U_p] \quad (14)$$

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 (15).

$$[(\exists x,y,p1,p2,k)\ L(x,x,p1,p2,A12,Gt,k)\Pi L(x,y,p1,p2,A12,Gt,k)\wedge x\neq y\wedge p1\neq p2 : ARG(Dep.1,x); ARG(Dep.2,y);] \quad (15)$$

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 sentence '*Mary carries a book*' is translated into (16).

$$(\exists y,p1,p2,k)L(Mary,Mary,p1,p2,A12,Gt,k)\Pi L(Mary,y,p1,p2,A12,Gt,k)\wedge Mary\neq y \wedge p1\neq p2 \wedge book(y) \quad (16)$$

For another example, the meaning description of the English preposition 'through' is also given by (17).

$$[(\exists x,y,p1,z,p3,g,k,p4,k0) (\underline{L(x,y,p1,z,A12,g,k)} \bullet L(x,y,z,p3,A12,g,k))\Pi L(x,y,p4,p4,A13,g,k0)\wedge p1\neq z\wedge z\neq p3 : ARG(Dep.1,z); IF(Gov=Verb)\rightarrow PAT(Gov,(1,1)); IF(Gov=Noun)\rightarrow ARG(Gov,y);] \quad (17)$$

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 (17) 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 S10, S11 and S12 are interpreted as (18), (19) and (20), respectively. The underlined parts of these formulas are the results of PAT operations. The expression (21) 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 (21').

(S10) The train runs through the tunnel.

$$(\exists x,y,p1,z,p3,k,p4,k0) (\underline{L(x,y,p1,z,A12,Gt,k)} \bullet L(x,y,z,p3,A12,Gt,k)) \Pi L(x,y,p4,p4,A13,Gt,k0)\wedge p1\neq z\wedge z\neq p3 \wedge train(y)\wedge tunnel(z) \quad (18)$$

(S11) The path runs through the forest.

$$(\exists x,y,p1,z,p3,k,p4,k0) (\underline{L(x,y,p1,z,A12,Gs,k)} \bullet L(x,y,z,p3,A12,Gs,k)) \Pi L(x,y,p4,p4,A13,Gs,k0)\wedge p1\neq z\wedge z\neq p3 \wedge path(y)\wedge forest(z) \quad (19)$$

(S12) The path through the forest is long.

$$(\exists x,y,p1,z,p3,x1,k,q,k1,p4,k0) (L(x,y,p1,z,A12,Gs,k) \bullet L(x,y,z,p3,A12,Gs,k)) \Pi L(x,y,p4,p4,A13,Gs,k0)\wedge L(x1,y,q,q,A02,Gt,k1) \wedge p1\neq z\wedge z\neq p3 \wedge q>k1 \wedge path(y)\wedge forest(z) \quad (20)$$

$$(\exists x1,y1,q,k1)L(x1,y1,q,q,A02,Gt,k1)\wedge q>k1 \quad (21)$$

$$(\exists x1,y1,k1)L(x1,y1,Long,Long,A02,Gt,k1) \quad (21')$$

For another example, consider such somewhat complicated sentences as S13 and S14. 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 (5), respectively. The concepts of S13 and S14 are given by (22) and (23), where 'A15' and '_' represent the attribute 'Trajectory' and abbreviation of the variables bound by existential quantifiers, respectively. Figure 7 shows an example of pictorial interpretation of (23).

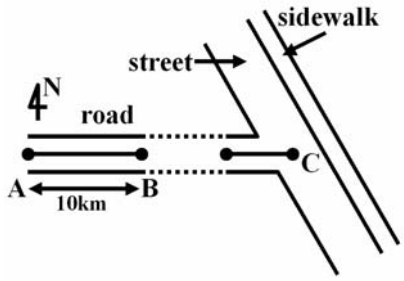

Fig.7. Pictorial interpretation of (23).

(S13) The *bus* runs 10km straight east from A to B, and *after a while*, at C it meets the street with the sidewalk.

$$(\exists x,y,z,p,q)(L(\_,x,A,B,A12,Gt,\_)\Pi L(\_,x,0,10km,A17,Gt,\_)\ \Pi L(\_,x,Point,Line,A15,Gt,\_)\Pi L(\_,x,East,East,A13,Gt,\_))\ \bullet\ \varepsilon_t\ \bullet\ (L(\_,x,p,C,A12,Gt,\_)\ \Pi L(\_,y,q,C,A12,Gs,\_)\Pi L(\_,z,y,y,A12,Gs,\_))\wedge bus(x)\wedge street(y)\wedge sidewalk(z)\wedge p\neq q \quad (22)$$

(S14) The *road* runs 10km straight east from A to B, and *after a while*, at C it meets the street with the sidewalk.

$$(\exists x,y,z,p,q)(L(\_,x,A,B,A12,Gs,\_)\Pi L(\_,x,0,10km,A17,Gs,\_)\Pi L(\_,x,Point,Line,A15,Gs,\_)\Pi L(\_,x,East,East,A13,Gs,\_))\bullet\varepsilon_s\bullet(L(\_,x,p,C,A12,Gs,\_)\Pi L(\_,y,q,C,A12,Gs,\_)\ \Pi L(\_,z,y,y,A12,Gs,\_))\wedge road(x)\wedge street(y)\wedge sidewalk(z)\wedge p\neq q \quad (23)$$

4 Cross-media translation

4.1 Functional requirements

The authors have 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.

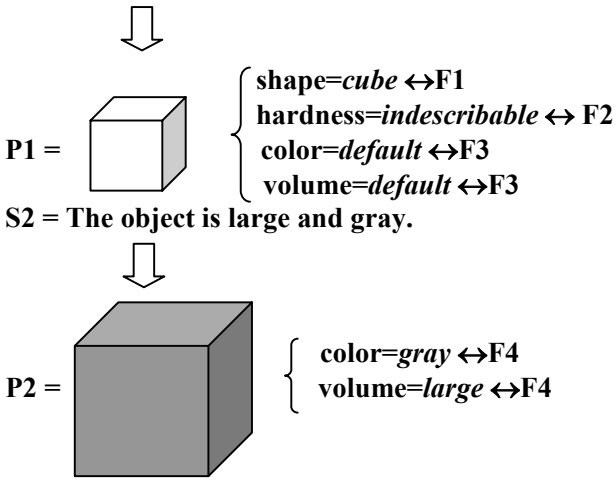

Fig.8. Systematic cross-media translation.

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.8.

4.2 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 (24), where *Wm, Am*$_i$, 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.

$$F(Wm)=\{Am_1, Am_2, \ldots, Am_n\} \quad (24)$$

Considering this relation, cross-media translation is one kind of mapping from the world describable by the source medium (*ms*) to that by the target medium (*mt*) and can be defined by the expression (25).

$$Y(Smt)=\psi(X(Sms)), \quad (25)$$

where

Sms: the maximal set of attributes of the world describable by the source medium *ms* ,

Smt: the maximal set of attributes of the world describable by the target medium *mt*,

X(Sms) :a locus formula about the attributes

belonging to *Sms*,

Y(Smt) : a locus formula about the attributes belonging to *Smt* ,

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.

4.3 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 Section 5).

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 (26) 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 (27) 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'*.

$$X \circ Y \rightarrow Z \qquad (26)$$

$$\begin{aligned} &\{X / (L(x,y,p,p,A,G,k) \wedge M(y)) \\ &\quad \wedge (L(z,O\#,P,P,A,G,K) \wedge M(O\#)), \\ &Y / p=P \wedge k=K, \\ &Z / L(x,y,P,P,A,G,K) \wedge M(y) \} \end{aligned} \qquad (27)$$

The function F4 is realized quite easily by ***memorizing the history of applications of default reasoning***.

5 Cross-media operations between text and picture

5.1 Attribute paraphrasing rules

Five kinds of APRs for this case are shown in Table 1 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 1. APRs for text-to-picture translation.

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

5.2 Implementation

The methodology mentioned above has been implemented on the intelligent system IMAGES-M [1] shown in Fig.9. IMAGES-M is one kind of expert system with five kinds of user interfaces besides the inference engine (IE) and the knowledge base (KB) as follows.

(1) Text Processing Unit (TPU),
(2) Speech Processing Unit (SPU),
(3) Picture Processing Unit (PPU),
(4) Action Data Processing Unit (ADPU),
(5) Sensory Data Processing Unit (SDPU).

These user interfaces can mutually convert information media and locus formulas in the

collaboration with IE and KB, and miscellaneous combinations among them bring forth various types of cross-media operations.

Figures 10-12 show several examples of cross-media operations between texts and pictures.

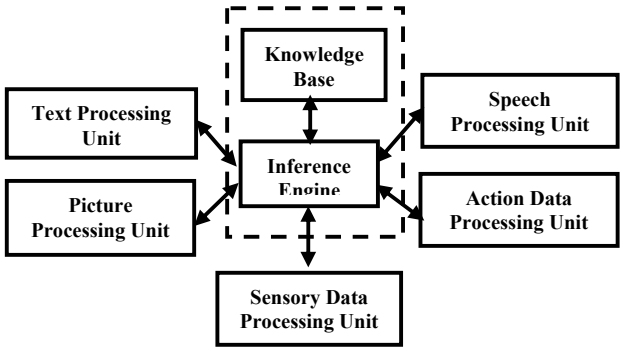

Fig.9. Configuration of IMAGES-M.

Karrigia eshte 3m ne te djathte te vazos.

猫は椅子の１ｍ下にいる。

Macja eshte e kuqe.
the small box is 1m to the left of the chair.
the big blue lamp is 2m above the pot.

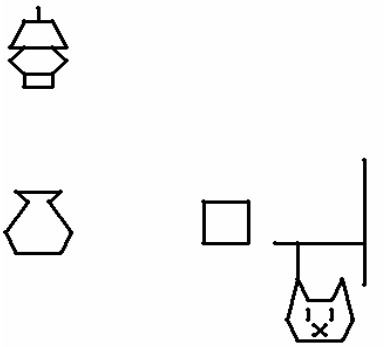

Fig.10. Text-to-picture translation.

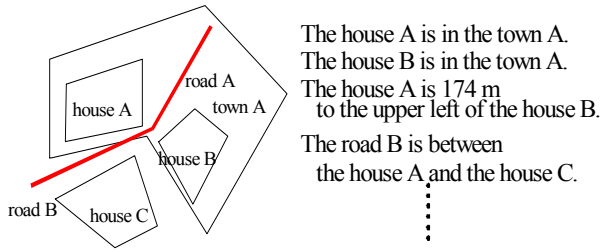

Fig.11. Picture-to-text translation.

H: ?猫 是 紅的 (Is the cat red ?)
S: 是 (yes)
H: ?何が椅子と花瓶の間にある
(What is between the chair and the flower-pot ?)
S: 箱 (box)
H: Is the box between the cat and the pot ?
S: NO
H: Eshte kutia midis maces dhe llampes ?
(Is the box between the cat and the lamp ?)
S: PO (yes)

Fig.12 Q-A about the picture in Fig.10 ('H': humanuser, 'S':IMAGES-M).

6 Discussions and conclusions

The MIDST is still under development and intended to provide a formal system, represented in $\boldsymbol{L}_{md}$, for natural semantics of space and time. This system is one kind of applied predicate logic consisting of axioms and postulates (e.g., **PRS** is Section 2) subject to human perceptive processes of space and time, while the other similar systems in Artificial Intelligence [15], [16], [17] are objective, namely, independent of human perception and do not necessarily keep tight correspondences with natural language.

The attribute spaces for humans correspond to the sensory receptive fields in their brains. At present, about 50 attributes and 6 categories of standards (e.g., Rigid standard, Species standard) [1] concerning the physical world have been extracted from a Japanese and an English thesaurus. Event patterns such as shown in Fig.3 are the most important for our approach and have been already reported concerning several kinds of attributes [2], [5].

The cross-media operations between texts in several languages (Japanese, Chinese, Albanian and English) and pictorial patterns like maps were successfully implemented on our intelligent system IMAGES-M. At our best knowledge, there is no other system that can perform cross-media operations in such a seamless way as ours [12], [13]. This leads to the conclusion that our locus formula representation has made the logical expressions of event concepts remarkably computable and has proved to be very adequate to systematize cross-media operations. This adequacy is due to its medium-freeness and its good correspondence with the performances of human sensory systems in both spatial and temporal extents while almost all other knowledge representation

schemes are ontology-dependent or spatial-event-unconscious.

Our future work will include establishment of learning facilities for automatic acquisition of word concepts from sensory data [5] and human-robot communication by natural language under real environments [14].

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] M.Yokota: "An approach to natural language understanding based on a mental image model," Proc. of the 2nd International Workshop on Natural Language Understanding and Cognitive Science, pp.22-31,2005.

[2] 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.

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

[4] 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.

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

[6] R.W. Langacker : Concept, Image and Symbol, Mouton de Gruyter, Berlin/New York, 1991.

[7] G.A. Miller, P.N.: Johnson-Laird : Language and Perception, Harvard University Press, 1976.

[8] A.Yamada, et.al.: "Reconstucting spatial image from natural language texts," Proc. of COLING 90, Nantes, 1992

[9] P.Olivier, J.Tsujii: "A Computational View of the Cognitive Semantics of Spatial Expressions," Proc. of ACL 94, Las Cruces, 1994.

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

[11] A.R.Shariff, M.Egenhofer, D.Mark: "Natural-Language Spatial Relations Between Linear and Areal Objects: The Topology and Metric of English-language Terms," International Journal of Geographical Information Science, 12-3,pp.215-246, 1998.

[12] 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, January, 1999.

[13] M.L.Kherfi, D.Ziou, A.Bernardi: "Image Retrieval from the World Wide Web : Issues, Techniques and Systems," ACM Computer Surveys, Vol.36-14, pp.35-67, 2004.

[14] M.Yokota, M.Shiraishi, G.Capi: "Human-robot communication through a mind model based on the Mental Image Directed Semantic Theory," Proc. of the 10th International Symposium on Artificial Life and Robotics (AROB '05), Oita, Japan, pp.695-698, 2005.

[15] J.F.Allen: "Towards a general theory of action and time," Artificial Intelligence, Vol.23-2, pp.123-154, 1984.

[16] D.V.McDermott: "A temporal logic for reasoning about processes and plans," Cognitive Science, Vol.6, pp.101-155, 1982.

[17] Y.Shoham: "Time for actions: on the relationship between time, knowledge, and action," Proc. of IJCAI 89, Detroit, MI, pp.954-959, 1989.