Towards integrated multimedia understanding for intuitive human-system interaction

Masao Yokota
Fukuoka Institute of Technology
yokota@fit.ac.jp

Abstract

The Mental Image Directed Semantic Theory (MIDST) has proposed an omnisensory mental image model and its description language $\boldsymbol{L_{md}}$ *intended to facilitate intuitive human-system interaction such that happens between non-expert people and GISs (Geographic Information Systems). This paper presents a systematic method for formulating and computing natural concepts (i.e., mental images) of physical reality in* $\boldsymbol{L_{md}}$ *and its application to spatial language understanding in view of cross-media operation on text and picture.*

1. Introduction

In the field of ontology, special attention has been paid to spatial language covering geography because its constituent concepts stand in highly complex relationships to underlying physical reality, accompanied with fundamental issues in terms of human cognition (for example, ambiguity, vagueness, temporality, identity, ...) appearing in varied subtle expressions [1]. Most of the traditional approaches to spatial language understanding have focused on computing purely objective geometric relations (i.e., topological, directional and metric relations) conceptualized as spatial prepositions or so, considering properties and functions of the objects involved [e.g., 2]. Such verb-centered expressions as S1 and S2, however, are assumed to reflect not much the purely objective geometrical relations but very much certain dynamism at human perception of the objects involved because they can refer to the same scene in the external world. This is also the case for S3 and S4 and we often encounter such intuitive spatial expressions in our daily life.

(S1) The path sinks to the brook.
(S2) The path rises from the brook.
(S3) The roads meet there.
(S4) The roads separate there.

Anyway, this fact may lead to a certain barrier preventing non-expert or ordinary people and computer systems from comprehensible communication in natural language in such a way as shown in Fig.1. Therefore, their semantic descriptions should be grounded in human perceptual representations, possibly, cognitively inspired and coping with all kinds of spatial expressions including such verb-centered ones as S1-S4 as well as preposition-centered ones. The Mental Image Directed Semantic Theory (MIDST) [3] has proposed a dynamic model of human perception yielding omnisensory image of the world and classified natural event concepts (i.e., event concepts in natural language) into two types of categories, 'Temporal Events' and 'Spatial Events'. These are defined as temporal and spatial changes (or constancies) in certain attributes of physical objects, respectively, with S1-S4 included in the latter. Both the types of events are uniformly analyzable as ***temporally*** parameterized loci in attribute spaces and describable in a formal language $\boldsymbol{L_{md}}$.

This paper presents a brief sketch of $\boldsymbol{L_{md}}$ and a systematic method to formulate and compute natural concepts of physical reality comprising spatial language semantics in order to facilitate intuitive huamn-system interaction, that is, interaction between non-expert people and computer systems such as GISs (Geographical Information Systems). This work is intended to model a more intuitive ontology of space and time by generalizing our concerned findings [e.g., 3-5] and to apply it to intuitive cross-media operation on text and picture.

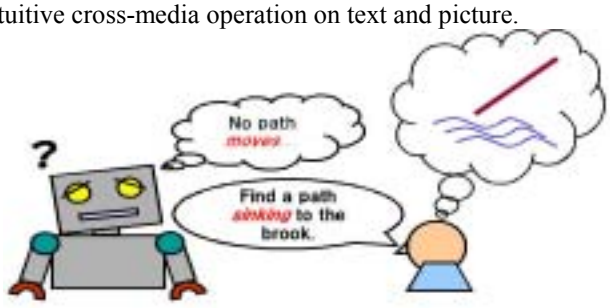

Fig.1. Miscommunication due to different perceptual groundings.

2. Mental Image Description Language L_{md}

2.1 Omnisensory Image Model

The MIDST treats word meanings 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. Such a locus is to be articulated by "Atomic Locus" with an ***absolute*** time-interval $[t_i, t_f]$ $(t_i < t_f)$ as depicted in Fig.2 (left) and formulated as (1).

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

This formula is called 'Atomic Locus Formula' whose first two arguments are often referred to as 'Event Causer (EC)' and 'Attribute Carrier (AC)', respectively. A logical combination of atomic locus formulas defined as a well-formed formula (i.e., wff) in predicate logic is called simply 'Locus Formula'.

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

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

When $g=G_t$ and $g=G_s$, the locus indicates monotonic change or constancy of the attribute in time domain and that in space domain, respectively. The former is called 'temporal event' and the latter, 'spatial event'. For example, the motion of the 'bus' represented by S5 is a temporal event and the ranging or extension of the 'road' by S6 is a spatial event whose meanings or concepts are formulated as (2) and (3), respectively, where 'A_{12}' denotes 'Physical Location'. These two formulas are different only at the term 'Event Type (i.e., *g*)'.

(S5) The bus runs from Tokyo to Osaka.

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

(S6) The road runs from Tokyo to Osaka.

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

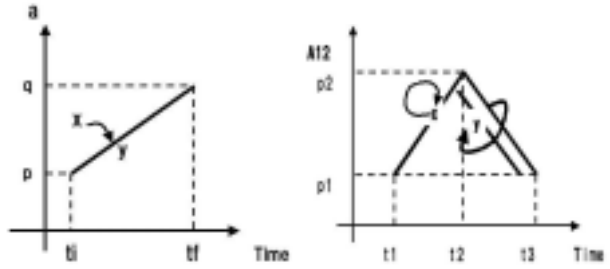

Fig.2. Atomic Locus (left) and Locus of 'fetch' (right).

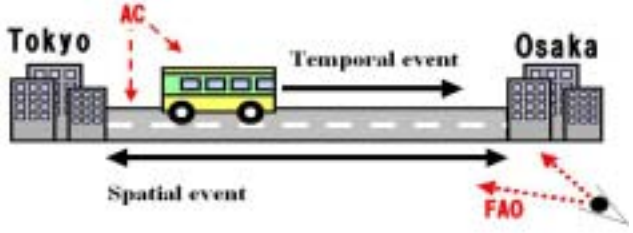

Fig.3. FAO movements and Event types.

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) [4]. To be brief, it is assumed that the FAO is fixed on the whole AC in a temporal event but *runs* about on the AC in a spatial event. According to this assumption, as shown in Fig.3, the *bus* and the FAO move together in the case of S5 while the FAO solely moves along the *road* in the case of S6.

2.2 Tempo-logical connectives

The definition of a tempo-logical connective K_i is given by **D1**, where τ_i, χ and K refer to one of ***purely*** temporal relations indexed by an integer '*i*', a locus, and an ordinary binary logical connective such as the conjunction '$\wedge$', respectively. The definition of each τ_i is provided with Table 1 implying the trivial theorem **T1,** where the durations of χ_1 and χ_2 are $[t_{11}, t_{12}]$ and $[t_{21}, t_{22}]$, respectively. This table shows the complete list of temporal relations between two intervals, where 13 types of relations are discriminated by the suffix '*i*' ($-6\leq i \leq 6$). This is in accordance with Allen's notation [6] which, to be strict, is for 'temporal conjunctions ($=\wedge_i$)' but not for pure 'temporal relations ($=\tau_i$)'.

D1. $\chi_1 K_i \chi_2 \leftrightarrow (\chi_1 K \chi_2) \wedge \tau_i(\chi_1, \chi_2)$

T1. $\tau_{-i}(\chi_2, \chi_1) \equiv \tau_i(\chi_1, \chi_2)$ $(\forall i \in \{0,\pm1,\pm2,\pm3,\pm4,\pm5, \pm6\})$

The temporal connectives used most frequently are the temporal conjunctions 'SAND ($\wedge_0$)' and 'CAND ($\wedge_1$)', standing for 'Simultaneous AND' and 'Consecutive AND', conventionally denoted by 'Π' and '•', respectively. Employing these connectives, for example, the English verb concept 'fetch' can be defined as (4) and depicted as Fig.2 (right). Furthermore, the underlined part of (4) stands for the concept of 'carry' and this relation can be formulated as (5) employing the temporal implication '$\supset_{-4}$', reading that an event 'fetch(x,y)' is necessarily *finished by* an event 'carry(x,y)'. This kind of formula is not an axiom but a theorem deducible from the definitions of event concepts in the formal system.

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

$fetch(x,y) \supset_{-4} carry(x,y)$ (5)

Table 1. List of temporal relations

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

2.3 Empty event

An 'Empty Event (EE)', denoted by 'ε', stands for nothing but for ***absolute*** time elapsing and is explicitly

defined as **D2** with the attribute 'Time Point (A_{34})' and the Standard of absolute time 'T_a'. Usually people can know only a certain ***relative*** time point by a clock that is seldom exact and that is to be denoted by another Standard in the $\boldsymbol{L_{md}}$. Hereafter, $\boldsymbol{\Delta}$ denotes the total set of absolute time intervals. According to this scheme, the suppressed absolute time-interval $[t_a, t_b]$ of a locus χ can be indicated as (6).

D2. $\varepsilon([t_i,t_j])\leftrightarrow(\exists x,y,g)L(x,y,t_i,t_j,A_{34},g,T_a)$,
where $[t_i, t_j]\in\boldsymbol{\Delta}\,(=\{[t_1, t_2] \mid t_1<t_2\ (t_1, t_2\in\boldsymbol{R})\})$. □
$\chi\Pi\varepsilon([t_a,t_b])$ (6)

3. Semantic description of physical reality

3.1 Event concepts

The semantic description of an event is compared to a movie film recorded through a floating camera because it is necessarily grounded in FAO's movement over the event. Therefore, as already pointed out, S1 and S2 can refer to the same scene in spite of their appearances, where what 'sinks' or 'rises' is FAO and whose conceptual descriptions are given as (7) and (8), respectively, where the special symbol '_' is defined by (9), standing for an anonymous variable bound by an existential quantifier, and 'A_{13}', '↑' and '↓' refer to the attribute 'Direction', and its values 'upward' and 'downward', respectively. Such a fact is generalized as '***Postulate of Reversibility of a Spatial Event*** (PRS)' that can be one of the principal inference rules belonging to people's intuitive knowledge about geography. This postulation is also valid for such a pair of S7 and S8 as interpreted approximately into (10) and (11), respectively. These pairs of conceptual descriptions are called ***equivalent in the PRS***, and the paired sentences are treated as ***paraphrases*** each other.

$(\exists y,p,z)L(\_,y,p,z,A_{12},G_s,\_)\Pi L(\_,y,\downarrow,\downarrow,A_{13},G_s,\_)$
$\wedge$path(y) $\wedge$brook(z) $\wedge p\neq z$ (7)

$(\exists y,p,z)L(\_,y,z,p,A_{12},G_s,\_)\Pi L(\_,y,\uparrow,\uparrow,A_{13},G_s,\_)$ $\wedge$path(y)
$\wedge$brook(z) $\wedge p\neq z$ (8)

$L(\ldots,\_,\ldots)\leftrightarrow(\exists\omega)L(\ldots,\omega,\ldots)$ (9)

(S7) Route A and Route B meet at the city.

$(\exists p,y,q)L(\_,Route\_A,p,y,A_{12},G_s,\_)\Pi$
$L(_,Route_B,q,y,A_{12},G_s,_)\wedge$city(y) $\wedge p\neq q$ (10)

(S8) Route A and Route B separate at the city.

$(\exists p,y,q)L(\_,Route\_A,y,p,A_{12},G_s,\_)\Pi$
$L(_,Route_B,y,q,A_{12},G_s,_)\wedge$city(y)$\wedge p\neq q$ (11)

For another example of spatial event, Fig.4 (up) 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 S9 using the preposition 'between' and formulated as (12), corresponding also to such concepts as 'row', 'line-up', etc.

(S9) □ is between Δ and ○.

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

For more complicated examples, consider S10 and S11. The underlined parts are deemed to refer to some events neglected in time and in space, respectively. These events correspond with skipping of FAOs and are called 'Temporal Empty Event' and 'Spatial Empty Event', denoted by 'ε_t' and 'ε_s' as Empty Events with $g=G_t$ and $g=G_s$ at (6), respectively. Their concepts are described as (13) and (14), where 'A_{15}' and 'A_{17}' represent the attribute 'Trajectory' and 'Mileage', respectively. From the viewpoint of cross-media reference, the formula (14) can refer to such a spatial event depicted as the still picture in Fig.4 (down) while (13) is to be interpreted into a motion picture.

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

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

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

$(\exists x,y,z,p,q)(L(\_,x,A,B,A_{12},G_s,\_)\Pi$
$L(\_,x,0,10km,A_{17},G_s,\_)\Pi L(\_,x,Point,Line,A_{15},G_s,\_)\Pi$
$L(\_,x,East,East,A_{13},G_s,\_))\bullet\varepsilon_s$ $\bullet(L(\_,x,p,C,A_{12},G_s,\_)$
$\Pi L(\_,y,q,C,A_{12},G_s,\_)\Pi L(\_,z,y,y,A_{12},G_s,\_))$
$\wedge$road(x)$\wedge$street(y)$\wedge$sidewalk(z)$\wedge p\neq q$ (14)

There are a considerable number of postulates of space and time to facilitate intuitive interaction between humans and IMAGES-M [4], and the PRS (Postulate of Reversibility of a Spatial Event) is one of the most important. This postulate can be formulated as (15) using '$\equiv_0$', where χ and χ^R is a locus formula and its 'reversal' for a certain spatial event, respectively. The recursive operations to transform χ into χ^R are defined by (16)-(18), where the reversed values p^R and q^R depend on the properties of p and q. For example, (14) is transformed into (19) to be verbalized as S12, where $p^R=p$ and $q^R=q$ for A_{12}; $p^R=-p$ and $q^R=-q$ for A_{13}.

$\chi^R.\equiv_0.\chi$ (15)
$(\chi_1\bullet\chi_2)^R\leftrightarrow\chi_2{}^R\bullet\chi_1{}^R$ (16)
$(\chi_1\Pi\chi_2)^R\leftrightarrow\chi_1{}^R\,\Pi\chi_2{}^R$ (17)
$(L(x,y,p,q,a,G_s,k))^R\leftrightarrow L(x,y,q^R,p^R,a,G_s,k)$ (18)

$(\exists x,y,z,p,q)(L(\_,x,C,p,A_{12},G_s,\_)\Pi$
$L(\_,y,C,q,A_{12},G_s,\_)\Pi L(\_,z,y,y,A_{12},G_s,\_))\bullet\varepsilon_s\bullet$
$(L(\_,x,B,A,A_{12},G_s,\_)\Pi L(\_,x,0,10km,A_{17},G_s,\_)\Pi$
$L(\_,x,Point,Line,A_{15},G_s,\_)\Pi L(\_,x,West,West,A_{13},G_s,\_))$
$\wedge$road(x)$\wedge$street(y)$\wedge$sidewalk(z)$\wedge p\neq q$ (19)

(S12) The road separates at *C* from the street with the sidewalk and, after a while, runs 10km straight west from *B* to *A*.

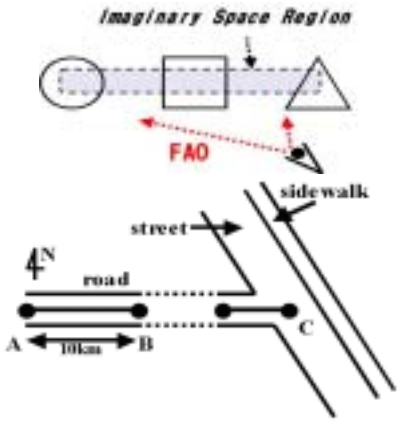

Fig.4. Complicated spatial events: 'row' (left) and 'example of road map' (right).

3.2 Object concepts

A physical object can be semantically defined as a combination of its properties and its relations with others. For example, the semantic descriptions of 'rain', 'wind' and 'air' can be given as (20)-(22), reading 'Rain is water attracted from the sky by the earth, makes an object wetter, is pushed an umbrella to by a human,…,' 'Wind is air, affects the direction of rain,… ,' and 'Air has no shape, no taste, no vitality, …,' respectively. The special symbols '*' and '/' are defined as (23) and (24) representing 'always' and 'no value', respectively.

$(\lambda x)\mathrm{rain}(x)\leftrightarrow(\lambda x)(\exists x_1,x_2,\ldots)\mathrm{L}(\_,x,x_1,x_1,A_{41},G_t,\_)$
$\prod\mathrm{L}(\mathrm{Earth},x,\mathrm{Sky},\mathrm{Earth},A_{12},G_t,\_)\prod\mathrm{L}(x,x_2,p,q,A_{25},G_t,\_)$
$\prod\mathrm{L}(x_3,x_4,x,x,A_{19},G_t,x_3)\wedge\mathrm{water}(x_1)$
$\wedge\mathrm{object}(x_2)\wedge\mathrm{human}(x_3)\wedge\mathrm{umbrella}(x_4)\wedge(p<q)\ldots$ (20)

$(\lambda x)\mathrm{wind}(x)\leftrightarrow(\lambda x)(\exists x_1,x_2,\ldots)\mathrm{L}(\_,x,x_1,x_1,A_{41},G_t,\_)$
$\wedge\mathrm{air}(x_1)\wedge(\mathrm{L}(x,x_2,p,q,A_{13},G_t,\_)\wedge\mathrm{rain}(x_2)\ldots$ (21)

$(\lambda x)\mathrm{air}(x)\leftrightarrow(\lambda x)(\ldots\wedge\mathrm{L}^*(\_,x,/,/,A_{11},G_t,\_)\wedge\ldots\wedge$
$\mathrm{L}^*(\_,x,/,/,A_{29},G_t,\_)\wedge\ldots\wedge\mathrm{L}^*(\_,x,/,/,A_{39},G_t,\_)\wedge\ \ldots)$ (22)

$X^*\leftrightarrow(\forall[p,q])X\ \Pi\ \varepsilon([p,q])$ (23)

$\mathrm{L}(\ldots,/,\ldots)\leftrightarrow\sim(\exists p)\ \mathrm{L}(\ldots,\omega,\ldots)$ (24)

4. Discussions and conclusions

The cross-media operations between texts in several languages and pictorial patterns like maps were successfully implemented on our intelligent system IMAGES-M. Figures 5 and 6 show examples of Q-A on a map and text-to-action translation by IMAGES-M, respectively, where several kinds of intuitive postulates such as PRS were effectively utilized. In this research area, it is most conventional that conceptual contents conveyed by information media such as language and picture are represented in computable forms independent of each other and translated via 'transfer' processes which are often specific to task domains [e.g., 7-9]. That is, at my best knowledge, there is no other system that can perform cross-media operation in such a seamless way as mine.

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

H: What is between the buildings A and B?
S: The railway D.
H: Where do the street A and the road B meet?
S: At the crossing C.
H: Where do the street A and the road B separate?
S: At the crossing C.

Fig.5. Q-A on a map between a human (H) and IMAGES-M (S).

Fig.6. Text-to-Action translation by IMAGES-M: 'Sit down *AND* wave your left hand' was interpreted as 'Sit down *BEFORE* waving your left hand.'

References

[1] Geo-ontology Concepts and Issues. Report of a workshop on Geo-ontology (Compiled and edited by Jenny Harding), Ilkley UK, Sept. 2002.

[2] Coventry, K. R., Prat-Sala, M., Richards, L. V.: "The interplay between geometry and function in the comprehension of 'over', 'under', 'above' and 'below'," Journal of Memory and Language, 4, 376-398, 2001.

[3] Yokota,M.: "An Approach to Integrated Spatial Language Understanding Based on Mental Image Directed Semantic Theory," Proc. of 5th Workshop on Language and Space, Bremen, Germany, Oct., 2005.

[4] Yokota,M.: "A psychological experiment on human understanding process of natural language," Trans. of IEICE Japan, J71D-10, pp.2120-2127, 1988.

[5] Yokota, M. & Capi,G.: "Cross-media Operations between Text and Picture Based on Mental Image Directed Semantic theory," WSEAS Trans. on INFORMATION SCIENCE and APPLICATIONS, Issue 10, 2, pp.1541-1550, Oct. 2005.

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

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

[8]Shariff,A.R.,Egenhofer,M.,Mark,D.:"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.

[9]Coradeschi,S., Saffiotti,A.: "An Introduction to the Anchoring Problem," Robotics and Autonomous Systems 43, pp.85-96, 2003.