2004－DPS－120　(16)
2004／11／5

Linguistic Interpretation of Human Motion Based on a Multimedia Description Language

Mikio Amano, Daisuke Hironaka, Seio Oda, Genci Capi, Masao Yokota

Faculty of Information Engineering
Fukuoka Institute of Technology,
E-mail: yokota@fit.ac.jp

Abstract

The Mental Image Directed Semantic Theory (MIDST) has proposed a formal language L_{md} that is employed for many-sorted predicate logic and whose semantics is given based on an omnisensual mental image model. This language is used for intermediate knowledge representation of multimedia contents and has been applied to integrated multimedia understanding This paper presents a brief sketch of L_{md} and its application to linguistic interpretation of human motion data obtained through a motion capture system, which we think will serve for a basis of linguistic summarization of immense video data such as movies.

1. Introduction

Rapid increase of matured societies will necessarily bring a large number of handicapped people due to aging and thereby serious shortage of workers for them such as nurses one of whose routines is to survey their actions through TV monitors in order to prevent accidents.

However it is unfeasible and privacy-invasive for human workers to watch TV monitors for such surveillance all day and night long. For such a purpose, certain kinds of intelligent systems can be very helpful that facilitate exploiting significant parts such as abnormal motions from video data and reporting it in natural language in real time and desirably linguistic summarization of immense amount of recorded ones.

M.Yokota, et al have proposed a formal language named L_{md} for multimedia information description in the Mental Image Directed Semantic Theory (MIDST) [1], [11] and thereby a methodology for automatic understanding of various information media through intermediate knowledge representation [2], [3]. The language L_{md} is employed for many-sorted first-order predicate logic containing one special predicate called ‘Atomic locus’ 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 such as words [4]-[10].

Employing L_{md}, D.Hironaka et al have been developing an integrated multimedia understanding system IMAGES-M that can facilitate, for example, cross-media reference including bi-directional translation between different kinds of media such as text and picture [19]. M.Amano et al have also been applying L_{md} to systematic linguistic interpretation of human motions [12].

There have already been reported a considerable number of works on computer interpretation of human motions including facial expressions [13]-[18]. At our best knowledge, almost all of their methods were very specific to a single goal, for example, improvement in animation data, and implemented as procedures without explicitly logical representation of human motions that we believe indispensable for systematic bi-directional translation between numerical motion data and linguistic expressions.

This paper presents a brief sketch of the multimedia description language L_{md} and its application to systematic linguistic interpretation of human motion data that are obtained through a motion capturing system.

2. Multimedia description language, L_{md}

2.1 Locus formulas

MIDST treats word meanings in association with mental images, not limited to visual but omnisensual, modeled as “Loci in Attribute Spaces” [1], [11]. 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.

A general locus is to be articulated by “Atomic Locus” formalized as the expression (1). This is a formula in many-sorted first-order predicate logic, where “L” is a predicate constant called ‘Atomic locus’ 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’).

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

The formula is called “Atomic Locus Formula” whose arguments are referred to as ‘Event Causer ’ and ‘Attribute Carrier ’, ‘Initial attribute value’, ‘Final attribute value’, ‘Attribute’, ‘Event pattern’, and ‘Standard’, respectively. For simplicity, a matter term is often put in the place of ‘Event Causer’, ‘Attribute Carrier’ or ‘Standard’ in order to represent its attribute value at the time.

The interpretation of the expression (1) is intuitively given as follows, where “matter” refers to “object ” or “event”.

“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 monotonous 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 formalized as (2) and (3), respectively, where the attribute is “physical location ” denoted by ‘A12’.

(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.

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

The expression (4) is the conceptual description of the English word “fetch”, implying such a temporal event that ‘*x1*’ goes for ‘*x2*’ and then comes back with it, where ‘Π’and ‘•’ are instances of the tempo-logical connectives, ‘SAND’ and ‘CAND’, standing for “Simultaneous AND” and “Consecutive AND”, respectively. In general, a series of atomic locus formulas with such connectives is simply called ‘Locus formula’.

$$\begin{aligned}&(\exists x1,x2,p1,p2,k)\ L(x1,x1,p1,p2,A12,Gt,k)\\ &\bullet (L(x1,x1,p2,p1,A12,Gt,k)\\ &\Pi\, L(x1,x2,p2,p1,A12,Gt,k))\wedge x1\neq x2\wedge p1\neq p2 \qquad (4)\end{aligned}$$

Furthermore, a very important concept called ‘Empty Event (EE)’ and symbolized as ‘ε’ must be introduced. An EE stands for nothing but for time collapsing and is explicitly defined as (5) with the attribute ‘Time Point (A34)’. It is essentially significant for the MIDST that ***every temporal relation can be represented by a combination of Empty Events, SANDs and CANDs***. For example, (6) represents ‘X_1 during X_2 ’.

$$\varepsilon \Leftrightarrow (\exists x,t,p,q,g,k)\ L(x,t,p,q,A34,g,k)\wedge time(t) \qquad (5)$$

$$(\varepsilon_1 \bullet X_1 \bullet \varepsilon_2)\ \Pi\, X_2 \qquad (6)$$

2.2 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 1. 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.

Table 1. A part of attributes extracted from linguistic expressions.

+The properties “S” and “V” represent “scalar” and “vector”, respectively.

Code	Attribute [Property]	Linguistic expressions for attribute values.
*A01	PLACE OF EXISTE NCE [V]	He is in Tokyo. The accident happened in Osaka.
*A02	LENGTH [S]	The stick is 2 meters long (in length).
	………………………………………	
A09	AREA [S]	The crop field is 10 square miles.
A10	VOLUME [S]	The box 10 cubic meters.
*A11	SHAPE [V]	The cake is round.
*A12	PHYSICAL LOCATION [V]	Tom moved to Tokyo.
*A13	DIRECTION [V]	The box is to the left of the chair.
*A14	ORIENTATION [V]	The door faces to south.
*A15	TRAJECTORY [V]	The plane circled in the sky.
*A16	VELOCITY [S]	The boy runs very fast.
*A17	DISTANCE [S]	The car ran ten miles.
A18	STRENGTH OF EFFECT [S]	He is very strong.
	………………………………………	
A32	COLOR [V]	The apple is red. Tom painted the desk white.
A33	INTERNAL SENSATION [V]	I am very tired.
*A34	TIME POINT [S]	It is ten o’clock.
	………………………………………	

Table 2. Standards of attribute values.

Categories of standards	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*'.

Correspondingly, six categories of standards shown in Table 2 have been extracted that are assumed necessary for representing values of each attribute in Table 1. In general, the attribute values represented by words are relative to certain standards as explained briefly in Table 2.

3. Principles of cross-media translation

3.1 Functional requirements

The authors have considered that systematic cross-media translation and which in turn 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.

3.2 Formalization

According to MIDST, any content conveyed by an information medium is assumed to be associated with the loci in certain attribute spaces and in turn the world describable by each medium can be characterized by the maximal set of such attributes. This relation is conceptually formalized by the expression (7), 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.

$$F(Wm)=\{Am_1, Am_2,\ldots, Am_n\} \quad (7)$$

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 (8).

$$Y(Smt)=\psi(X(Sms)), \quad (8)$$

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* ,

and

ψ : the function for transforming *X* into *Y*, so called, 'Locus formula paraphrasing function' which is designed to realize all the functions F1-F4 by inference processing at the level of locus formula representation.

According to the formalization here, the cross-media translation from human motion data (HMD) to text can be basically specified as Table 3

3.3 Locus formula paraphrasing function ψ

In order to satisfy F1, a certain set of '***Attribute paraphrasing rules (APRs)***', so called, are defined ***at every pair of source and target media*** (See Section 4.5).

Table 3. Fundamental specifications for HMD-to-text translation.

	Categories of media	Maximal sets of attributes - Sms and Smt	Categories of standards	UI-Devices
Source medium	Human motion data	Sms ={A12, A34}	Rigid Standard	MC system
Target medium	Natural language texts	Smt = All of Table 1	All of Table 2	TV monitor

The function F2 is satisfied 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 the function F3, ***default reasoning*** is employed. That is, such an inference rule as defined by the expression (9) 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 (10) 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 *'M'*.

$$X \circ Y \rightarrow Z \qquad (9)$$

$$\begin{aligned} \{\ & X / (L(x,y,p,p,A,G,k) \wedge M(y)) \\ & \quad \wedge (L(z,O\#,P,P,A,G,K) \wedge M(O\#)), \\ & Y / p=P \wedge k=K, \\ & Z /\ L(x,y,P,P,A,G,K) \wedge M(y)\ \} \end{aligned} \qquad (10)$$

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

4. Human motion data to text translation

4.1 Structural description of human body

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 (11) and (12) 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.

$$\begin{aligned} & (\lambda x)arm(x) \Leftrightarrow (\lambda x)(\exists y1,y2,k) \\ & L(x,x,y1,y2,A12,Gs,k) \wedge shoulder(y1) \wedge hand(y2) \end{aligned} \qquad (11)$$

$$\begin{aligned} & (\lambda x)wrist(x) \Leftrightarrow (\lambda x)(\exists y1,y2,y3,y4,k) \\ & (L(y1,y1,y2,x,A12,Gs,k) \bullet L(y1,y1,x,y3,A12,Gs,k)) \\ & \wedge body\text{-}part(y1) \wedge forearm(y2) \wedge hand(y3) \end{aligned} \qquad (12)$$

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 S3 [1].

(S3) The left arm moved away from the left shoulder and the left hand.

4.2 Conceptual description of human motion

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 the verb 'nod' is given by (13) 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.

$$\begin{aligned} & (\lambda x)\ nodding(x) \Leftrightarrow (\lambda x)\ (\exists y1,y2,k1,k2,k3) \\ & L(y1,\{y1,y2\},x,x,A01,Gt,k1) \\ & \ \Pi L(y1,y2,Down,Down,A13,Gt,k2) \\ & \ \Pi L(y1,y2,Forward,Forward,A13,Gt,k3) \\ & \wedge person(y1) \wedge head(y2) \wedge motion(x) \end{aligned} \qquad (13)$$

4.3 Motion data acquisition

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 capture system (MC system) at a sampling rate. Figure 1 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 (11) and (12). Real motion data were graphically interpreted according to the model as shown in Fig.3.

The smallest motion datum can be formally denoted 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.

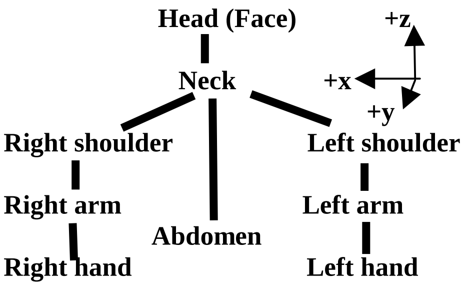


Fig.1 Wire frame model of upper half of human body

4.4 Motion meaning representation (MMR)

For example, a large number of motion data of the subject's head over a time interval are digested into a locus formula such as (14) where 'Tom' is the default name of the subject and Ps 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 rigid standards specific to the motion capturing system.

$$L(Tom,Head,P_1,P_2,A12,Gt,Mc)\bullet$$
$$L(Tom,Head,P_2,P_3,A12,Gt,Mc)\bullet\ldots\bullet$$
$$L(Tom,Head,P_{n-2},P_{n-1},A12,Gt,Mc)\bullet$$
$$L(Tom,Head,P_{n-1},P_n,A12,Gt,Mc) \qquad (14)$$

4.5 Attribute paraphrasing rule (APR)

Human motion data gained through a motion capture system associate limitedly with the sense 'sight' and its related attributes are A12 (physical position) and A34 (Time point) in Table 1.

In translation between motion data and texts, these two attributes and those marked with '*' in Table 1 are to be paraphrased with each other according to 'Attribute paraphrasing rules (APRs)' such as (15)-(17), where the left and right hands of the symbol '⇔'refer to the attributes concerning to MMRs and TMRs, respectively.

$$(\exists p,q)L(y1,y2,p,q,A12,Gt,Mc)\wedge q\neq p$$
$$\wedge p=(x_p,y_p,z_p)\wedge q=(x_q,y_q,z_q)$$
$$\Leftrightarrow(\exists x,k)L(y1,\{y1,y2\},x,x,A01,Gt,k)\wedge motion(x) \qquad (15)$$

$$(z_q\text{-}z_p<0, A12)\Leftrightarrow(Down, A13) \qquad (16)$$

$$(y_q\text{-}y_p>0, A12)\Leftrightarrow(Forward, A13) \qquad (17)$$

4.6 Text meaning representation (TMR)

Based on APRs (15)-(17), the MMR (14) is unified with (13), namely, the conceptual description of the verb 'nod', which yields the expression (18) called 'Text meaning representation (TMR)'.

$$(\exists k1,k2,k3)L(Tom,\{Tom,Head\},x,x,A01,Gt,k1)$$
$$\Pi L(Tom,Head,Down,Down,A13,Gt,k2)$$
$$\Pi L(Tom,Head,Forward,Forward,A13,Gt,k3)$$
$$\wedge person(Tom)\wedge head(Head)\wedge motion(x) \qquad (18)$$

The sentence 'Tom nodded.' is to be generated from this TMR using the sentence pattern of 'nod' which is generalized as '***y1* nod**' indicating the correspondence between the subject of the verb and the term '*y1*' in its conceptual description (13).

4. Experiment

The methodology mentioned above has been implemented on the intelligent system IMAGES-M shown in Fig.2. IMAGES-M is one kind of expert system equipped with five kinds of user interfaces besides the inference engine (IE) and the knowledge base (KB) as follows.

(U1) Text Processing Unit (TPU),
(U2) Speech Processing Unit (SPU),
(U3) Picture Processing Unit (PPU),
(U4) Animation Processing Unit (APU), and
(U5) Sensory Data Processing Unit (SDPU).

For motion-to-text translation, SDPU and TPU collaborate exclusively among these interfaces. In this case, the system works in a top-down way, that is, it tries to find such motions that can be verbalized only by using the verbs stored in KB in advance. The details of this process are as follows.

(STEP-1) SDPU takes in motion data from the motion capture system and digests it into an MMR and transfers it to IE.

(STEP-2) IE, using the conceptual descriptions of verbs and APRs stored in KB, tries to detect any verb concept within the MMR, where detection of a verb means success in generation of its TMR.

(STEP-3) TPU receives the TMR from IE and verbalizes it using the sentence patterns of the detected verb.

Figure 3-1 to 3 are graphical interpretations of the real motion data at the time points t1, t2 and t3, respectively. The sets of real motion data over time intervals [t1, t2] and [t2, t3] were translated into the texts in Figure.4-a and b, respectively.

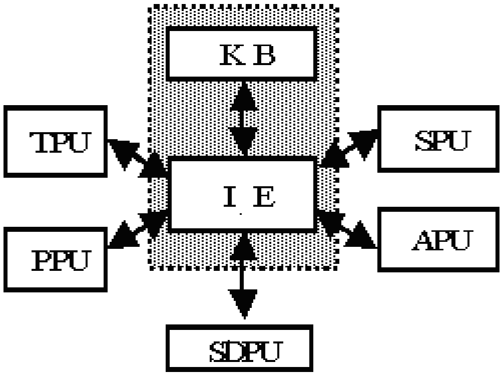
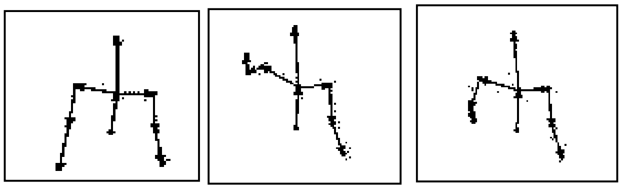


Fig. 2 Configuration of IMAGES-M



(1) Data at t1 (2) Data at t2 (3) Data at t3

Fig.3 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 t1 to t2.

……………
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 t1 to t2.

Fig.4 Texts generated from real motion data.

5. Discussion and conclusion

We have proposed a methodology for systematic linguistic interpretation of human motion data based on MIDST, implemented it on the intelligent system IMAGES-M and confirmed its validity for about 30 verb concepts such as ‘raise’ and ‘nod’. Our work’s most remarkable advance to the others resides in the transparency of the processing algorithms due to the formal language L_{md}.

The future problems of our project are as follows:

(P1) Treatment of more complicated verb concepts such as ‘crouch’ and ‘crawl’.

(P2) Realization of the reverse process, namely, text-to-motion translation.

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, Project number 14580436.

References

[1] 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.

[2] D. Hironaka, S. Oda, K. Ryu & M. Yokota : “Mutual Conversion of Sensory Data and Texts by an Intelligent System IMAGES-M’’, Proc. of the 8th International Symposium on Artificial Life and Robotics (AROB ’03), pp.141-144, 2003.

[3] 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.

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

[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, in 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] M.Yokota & D.Hironaka : Cross-media Translation Based on Mental Image Directed Semantic Theory toward More Comprehensible Multimedia Communication, Proc. of IEEE AINA-2004, Fukuoka, Japan, March 2004.

[12] M.Amano & M.Yokota : On linguistic recognition of human motions based on the Mental Image Directed Semantic Theory, IEICE Technical report, TL-101-484, pp.7-12, 2001.

[13] K.Mase : Recognition of facial expression from optical flow, IEICE Transactions, Vol. E 74-10, pp.3474-3483,1991.

[14] T.B. Moeslund & E.Granum.: A Survey of Computer Vision-Based Human Motion Capture. *Computer Vision and Image Understanding: CVIU*, 81-3, pp.231–268, 2001.

[15] Y.Yacoob & L.S.Davis : Labeling of human face components from range data, IEEE CVPR, 592-593, 1993.

[16] H.Ren & G.Xu : Human Action Recognition in Smart Classroom, Proc. of Int. Conf. on Automatic Face and Gesture Recognition, pp. 417–422, 2002.

[17] H.Sidenbladh, M.Black & L.Sigal : Implicit Probabilistic Models of Human Motion for Synthesis and Tracking, Proc. of European Conf. on Computer Vision, number 2350 in LNCS, pp.784–800, Springer Verlag, 2002.

[18] J.Sullivan & S.Carlsson : Recognizing and Tracking Human Action, Proc. of European Conf. on Computer Vision, number 2350 in LNCS, pp. 629–644, Springer Verlag, 2002..

[19] D.Hironaka & M.Yokota : Multimedia description language for more intelligent networking, Proc. of 15th workshop on DEXA, pp.318-323, 2004.