The Thirteenth International Symposium on Artificial Life and Robotics 2008(AROB 13th '08),
B-Con Plaza, Beppu, Oita, Japan, January 31-February 2, 2008

Cross-media translation of human motion into text and text into animation based on Mental Image Description Language *Lmd*

Huang He, Li Fan, Kaoru Sugita and Masao Yokota

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

Abstract: The Mental Image Directed Semantic Theory (MIDST) has proposed an omnisensory mental image model and its description language L_{md}. This language can provide intelligent systems with integrated formulation and computation of multimedia contents. That is, L_{md} can work as a pivot language among various information media such as text, picture, gesture etc. and can facilitate cross-media operation among them. This paper presents a brief sketch of the systematic cross-media translation based on L_{md} and its application to cross-media interpretation of human motion data into text and text into animation.

Keywords: Cross-media translation, Knowledge representation language, Human motion, Text, Animation.

I. INTRODUCTION

There have already been reported a considerable number of works on computer interpretation of human motions including facial expressions [1]-[6]. 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 multidirectional translation among numerical human motion data, text and animation. The Mental Image Directed Semantic Theory (MIDST) [7,8] has proposed an omnisensory mental image model and its description language L_{md}. This language can provide intelligent systems with integrated formulation and computation of multimedia contents. That is, L_{md} can work as a pivot language among various information media such as text, picture, gesture etc. and can facilitate cross-media operation among them [9].

This paper presents a brief sketch of the systematic cross-media translation based on L_{md} and its application to cross-media interpretation of human motion data into text and text into animation.

II. CROSS-MEDIA TRANSLATION

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.

The functional requires above are totally exemplified as follows. The text consisting such two sentences as 'There is a hard cubic object' and 'The object is large and red' can be translated into a still picture in such a way as shown in Fig.1.

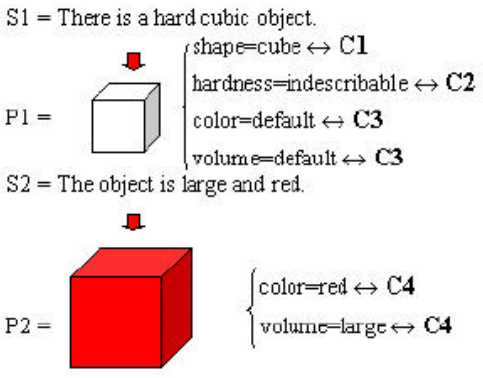

Fig.1. Systematic cross-media translation

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 (1), 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.

$$F(Wm)=\{Am_1, Am_2, \ldots, Am_n\} \quad (1)$$

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 (28).

$$Y(Sm_t)=\psi(X(Sm_s)), \quad (2)$$

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

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***.

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 (3) 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 (4) 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'*.

$$X \circ Y \rightarrow Z \quad (3)$$

$$\{X / (L(x,y,p,p,A,G,k) \wedge M(y)) \wedge (L(z,O\#,P,P,A,G,K) \wedge M(O\#)),\ Y / p=P \wedge k=K,\ Z/L(x,y,P,P,A,G,K) \wedge M(y)\} \quad (4)$$

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

III. MOTION DATA INTO TEXT

1. Structural description of human body

The human body can be described in a computable form using locus formulas [8]. 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 (5) and (6) 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, where A_{12} is the attribute of 'Physical location'.

$$(\lambda x)\text{arm}(x) \Leftrightarrow (\lambda x)(\exists y_1, y_2, k)\ L(x,x,y_1,y_2,A_{12},G_s,k) \wedge \text{shoulder}(y_1) \wedge \text{hand}(y_2) \quad (5)$$

$$(\lambda x)\text{wrist}(x) \Leftrightarrow (\lambda x)(\exists y_1, y_2, y_3, y_4, k)\ (L(y_1,y_1,y_2,x,A_{12},G_s,k) \bullet L(y_1,y_1,x,y_3,A_{12},G_s,k)) \wedge \text{body-part}(y_1) \wedge \text{forearm}(y_2) \wedge \text{hand}(y_3) \quad (6)$$

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 S1 [10].

(S1) The left arm moved away from the left shoulder and the left hand.

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 (7) 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.

$$(\lambda x)\ \text{nodding}(x) \Leftrightarrow (\lambda x)\ (\exists y_1, y_2, k_1, k_2, k_3)\ L(y_1,\{y_1,y_2\},x,x,A_{01},G_t,k_1)\ \Pi L(y_1,y_2,\text{Down},\text{Down},A_{13},G_t,k_2)\ \Pi L(y_1,y_2,\text{Forward},\text{Forward},A_{13},G_t,k_3)\ \wedge \text{person}(y_1) \wedge \text{head}(y_2) \wedge \text{motion}(x) \quad (7)$$

3. Motion data acquisition

As for our experiment, colored markers were put on the parts of the human body and their position data (i.e. 3D coordinates) were taken in through a motion capturing system at a certain sampling rate. Figure 2 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 (5) and (6). Real motion data were graphically interpreted according to the model as shown in Fig.4.

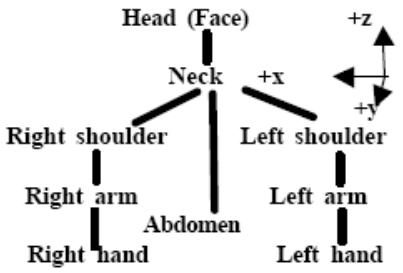

Fig.2. Wire frame model of upper half of human body

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.

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 (8) 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.

$$L(Tom,Head,P_1,P_2,A_{12},G_t,M_c)\bullet$$
$$L(Tom,Head,P_2,P_3,A_{12},G_t,M_c)\ \bullet\ldots\bullet$$
$$L(Tom,Head,P_{n-2},P_{n-1},A_{12},G_t,M_c)\ \bullet$$
$$L(Tom,Head,P_{n-1},P_n,A_{12},G_t,M_c) \quad (8)$$

5. Attribute paraphrasing rule (APR)

Human motion data gained through a motion capturing system associate limitedly with the sense 'sight' and its related attributes are A_{12} (physical position) and A_{34} (Time point).

In translation between motion data and texts, these two attribute are to be paraphrased with each other according to 'Attribute paraphrasing rules (APRs)' such as (9)-(11), where the left and right hands of the symbol '⇔'refer to the attributes concerning MMRs and TMRs, respectively. A_{01} and A_{13} are the attributes of 'Place of existence' and 'Direction', respectively.

$$(\exists p,q)L(y_1,y_2,p,q,A_{12},G_t,M_c)\ \wedge q\neq p$$
$$\wedge p=(x_p,y_p,z_p)\wedge q=(x_q,y_q,z_q)\ \Leftrightarrow$$
$$(\exists x,k)L(y_1,\{y_1,y_2\},x,x,A_{01},G_t,k)\wedge motion(x) \quad (9)$$
$$(z_q-z_p<0,\ A_{12})\ \Leftrightarrow\ (Down,\ A_{13}) \quad (10)$$
$$(y_q-y_p>0,\ A_{12})\ \Leftrightarrow\ (Forward,\ A_{13}) \quad (11)$$

6. Text meaning representation (TMR)

Based on APRs (9)-(11), the MMR (8) is unified with (7), namely, the conceptual description of the verb 'nod', which yields the expression (12) called 'Text meaning representation (TMR)'.

$$(\exists x,k_1,k_2,k_3)L(Tom,\{Tom,Head\},x,x,A_{01},G_t,k_1)$$
$$\Pi L(Tom,Head,Down,Down,A_{13},G_t,k_2)$$
$$\Pi L(Tom,Head,Forward,Forward,A_{13},G_t,k_3)$$
$$\wedge person(Tom)\ \wedge head(Head)\wedge motion(x) \quad (12)$$

The sentence 'Tom nodded.' is to be generated from this TMR using the sentence pattern of 'nod' which is generalized as 'y_1 **nods**' indicating the correspondence between the subject of the verb and the term 'y_1' in its conceptual description (7).

7. Experiment

The methodology mentioned above has been implemented on the intelligent system IMAGES-M [9] shown in Fig.3. 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). Figure 4-1 to 3 are graphical interpretations of the real motion data at the time points t_1, t_2 and t_3, respectively.

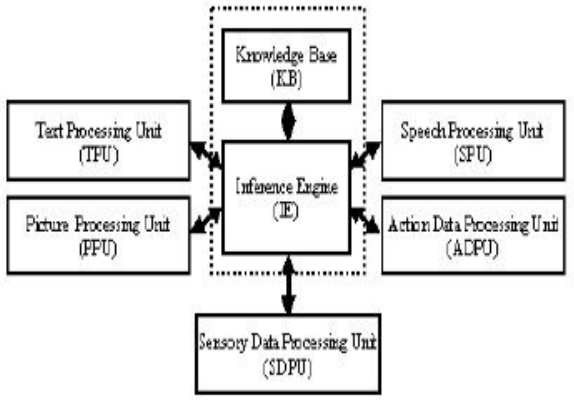

Fig. 3. Configuration of IMAGES-M

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

Fig.4. Graphical interpretations of real motion data

For example, the text generated from the data from t1 to t2 is as follows.

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 b ent the right arm.
……

The following is the text generated from the data from t2 to t3.

……

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

IV. TEXT INTO ANIMATION

Each of such human's/robot's motions (M_k) as 'walk' and 'bow' is given as an ordered set of its standardized characteristic snapshots (S_k) called 'Standard Motion' and defined by (13). In turn, a family (F_X) of S_ks is called 'Family of Standard Motions' and defined by (14), where the suffix 'X' refers to 'human (X=H)' or 'robot (X=R)'.

$$S_k=\{M_{kS}, \ldots, M_{kE}\} \quad (13)$$
$$F_X=\{S_1, S_2, \ldots, M_N\} \quad (14)$$

For example, the L_{md} expression of human walking in default is given by (15), reading that a human moves by his/her legs making his/her shape (A_{11}) change monotonically from $Walk_S$ to $Walk_E$.

$$(\exists x,y,p_1,p_2,q_1,q_2)\, L(\_,y,x,x,A_{01},G_t,\_)\Pi L(y,x,q_1,q_2,A_{12},G_t,\_)\Pi\, L(x,x,Walk_S,Walk_E,A_{11},G_t,F_H) \wedge q_1 \neq q_2 \wedge human(x) \wedge legs(y) \quad (15)$$

For another example, the L_{md} expression (16) is for the robotic motion of head shaking in default, reading that a robot affects its head in the 'Orientation (A_{14})', making its shape change monotonically from $Shake\_head_S$ to $Shake\_head_E$. The shape values are given in a computable form general enough to reconstruct any human motion in 3D graphics or so. Figure 4 shows an example of its interpretation in 3D graphics by our intelligent system IMAGES-M, which is also an example of cross-media translation from the text 'The robot shakes its head' into the animation.

$$(\exists x,y,p_1,p_2)L(\_,y,x,x,A_{01},G_t,\_)\Pi L(x,y,p_1,p_2,A_{14},G_t,\_)\Pi\, L(x,x,Shake\_head_S,Shake\_head_E,A_{11},G_t,F_R) \wedge robot(x) \wedge head(y) \quad (16)$$

The attributes extracted from natural language words are essentially for human sensors or actuators and therefore the locus formula as human motion should be translated into its equivalent concerning the attributes specific to the robot's organs.

For example, an atomic locus of the robot's 'Shape (A_{11})' specified by the human should be paraphrased into a set of atomic loci of the 'Angularity (A_{45})' of each joint in the robot. For another example, 'Velocity (A_{16})' for the human into a set of change rates in 'Angularity (A_{45})' over 'Duration (A_{35})' (i.e., A_{45}/A_{35}) of the robot's joints involved. These knowledge pieces are called 'Attribute Paraphrasing Rules (APRs)' as already mentioned above.

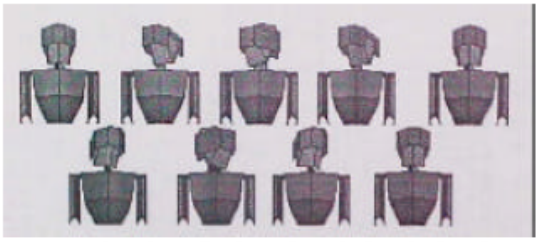

Fig. 5. 3D animation of 'The robot shakes its head.'

VI. CONCLUSION

We have proposed a methodology for systematic cross-media operations of human motion based on L_{md}, and implemented it on the intelligent system IMAGES-M and confirmed its validity for about 40 verb concepts such as 'raise' and 'nod'.

Our work's most remarkable advance to the others resides in the transparency of the description of word meanings and the processing algorithms due to the formal language L_{md}. In turn, this feature results in higher modularity of the program and higher order processing of human motion, for example, inference based on the knowledge formalized in L_{md}.

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] Mase K(1991), Recognition of facial expression from optical flow. IEICE Transactions, Vol. E 74-10: 3474-3483
[2] Moeslund TB, Granum E (2001), A Survey of Computer Vision-Based Human Motion Capture. *Computer Vision and Image Understanding: CVIU*, 81-3: 231–268
[3] Yacoob Y, Davis LS (1993), Labeling of human face components from range data. IEEE CVPR, 592-593
[4] Ren H, Xu G (2002), Human Action Recognition in Smart Classroom. Proc. of Int. Conf. on Automatic Face and Gesture Recognition, 417–422
[5] Sidenbladh H, Black M, Sigal L (2002), Implicit Probabilistic Models of Human Motion for Synthesis and Tracking. Proc. of European Conf. on Computer Vision, LNCS2350:784–800, Springer Verlag
[6] Sullivan J, Carlsson S (2002), Recognizing and Tracking Human Action. Proc. of European Conf. on Computer Vision, LNCS2350: 629–644, Springer Verlag
[7] Yokota M (2008), Intuitive spatiotemporal representation based on Mental Image Description Language *Lmd*. The Thirteenth International Symposium on Artificial Life and Robotics 2008(AROB 13th '08), Beppu, Oita, Japan
[8] Yokota M (2006), 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: 1124-1133
[9] Yokota M, Capi G (2005), Cross-media Operations between Text and Picture Based on Mental Image Directed Semantic Theory. WSEAS Transactions on Information Science and Applications, 10-2: 1541-1550
[10] Yokota M (2005), An approach to integrated spatial language understanding based on Mental Image Directed Semantic Theory. Proc. of 5th Workshop on Language and Space, Bremen, Germany