Frequent Motion Pattern Extraction for Motion Recognition in Real-time Human Proxy

Arita, Daisaku
Department of Intelligent Systems, Kyushu University

Yoshimatsu, Hisato
Department of Intelligent Systems, Kyushu University

Taniguchi, Rin-ichiro
Department of Intelligent Systems, Kyushu University

出版情報：Proceedings of JSAI Workshop on Conversational Informatics, pp.25-30, 2005-06. 人工知能学会
バージョン：
権利関係：

Frequent Motion Pattern Extraction for Motion Recognition in Real-time Human Proxy

Daisaku Arita, Hisato Yoshimatsu, and Rin-ichiro Taniguchi

Department of Intelligent Systems, Kyushu University
6-1, Kasuga-koen, Kasuga, Fukuoka, 816-8580, JAPAN
{arita,hisato,rin}@limu.is.kyushu-u.ac.jp

Abstract. We have proposed a concept of Real-time Human Proxy (RHP), which acquires, symbolizes, transfers and represents human information for avatar-based communication. For applying RHP to avatar-based communication, it is necessary for system developers to enumerate human actions to be acquired, symbolized and represented. In this paper, we propose a new method to extract frequently occurring human motion patterns from human motion acquired by an MCS for supporting system developers' enumeration. The method employs the idea of motif which is applied to one-dimensional data. We extend the idea to multi-dimensional data, or human motion information. Also we present experimental results.

1 Introduction

There are several researches on virtual environments for distant interaction. In these researches, a 3-D virtual space is reconstructed, in which each participant is represented as an avatar by computer graphics techniques. Through the reconstructed virtual space, each participant can see and hear other participants' activities from the position where his/her avatar is represented. Therefore, it is called avatar-based interaction.

In avatar-based interaction, an avatar is expected to reflect activities of a participant into a virtual space as if he/she were there. Nevertheless, legacy input devices, such as a keyboard and a mouse, are not sufficient to acquire participant's activities in aspects of quality and quantity. Using such devices, a participant has to intentionally keep feeding their own activities into a system by hand, and acquired information may be neither precise nor rich. To solve this problem, as an input device, we have developed a vision-based motion capture system (MCS)[1]. Using the MCS, we can acquire rich information of participants without compelling them to do annoying operations.

Though motion information acquired by an MCS is very suitable for controlling an avatar behaving just like a participant, it requires avatar's physique is just same as a participant. On the other hand, there is no such problem if an avatar is controlled according to action symbols, i.e. a system recognizes participant's human action, a series of human motions with one meaning, from motion information to get the labels of human actions, called action symbols, and generates motion information suited to avatar's physique from action symbols.

According to the above idea, we have proposed a concept of Real-time Human Proxy (RHP), which acquires, symbolizes, transfers and represents human information for avatar-based interaction[2]. As the first step of RHP, we focus on nonverbal, or body movement information of humans. For applying RHP to avatar-based interaction, it is necessary for system developers to choose which human actions should be acquired, symbolized and represented. In this paper, we propose a new method to extract frequently occurring human motion patterns from human motion acquired by an MCS for supporting system developers' choice.

2 Real-time Human Proxy

Real-time Human Proxy (RHP) is a new concept for avatar-based interaction, which makes avatars act more meaningfully, or expressively, referring to action information acquired by an MCS. As the first step, we currently focus on acquisition and representation of human action, or nonverbal information. In the acquisition process, we symbolize the human action information under a given communication environment, such as classroom. In the presentation process, the symbols acquired are presented, or visualized, which are augmented based on the knowledge of the environment. Figure 1 shows a concept of RHP.

Motion capture We have developed an MCS which uses eight or nine cameras and a PC-cluster[1]. The MCS can acquire 3-D positions of human parts (a head, hands, feet, elbows, knees and body) and 3-D pose of a head in real-time.

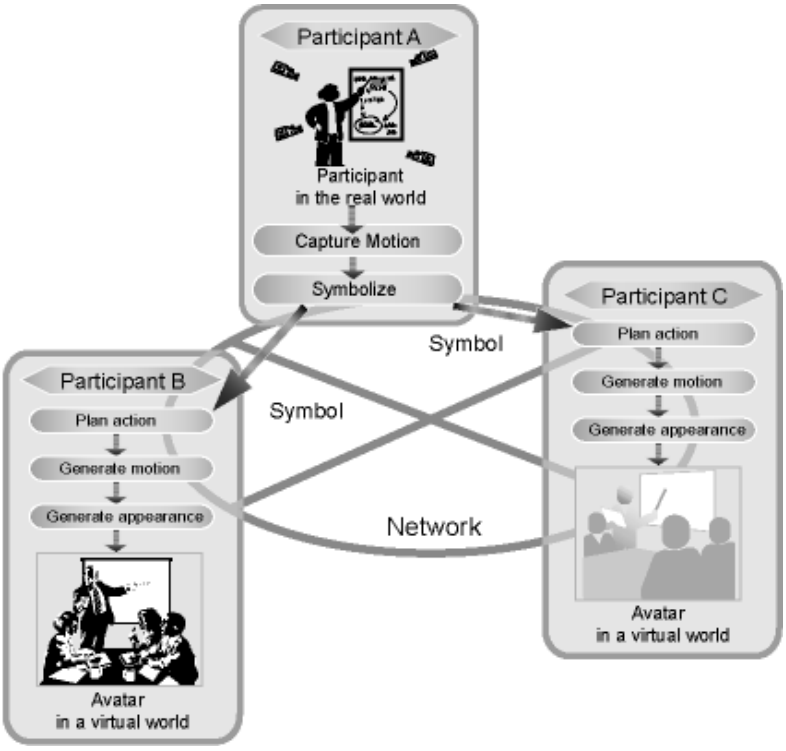


Fig. 1. The concept of RHP.

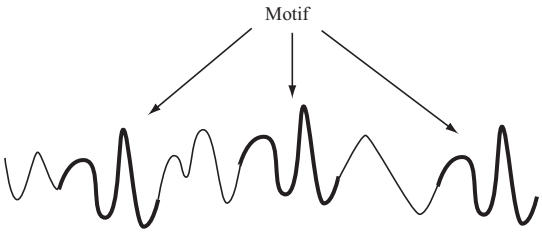


Fig. 2. Motif

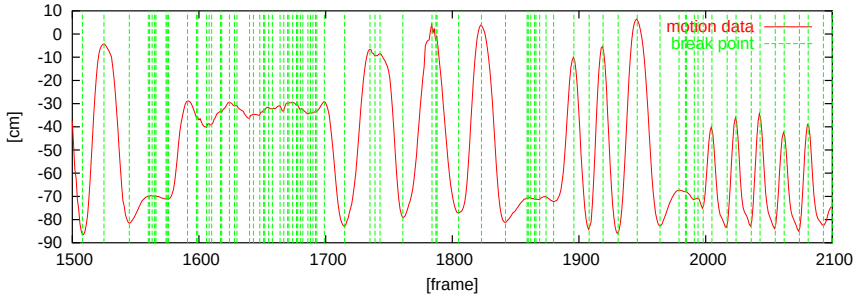
Symbolization On RHP, we acquire human actions instead of human motions. We categorize motion sequences into pre-defined actions, expressing them as symbols. Each symbol is formed by a label of an action and its parameters, such as "walking (p_x, p_y, ν_x, ν_y)" where p_x and p_y are the position, ν_x and ν_y are the velocity of a participant. After recognizing human actions from captured motion data, the system transfers the symbols to the representation side of a virtual space.

Avatar with pre-defined knowledge We define that an avatar is an object which is participant's substitute in a virtual space. An avatar has pre-defined knowledge to generate its motion and appearance from symbols[3].

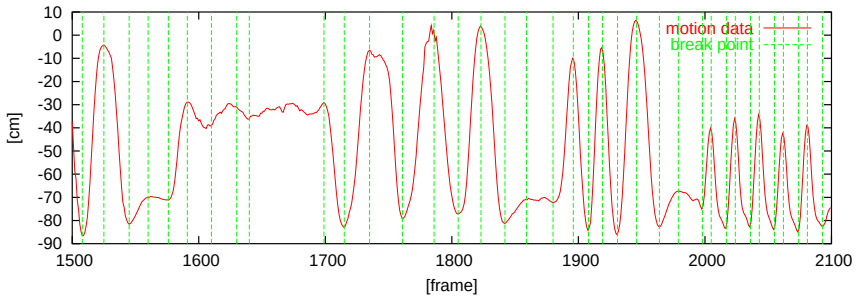
Representation of Virtual Space Generated appearance is represented in a virtual space. A participant is able to see the virtual space in which any participants, including him/herself, are represented as avatars.

3 Frequently Occurring Human Motion Pattern Extraction

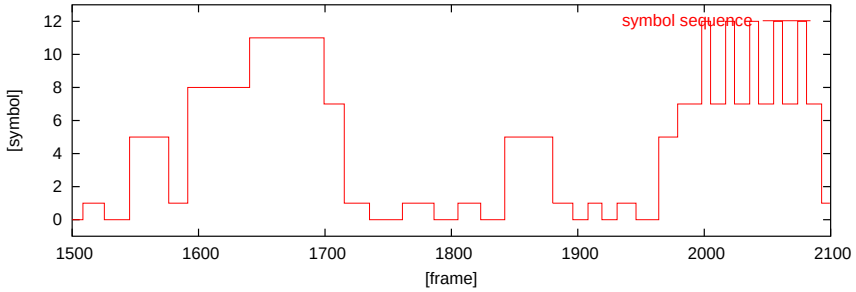
For RHP, it is necessary for system developers to enumerate human actions in advance to be acquired, symbolized and represented according to the situation of interaction. For example, actions to be enumerated for distant learning are "pointing", "hand raising", "nodding", "head shaking", "face touching (usually means anxiety and strain)", "arm folding (sometimes means caution and doubt)", and so on. This enumeration is very important to realize natural interaction. However, it is difficult to enumerate all actions necessary for the situation.



(a) Segmentation at maxima and minima.



(b) Motion units.



(c) Clustering results. The vertical axis means symbols.

Fig. 3. Segmentation and clustering of height of a right hand.

In this paper, we propose a new method to extract frequently occurring human motion patterns, called “motifs” defined in [4] as frequently occurring patterns in some time sequences shown in Fig. 2, from human motion sequences acquired by an MCS. Our method makes it easier for system developers to enumerate human actions since they can choose human actions from extracted motifs.

First, we will explain a method to extract motifs from one-dimensional human motion sequences. Secondly we explain how to extend the method to multi-dimensional human motion sequences. Lastly, we explain human motion sequences used for our system.

3.1 Extracting Motifs from One-dimensional Human Motion Sequences

Motifs are extracted from one-dimensional human motion sequences as follows.

Segmentation of human motion sequences Human motion sequences are segmented into motion units, which are the minimum elements of motion sequences, at maxima and minima of motion sequences. This is because

maxima and minima, which are turning points of human motion direction, should be segmentation points. Fig. 3 (a) shows segmentation points of a motion sequence (height of a right hand). However, regarding all maxima and minima as segmentation points, noises of motion capture and small trembles of motion causes over-segmentation. Then, it is necessary to eliminate meaningless segmentation points, whose times are too close to neighbors or whose values are less different from neighbors. Fig. 3 (b) shows motion units whose segmentation points are thinned out.

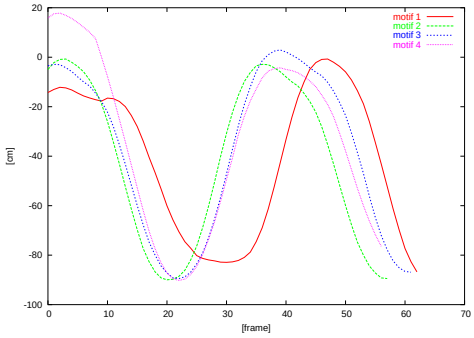


Fig. 4. Extracted motif.

Clustering of motion units After segmentation, motion units are clustered. Similar motion units are clustered a same group which is represented by a symbol. This clustering converts a human motion sequence to a symbol sequence. Clustering of similar motion units is realized by the Nearest Neighbor method. Similarity between motion units is measured by the Dynamic Programming technique. Fig. 3 (c) shows a symbol sequence.

Extraction of frequently occurring patterns Frequently occurring symbol patterns in symbol sequences are extracted as motifs. Each Motif is extracted to maximize its length, or the number of symbols in it, as much as possible. Fig. 4 shows a sample of motifs whose length is four. Though each motion sequence is individually expanded along the horizontal time axis, the difference is overcome by the Dynamic Programming technique.

3.2 Extracting Motifs from Multi-dimensional Human Motion Sequences

A human motion sequence, described in the next section, is a multi-dimensional sequence, three-dimensional position sequences of multiple human joints. Then, we have to extend the motif extraction method described above to multi-dimensional data. First, Segmentation and clustering is applied to each human motion sequence. Since segmentation points of human motion sequences are different from each other, symbol sequences are not synchronized (See Fig. 5 (a)). Secondly, multiple symbol sequences are synchronized by re-segmenting at points where at least one symbol sequence is segmented (See Fig.5 (b)). Thirdly, a new symbol sequence is generated by combining multiple symbol sequences and relabeling (See Fig.5 (c)). Symbols whose temporal lengths are too short are eliminated to avoid extracting useless motifs.

S_1	A	B	C	A	B
S_2	D	E	F	E	

(a) unsynchronized symbol sequences

S_1	A	A	B	C	C	A	A	B
S_2	D	E	E	E	F	F	E	E

(b) synchronized symbol sequences

S_3	a	b	c	d	e	f	b	c

(c) Combined symbol sequence

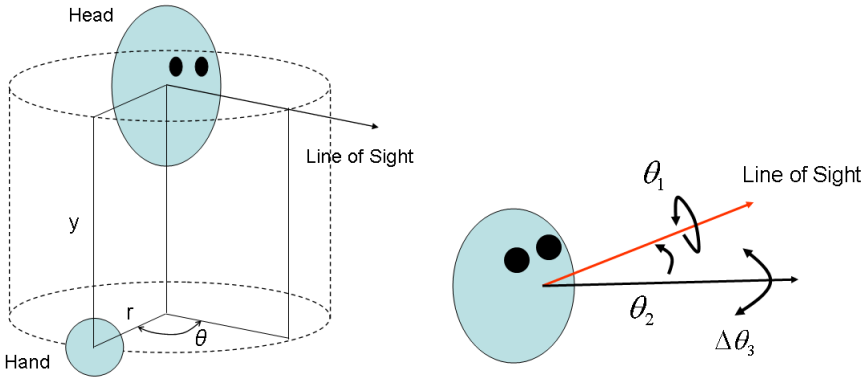
Fig. 5. Synchronization with multiple symbol sequences.

3.3 Multi-dimensional human motion sequence

Our MCS[1] can capture the three-dimensional positions of the whole body of a human. At the first step, however, we deal with the upper half of the human body (head, elbows and hands), i.e. 15-dimensional human motion sequence. Captured positions are expressed in three-dimensional rectangular coordinates. This means that captured positions are changed depending on the human direction though he/she makes a same motion. Then, we employ human-head-centered cylindrical coordinates like Fig. 6 (a). For a head, we employ such coordinates as Fig. 6 (b).

4 Experiments

In this section, we show experimental results of motif extraction described above. Input human motion sequences are about 40000 frames (12 fps). The sequences include such actions as shown in Table 1. Some of these actions



(a) Cylindrical coordinates (r, θ, y) for hand. The origin is at the head center. r and y are horizontal and vertical distances from the origin respectively. θ is rotation angle from the line of sight.

(b) Coordinates $(\theta_1, \theta_2, \theta_3)$ for head. The origin is the head center. θ_1, θ_2 and θ_3 are roll, yaw and pitch angles respectively. Instead of θ_3, $\Delta\theta_3$ is used as motion information.

Fig. 6. Coordinates

pointing by a right hand	rotating a right arm	raising a right hand
pointing by a left hand	clapping hands	raising a left hand
knocking by a left hand	sticking both hands ahead	raising both hands
waving a right hand beside a head	rotating shoulders and arms	nodding a head
waving both hands beside a head		shaking a head

Table 1. Human actions.

consists of outward motion, middle motion, and homeward motion. For example, “raising a hand” consists of raising a hand, keeping a hand up, and lowering a hand. Then, such actions should be extracted as two motifs corresponding to an outward motion and a homeward motion. A middle motion can not be extracted since it has no motion pattern.

The experimental results shows that all actions except head actions are extracted as motifs. The reason why head actions are not extracted may be that a head often stands still and there are noises of our MCS, then miss-segmentation of head motion sequences occurs. Fig.7 shows samples of extracted motifs. However, since our current system extracts a huge amount of motifs almost of which are useless, system developers still have to make a very hard job to choose human motions out of motifs. Then, we are planning to decrease the number of extracted motifs.

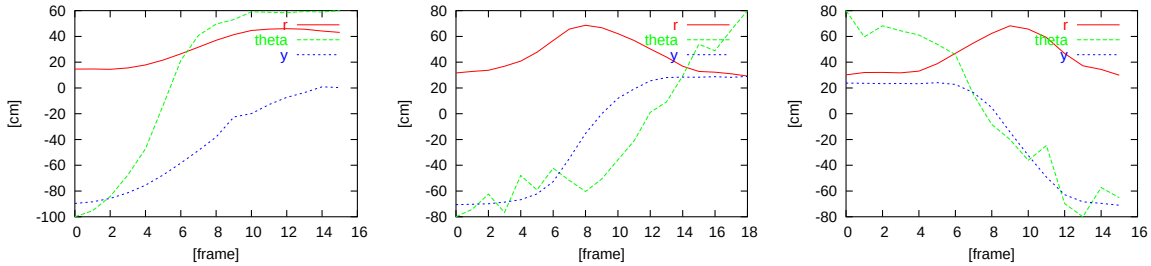
Fig.8 shows snapshots of interaction using RHP. Each actions are recognized by an HMM recognizer, which learns extracted motifs. According to recognition results, his avatar, a standing avatar in the display, makes actions.

5 Conclusion

In this paper, we propose a motif extraction method from human motion sequences for Real-time Human Proxy. Motif extraction makes it easier for system developers to enumerate human actions without omission since they can choose human actions from extracted motifs.

References

1. Date, N., Yoshimoto, H., Arita, D., Taniguchi, R.: Real-time human motion sensing based on vision-based inverse kinematics for interactive applications. In: Proc. of International Conference on Pattern Recognition. Volume 3. (2004) 318–321
2. Arita, D., Yoshimatsu, H., Hayama, D., Kunita, M., Taniguchi, R.: Real-time human proxy: An avatar-based interaction system. In: CD-ROM Proc. of International Conference on Multimedia and Expo. (2004)

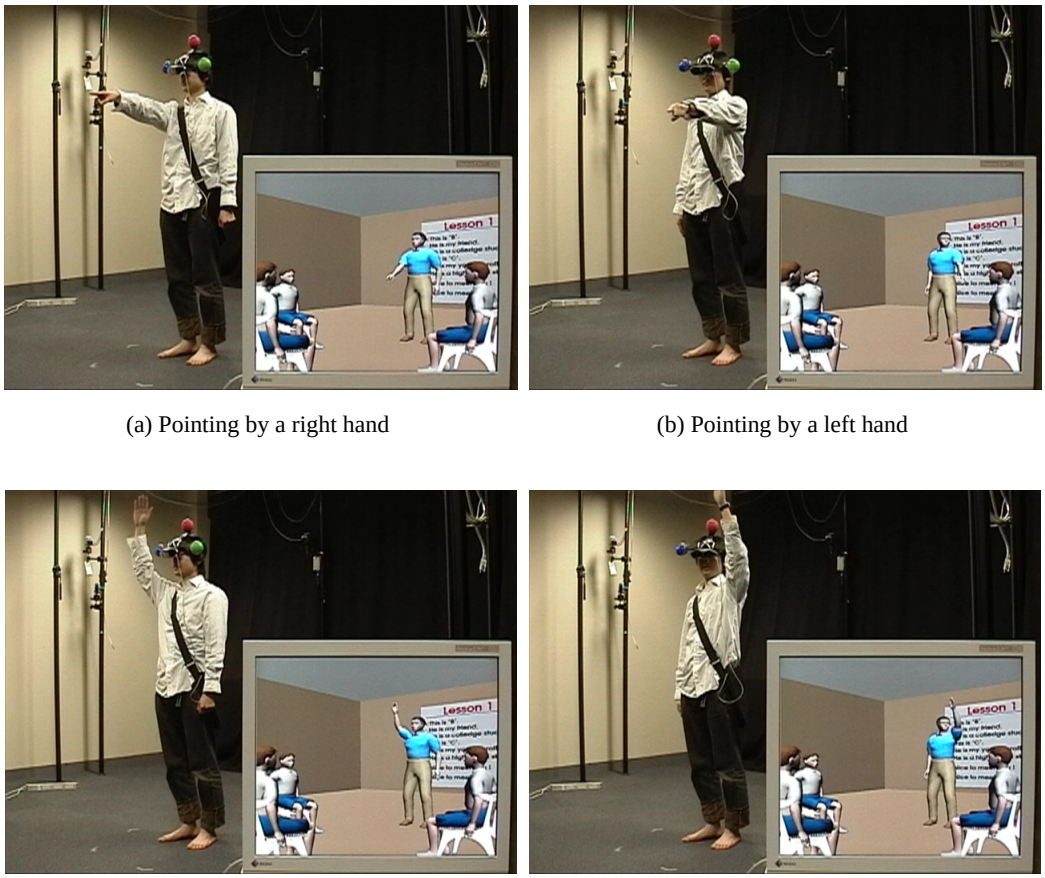


(a) Pointing by a right hand(outward)

(b) Raising a right hand (outward)

(c) Raising a right hand (homeward)

Fig. 7. Extracted Motifs.



(a) Pointing by a right hand

(b) Pointing by a left hand

(c) Raising a right hand

(d) Raising a left hand

Fig. 8. Snapshots of RHP.

3. Hayama, D., Yoshimatsu, H., Yoshimoto, H., Arita, D., Taniguchi, R.: Avatar generation for real-time human proxy. In: Proc. of 10th International Conference on Virtual Systems and Multimedia. (2004) 386–395
4. Lin, J., Keogh, E., Lonardi, S., Patel, P.: Finding motifs in time series. In: Proc. of the 2nd Workshop on Temporal Data Mining. (2002) 53–68