Electronic Preprint for Journal of Information Processing Vol.21 No.2

Technical Note

Indexing of Motion Capture Data Using Feature Vectors Derived from Posture Variation

Takeshi Miura[1,a)] Naho Matsumoto[2] Takaaki Kaiga[1,3] Hiroaki Katsura[2]
Katsubumi Tajima[1] Hideo Tamamoto[1]

Received: October 22, 2012, Accepted: January 11, 2013

Abstract: Recently several large-scale databases of motion-capture data streams have been constructed. We present a novel method to index motion-capture data streams in such databases. We pay attention to posture variation; the impression of the visual aspect of the whole body is regarded as important. The spatial distribution of body segments is statistically summarized as a feature vector having only 12 dimensions. The experimental results showed that the feature vector we introduced provided properties comparable to those of the methods previously proposed, even though its dimensionality is extremely low.

Keywords: motion capture, information retrieval, indexing, similarity

1. Introduction

Recently motion-capture (Mocap) data streams have attracted much attention due to their high reproducibility for human motions. Several large-scale databases of Mocap data streams have been constructed in the past few years [1], [2]. Utilizing Mocap databases allows us to easily create realistic computer animations of human-like characters.

The use of a fast information retrieval system is required for database management. Indexing documents in a database is known as one of the methods to realize fast information retrieval. In this paper, we present a novel method to index Mocap data streams; the developed method is used for similarity retrieval. We pay attention to posture variation; the impression of the visual aspect of the whole body is regarded as important. In the first stage of indexing, the spatial distribution of body segments is quantified at every frame by statistically analyzing the positions of body segments. Then, the tendency of all the frames in a data stream is statistically summarized as a 12-dimensional feature vector; this vector corresponds to a document vector in information retrieval.

To evaluate the developed method, we conducted an experiment in which a set of Mocap data streams selected from multiple motion categories was used. The experimental results showed that the feature vector we introduced provided properties comparable to those of the methods previously proposed, even though its dimensionality is extremely low.

The remainder of this paper is organized as follows. We first review the related work in Section 2. In Section 3, we describe the derivation of the feature vector. We verify the effectiveness of the developed method in Section 4. Conclusions are finally summarized in Section 5.

2. Related Work

It is well known that dynamic time warping (DTW) is often used to evaluate the similarity between Mocap data streams [3]. DTW directly compares every pair of frames each extracted from each of the data streams compared. This causes a significant disadvantage in computational complexity, namely quadratic time complexity [4].

Krüger et al. [5] reported a trial to reduce the time complexity of the frame-comparison approach. However, the proposed procedure requires a large amount of space to store the data of the frames similar to each of all the frames in a query motion. As for the present method, in contrast, only 12 data are stored as those representing each Mocap data stream.

A number of researchers have proposed several indexing methods utilizing some sort of features of Mocap data streams. Onuma et al. [6] developed FMDistance in which the feature vector representing the kinetic energy of joint motions was used. Li et al. [7] employed singular value decomposition (SVD) to extract the geometric structure of a Mocap-data matrix. The feature vectors of the above methods are much longer than that of the present method, as will be shown later.

Preprocessing of a database has also been examined: clustering [8], preparing binary geometric features [9], extracting hierarchically-structured motion patterns [10], etc. These approaches require a relatively large number of procedures such as updating newly added data streams [8], manually selecting motion features [9], spatially and temporally segmenting motion sequences [10], etc. On the other hand, the present method does not require the preprocessing of an entire database; only indexing individual Mocap data streams is needed.

1 Graduate School of Engineering and Resource Science, Akita University, Akita 010–8502, Japan
2 Faculty of Education and Human Studies, Akita University, Akita 010–8502, Japan
3 Digital Art Factory, Warabi-za Co., Ltd., Semboku, Akita 014–1192, Japan
a) miura@ipc.akita-u.ac.jp

3. Derivation of the Feature Vector

As mentioned in Section 1, we focus on the impression of the visual aspect of the whole body. Here, we index Mocap data streams under the assumption that the impression depends on the spatial distribution of body segments.

Consider the constellation of the joints and end effectors shown in **Fig. 1**: shoulders, elbows, wrists, fingers, knees, ankles, toes, neck and head. End effectors are hereafter regarded as joints for simplicity. The position of each joint is described in the coordinate system fixed to the pelvis, and normalized by the height of the body to reduce the influence of difference in body constitution.

We first quantify the distribution of body segments at each frame using the variance-covariance matrix of joint coordinates:

$$\boldsymbol{\Sigma}(n) = \begin{bmatrix} \sigma_{xx}(n) & \sigma_{xy}(n) & \sigma_{xz}(n) \\ \sigma_{yx}(n) & \sigma_{yy}(n) & \sigma_{yz}(n) \\ \sigma_{zx}(n) & \sigma_{zy}(n) & \sigma_{zz}(n) \end{bmatrix} \tag{1}$$

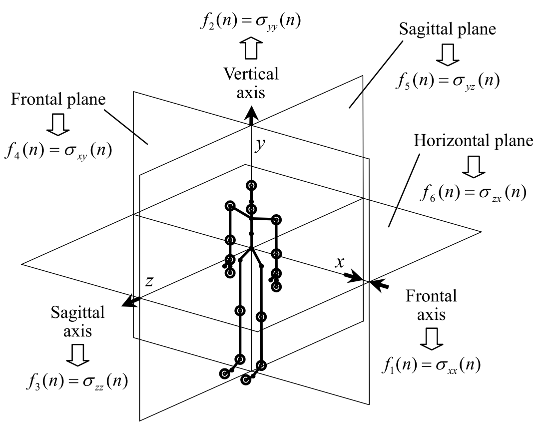
$$\sigma_{ab}(n) = \frac{1}{J}\sum_{j=1}^{J}\left\{p_{aj}(n) - \bar{p}_a(n)\right\}\left\{p_{bj}(n) - \bar{p}_b(n)\right\},$$

$$\bar{p}_a(n) = \frac{1}{J}\sum_{j=1}^{J} p_{aj}(n) \qquad (a, b : x, y \text{ or } z)$$

where $p_{aj}(n)$ is the a-coordinate of the jth joint at the nth frame and J is the number of joints selected ($J = 16$), respectively. Since $\boldsymbol{\Sigma}(n)$ is symmetric, only six elements are needed to describe the distribution of body segments; we select the elements corresponding to the axes and planes of movement [11] (see Fig. 1). We adopt these elements as the components of the feature vector $\boldsymbol{f}(n)$ characterizing a posture in each frame:

$$\begin{aligned} \boldsymbol{f}(n) &= \begin{bmatrix} f_1(n) & f_2(n) & f_3(n) & f_4(n) & f_5(n) & f_6(n) \end{bmatrix}^T \\ &= \begin{bmatrix} \sigma_{xx}(n) & \sigma_{yy}(n) & \sigma_{zz}(n) \\ & \sigma_{xy}(n) & \sigma_{yz}(n) & \sigma_{zx}(n) \end{bmatrix}^T \end{aligned} \tag{2}$$

To estimate the tendency throughout an entire data stream, we statistically summarize the feature vectors obtained from all the frames in the data stream as follows:



◎ : Joints and end effectors used to quantify the distribution of body segments

Fig. 1 Quantification of the spatial distribution of body segments.

$$\boldsymbol{F} = \begin{bmatrix} F_1 & F_2 & \cdots & F_{12} \end{bmatrix}^T = \begin{bmatrix} \bar{\boldsymbol{f}} \\ \bar{\boldsymbol{s}} \end{bmatrix} \tag{3}$$

$$\bar{\boldsymbol{f}} = \begin{bmatrix} \bar{f}_1 & \bar{f}_2 & \cdots & \bar{f}_6 \end{bmatrix}^T, \quad \bar{f}_i = \frac{1}{N}\sum_{n=1}^{N} f_i(n),$$

$$\bar{\boldsymbol{s}} = \begin{bmatrix} \bar{s}_1 & \bar{s}_2 & \cdots & \bar{s}_6 \end{bmatrix}^T, \quad \bar{s}_i = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\left\{f_i(n) - \bar{f}_i\right\}^2}$$

where N is the number of frames and $\bar{\boldsymbol{f}}$ and $\bar{\boldsymbol{s}}$ are the mean and standard deviation of $\boldsymbol{f}(n)$'s, respectively. We finally employ the 12-dimensional vector $\boldsymbol{F}$ as the feature vector representing a Mocap data stream. Since every component of $\boldsymbol{F}$ has the identical physical dimension (squared length), similarity between Mocap data streams can be evaluated by the Euclidean distance between $\boldsymbol{F}$'s.

The calculation of Eq. (1) for a Mocap data stream requires $9JN$ computations, and that of Eq. (3) requires DN computations ($D = 12$, invariant with respect to J which can be changed as the need arises, and in general $D \ll 9J$). As a result, the computational complexity of calculating $\boldsymbol{F}$ becomes $O(JN)$. As for the Euclidean distance between $\boldsymbol{F}$'s, $O(D)$ is required.

4. Experimental Results

We report the experimental results in this section. The Mocap data streams used in the experiment are shown in **Table 1** (138 data streams classified into 17 categories, downloaded from Carnegie-Mellon Mocap Database [1]). We compared the present method with FMDistance [6], kWAS [7] and PCA similarity factor [12]. These methods have the following properties in common with the present method:

(1) A Mocap data stream is represented as a feature vector having a fixed length.

(2) Preprocessing of a database is not required.

Table 2 shows the dimensionality of the feature vectors and computational complexity [*1]. It is noted that the dimensionality

Table 1 Motion-capture data streams used in the experiment.

Label	Category	Data	Number of data
A	Walk	07_01–07_03, 07_06–07_11	9
B	Walk (slow)	07_04, 07_05, 08_04, 37_01	4
C	Walk (on uneven terrain)	36_10–36_20	11
D	Marching	138_01–138_10	10
E	Run	09_01–09_09	9
F	Jump	118_01–118_10	10
G	Climb ladder	13_33, 13_34, 14_33–14_35	5
H	Golf (swing)	64_01–64_10	10
I	Soccer (kick ball)	10_01–10_03, 10_05, 10_06, 11_01	6
J	Basketball (forward dribble)	06_02–06_05	4
K	Boxing	14_01–14_03, 15_13, 17_10	5
L	Modern dance	05_02–05_14	13
M	Chicken dance	18_15, 19_15, 20_01, 21_01, 143_34	5
N	Salsa dance	61_01–61_10	10
O	Breaking	85_01–85_08, 85_10	9
P	Charleston	93_03–93_06, 93_08	5
Q	Indian dance	94_01–94_13	13

Total: 138

Downloaded from http://mocap.cs.cmu.edu.

*1 As for PCA similarity factor, we set the number of DOF to be identical to that of kWAS, and the number of principal components to be six which is large enough to give almost all Mocap data streams the over-80-percent contribution rate.

Table 2 Dimensionality of feature vectors and computational complexity.

Method	Dimensionality of feature vector	Computational complexity	
		Calculation of feature vector	Calculation of distance
Present method	$D = 12$	$O(JN)$	$O(D)$ (Euclidean distance)
FMDistance	$D = 61$ $(= L)$	$O(LN)$	$O(D)$ (Euclidean distance)
kWAS	$D = 330$ $(= (L + 1)k,\ L = 54,\ k = 6)$	$O(L^2N)$ (for SVD)	$O(Lk)$ ((inner product of L-dimensional vectors)$\times k$)
PCA similarity factor	$D = 324$ $(= Lk,\ L = 54,\ k = 6)$	$O(L^2N)$ (for PCA)	$O(Lk^2)$ (product of $(k \times L)$ and $(L \times k)$ matrices)

N: Number of frames, L: DOF of Mocap data, J: Number of joints, k: Number of singular values (or principal components).

Table 3 Results of supervised classification and unsupervised clustering.

	Present method			FMDistance			kWAS			PCA similarity factor		
	Supervised	Unsupervised 18 clusters		Supervised	Unsupervised 17 clusters		Supervised	Unsupervised 16 clusters		Supervised	Unsupervised 18 clusters	
	Error	R_c	P_c	Error	R_c	P_c	Error	R_c	P_c	Error	R_c	P_c
A	1	1.000	0.750	1	1.000	0.692	0	1.000	0.129	1	1.000	0.692
B	2	0.625	0.210	3	1.000	0.308	2	1.000	0.057	3	1.000	0.308
C	0	1.000	1.000	0	1.000	0.917	0	1.000	0.157	0	1.000	0.500
D	0	1.000	1.000	0	1.000	1.000	0	1.000	0.143	0	1.000	0.455
E	0	1.000	1.000	0	1.000	1.000	0	1.000	0.129	0	1.000	1.000
F	0	0.500	0.857	0	1.000	0.625	0	1.000	0.143	0	1.000	1.000
G	0	0.520	0.221	1	0.680	0.233	0	1.000	0.192	0	1.000	1.000
H	0	1.000	1.000	0	1.000	1.000	0	1.000	1.000	0	1.000	1.000
I	1	0.722	0.391	0	1.000	0.600	0	1.000	0.086	0	1.000	0.667
J	0	1.000	0.286	1	1.000	0.400	0	1.000	0.057	1	0.625	0.778
K	0	1.000	0.357	1	0.680	0.567	0	1.000	0.192	0	1.000	0.833
L	0	0.456	0.890	1	0.361	0.676	5	0.219	1.000	4	0.609	0.859
M	0	1.000	1.000	1	0.680	0.552	0	1.000	0.192	1	1.000	1.000
N	0	1.000	0.909	0	0.580	0.925	0	1.000	0.385	0	1.000	0.909
O	0	0.407	1.000	2	0.481	0.602	3	0.160	0.677	2	0.481	0.804
P	3	0.360	0.148	3	0.360	0.511	3	1.000	0.071	3	0.360	0.498
Q	0	0.858	1.000	2	0.609	0.719	1	1.000	1.000	0	0.858	0.868
Total	$A_e = 0.949$	0.797	0.804	$A_e = 0.884$	0.781	0.725	$A_e = 0.899$	0.872	0.413	$A_e = 0.891$	0.882	0.790
		$F_{measure} = 0.800$			$F_{measure} = 0.752$			$F_{measure} = 0.561$			$F_{measure} = 0.834$	

of the present method is extremely low compared with the other methods. As for computational complexity, all the methods take linear time with respect to N to calculate a feature vector; on the other hand, the complexity of calculating the distance between Mocap data streams does not depend on N in all the cases.

We verified the effectiveness of the above methods in two stages: supervised classification and unsupervised clustering. In supervised classification, we evaluated the results given by the 1-nearest-neighbor classifier [13] using the empirical accuracy A_e obtained from 1-fold cross-validation [13]. As for unsupervised clustering, we used the hierarchical clustering algorithms [14]; Ward's method was applied to both the present method and FMDistance which use Euclidean distance, whereas the group average method was applied to both kWAS and PCA similarity factor which use nonmetric similarity measures. The number of clusters was determined by maximizing the Bayesian information criterion for the Gaussian mixture clustering model [14] (present method and FMDistance) or by Mojena's stopping rule [15] (kWAS and PCA similarity factor). To evaluate the results of clustering, we used the parameter $F_{measure}$ [16].

$F_{measure}$ is given as a combination of the parameters recall R and precision P as follows:

$$F_{measure} = \frac{2RP}{R + P} \tag{4}$$

$$R = \frac{\sum_{c=1}^{C} M_c R_c}{M}, \quad P = \frac{\sum_{c=1}^{C} M_c P_c}{M},$$

$$R_c = \frac{\sum_{q=1}^{Q} M_{q,c} R_{q,c}}{M_c}, \quad P_c = \frac{\sum_{q=1}^{Q} M_{q,c} P_{q,c}}{M_c},$$

$$R_{q,c} = \frac{M_{q,c}}{M_c}, \quad P_{q,c} = \frac{M_{q,c}}{\sum_{c'=1}^{C} M_{q,c'}}$$

where M is the total number of samples (feature vectors in this case), M_c is the number of samples in the cth ground-truth category, $M_{q,c}$ is the number of samples of the cth ground-truth category in the qth cluster, C is the total number of the ground-truth categories and Q is the total number of clusters, respectively.

The experimental results are shown in **Table 3**. In supervised classification, the present method gave the highest value of A_e. As for unsupervised clustering, PCA similarity factor gave the highest value of $F_{measure}$; however, the difference of the present method from PCA similarity factor is only 0.034. Although the dimensionality of the feature vector we introduced is extremely low, the present method provided properties comparable to those of the other methods. This suggests that the efficiency of the present method is considerably high.

It should also be pointed out, on the other hand, that the present method has several limitations; a typical one is that whole-body locomotion is not considered. This may have caused the confusion between the categories "Walk (slow)" (B) and "Climb ladder" (G) in unsupervised clustering. The factor that motion speed is not incorporated is also noted; this may have caused the confusion between the categories "Walk" (A) and "Walk (slow)" (B).

5. Conclusions

The main contribution of this study is the dimensionality reduction of the feature vector used for similarity retrieval in Mocap databases; this was accomplished without significant performance degradation. It is hoped that the present method will help in improving Mocap-database management systems. However, the issue that several motion characteristics such as whole-body locomotion and motion speed are not incorporated still remains unresolved. Further work is necessary to resolve this issue.

Acknowledgments This study was supported by the “Adaptable and Seamless Technology Transfer Program through Target-driven R&D (A-STEP)” of Japan Science and Technology Agency.

References

[1] Carnegie-Mellon Mocap Database, available from ⟨http://mocap.cs.cmu.edu⟩.

[2] Animazoo Motion Capture Systems and Technology, available from ⟨http://www.animazoo.com/⟩.

[3] Hachimura, K.: Digital Archiving of Dancing, *Review of the National Center for Digitization*, Vol.8, pp.51–66 (2006).

[4] Strle, B., Možina, M. and Bratko, I.: Qualitative Approximation to Dynamic Time Warping Similarity between Time Series Data, *Proc. QR2009* (2009).

[5] Krüger, B., Tautges, J., Weber, A. and Zinke, A.: Fast Local and Global Similarity Searches in Large Motion Capture Databases, *Proc. SCA10*, pp.1–10 (2010).

[6] Onuma, K., Faloutsos, C. and Hodgins, J.K.: FMDistance: A Fast and Effective Distance Function for Motion Capture Data, *Proc. EUROGRAPHICS 2008*, Short Papers (2008).

[7] Li, C. and Prabhakaran, B.: A Similarity Measure for Motion Stream Segmentation and Recognition, *Proc. MDM/KDD 2005*, pp.89–94 (2005).

[8] Pradhan, N.G., Li, C. and Prabhakaran, B.: Hierarchical Indexing Structure for 3D Human Motions, *Proc. MMM 2007*, Part I, pp.386–396 (2007).

[9] Müller, M., Röder, T. and Clausen, M.: Efficient Content-Based Retrieval of Motion Capture Data, *ACM Trans. Graphics* (*Proc. ACM SIGGRAPH 2005*), Vol.24, No.3, pp.677–685 (2005).

[10] Deng, Z., Gu, Q. and Li, Q.: Perceptually Consistent Example-Based Human Motion Retrieval, *Proc. I3D 2009*, pp.191–198 (2009).

[11] Bartlett, R.: *Introduction to Sports Biomechanics*, 2nd Edition, Routledge (2008).

[12] Krzanowski, W.J.: Between-Groups Comparison of Principal Components, *Journal of the American Statistical Association*, Vol.74, No.367, pp.703–707 (1979).

[13] Mullin, M. and Sukthankar, R.: Complete Cross-Validation for Nearest Neighbor Classifiers, *Proc. ICML 2000*, pp.639–646 (2000).

[14] Gan, G., Ma, C. and Wu, J.: *Data Clustering: Theory, Algorithms, and Applications*, SIAM (2007).

[15] Mojena, R.: Hierarchical Grouping Methods and Stopping Rules: An Evaluation, *The Computer Journal*, Vol.20, No.4, pp.359–363 (1977).

[16] Tan, P.N., Steinbach, M. and Kumar, V.: *Introduction to Data Mining*, Addison Wisley (2005).

Takeshi Miura received his D.Eng. degree in electrical engineering from Hokkaido University in 1998. He is currently an associate professor in the Department of Electrical and Electronic Engineering, Graduate School of Engineering and Resource Science, Akita University.

Naho Matsumoto received her M.A. of sport science degree in sport pedagogy from Tsukuba University in 2003. She is currently an associate professor in the Department of Education and Human Studies, Graduate School of Pedagogy Science, Akita University. Her research interests include dance education, curriculum and instruction in sport pedagogy.

Takaaki Kaiga received his M.E. degree in mechanical engineering from Ibaraki University in 1995. Since 1996, he has been with the computer division of Digital Art Factory, Warabi-za Co., Ltd.

Hiroaki Katsura is currently a professor in the Department of Music Education, Course of School Subject Teaching, Program of School Education, Faculty of Education and Human Studies, Akita University.

Katsubumi Tajima received his D.Eng. degree in electrical engineering from Tohoku University in 1998. He is a professor in the Cooperative Major in Life Cycle Design Engineering, Graduate School of Engineering and Resource Science, Akita University.

Hideo Tamamoto received his D.Eng. degree in electrical engineering from the University of Tokyo in 1976. He is currently a professor in the Cooperative Major in Life Cycle Design Engineering, Graduate School of Engineering and Resource Science, Akita University. His research interests include design-for-testability of logic circuits, archiving and handing-down technique for traditional folk dances, and e-learning system.