

# Human motion capture data retrieval based on semantic thumbnail

Xin Wang<sup>1</sup> · Liangxiu Chen<sup>1</sup> · Jiali Jing<sup>1</sup> ·  
Herong Zheng<sup>1</sup>

Received: 10 January 2015 / Revised: 4 May 2015 / Accepted: 20 May 2015 /  
Published online: 29 May 2015  
© Springer Science+Business Media New York 2015

**Abstract** We present a method for the efficient retrieval and browsing of immense amounts of realistic 3D human body motion capture data. The proposed method organizes motion capture data based on statistical K-means (SK-means), democratic decision making, unsupervised learning, and visual key frame extraction, thus achieving intuitive retrieval by browsing thumbnails of semantic key frames. We apply three steps for the efficient retrieval of motion capture data. The first is obtaining the basic type clusters by clustering motion capture data using the novel SK-means algorithm, and after which, immediately performing character matching. The second is learning the retrieval information of users during the retrieval process and updating the successful retrieval rate of each data; the search results are then ranked on the basis of successful retrieval rate by democratic decision making to improve accuracy. The last step is generating thumbnails with semantic generalization, which is conducted by using a novel key frame extraction algorithm based on visualized data analysis. The experiment demonstrates that this method can be utilised for the efficient organization and retrieval of enormous motion capture data.

**Keywords** Motion capture data · Visualized data analysis · Thumbnail · Retrieval

## 1 Introduction

Huge amounts of realistic 3D human motion capture data are produced by the popular use of motion capture devices. These data are extensively used in computer games, animation,

---

✉ Xin Wang  
xinw@zjut.edu.cn

Liangxiu Chen  
varko.cheng@yahoo.com

Jiali Jing  
muhuanguai@hotmail.com

Herong Zheng  
hailiang@zjut.edu.cn

<sup>1</sup> Zhejiang University of Technology, Hangzhou, China

medical simulation, and so on. However, motion capture data gathered at high frequency are characterized by several disadvantages, including the huge amount and high cost of data, vast redundant information, the lack of structured information, and the poor re-usability of motion capture data. Faced with these issues, developers must find ways of efficiently managing and reusing data. As such, the main research target in this field is to design an algorithm through which we can query and browse the database efficiently.

In recent years, studies have shown that the range of applications of motion capture data is still being expanded. Many industries, such as sports training, rehabilitation, and disease detection, are currently using these emerging multimedia data. Motion capture data allows for enhanced creativity in animation and film production. Although the study of motion capture data and the development of related applications are still at early stages today, multimedia data, just like image, video and voice, are expected to become widely popular in the future. In other words, everyone would be able to create with the use of their knowledge and creativity.

Both experts and amateurs of animation production cannot perform their tasks well without searching, browsing and accumulating materials. Thus, efficient, accurate and reliable methods are necessary for the retrieval and reuse of extensive motion capture data. Motion retrieval is an essential link between motion synthesis and mixing technology, and motion capture data are complex high-dimensional vectors that vary with time. At present, the similarity between two motions is difficult, but nonetheless vital, to compare with the use of popular motion retrieval techniques based on content. However, motions that are similar in value are not necessarily similar in logic, thereby suggesting that simple measurements are ineffective for motion capture data. Many researchers have proposed similarity measurements based on motion segment or posture. During early days of motion retrieval study, human motion capture data are mostly treated as continuous high-dimensional vectors. Jeff et al. [12] have presented an Adaptive Feature Selection method that abstracts the characteristics of the query by a Linear Regression Model, and different feature subsets can be selected according to the properties of the specific query. To support the real-time interface, a specialized encoding of the motions and the hand-drawn query is required. Naoki et al. [11] have presented a system that allows the user to retrieve a particular sequence by performing an approximation of the motion with an instrumented puppet. Chao et al. [3] introduce a hierarchical encoding scheme based on a set of orthonormal spherical harmonic basis functions.

In recent years, an increasing number of researchers tend to discretize motion capture data, and then incorporate discrete frameworks such as string matching and text retrieval to solve motion retrieval. In Zhang's approach [16], geometric features are first used to describe motion capture data; then, these geometric features are clustered based on K-means to obtain the motion vocabulary (cluster centres); and finally, the term frequency-inverse document frequency technology and the vector space model are used to retrieve the motion. Deng et al. [4] have proposed a content-based retrieval method. In this method, the human body is divided into pieces according to the thickness of the granularity; the human body pieces are clustered and analysed, and motions are represented by establishing a hierarchy; finally, the distance between the motions is calculated by using Knuth-Morris-Pratt algorithm and similarity passing. Wu et al. [13] considered human body to consist of three separate parts, namely, upper limbs, lower limbs, and trunk. They use self-organizing map and Smith-Waterman algorithm on the parts of the body, and then they use hierarchical clustering to obtain the final theme of motion and consequently achieve motion retrieval. Analysis and retrieval of human motion has experienced a development from continuous to discrete. Treating motion as

discrete leads to the loss of parts of details, but the semantic level represented by data is enhanced. Furthermore, compare with motion editing and synthesis, motion retrieval focuses less on details of data. Therefore, ignoring details and extracting the backbone of information when calculating the similarity between motions would be more helpful in solving the problems.

The key posture that the key frame corresponds to is usually the limit of human motion in posture and also the most symbolic posture with the most abundant contents. This posture plays an essential role in motion fragment. Therefore, a motion can be summarized into semantic information with a sequence of key postures which have been organized. Methods for extracting the key frame of existing motion data can generally be divided into two categories. One is the use of interval sampling to extract the key frame. However, this approach results in oversampling in the slow parts of motion and under-sampling in the parts with drastic changes because the motion rhythm or motion frequency is not the same as that in the original motion. Consequently, details of the original motion are lost. The other category is the self-adaptive key frame extraction method, with which the extracted key frame can automatically represent motion sequences on the basis of the content characteristics of motion capture data. This category is mainly divided into two specific methods: one based on clustering and one based on curve simplification. To perform content analysis of a video, Zhuang et al. [5] used unsupervised clustering method to extract the key frame. Liu et al. [8] used a clustering method that considers the first frame as the key frame with which an index tree of motion is built. Loy et al. [9] adopted the same clustering method but considered the centre of each class as the key frame. Key frames extracted by clustering method can more accurately describe the content of the original samples because clustering method can efficiently group similar samples into the same category. However, this method rarely takes into account the temporal relations among the samples, thereby easily causing distortions in the analysis of motion sequences. Lim et al. [6] was the first to propose a key frame extraction method based on curve simplification [10]. This method takes the motion capture data of each frame as a data point in a high-dimensional space, and the concave and convex points in this curve are regarded as key frames of motion sequences. The disadvantages of this method are that the temporal relations of motion sequences are ignored and that coupling between different motions is conducted. Using simple Euclidean distance to measure the similarity between different data frames cannot accurately reflect the actual difference. Zhang et al. [17] used a similar method to extract the key frame of motion capture data; however, they measured the similarity between data frames by using the amplitude of curve. Yang Tao et al. [15] proposed a key frame extraction of motion capture data based on hierarchical curve simplification. They introduced the bone angle as a motion characteristic for determining the candidate key frames that would be selected as the final set of key frames based on the hierarchical curve simplification algorithm. Bulut et al. [2] used two steps to extract the key frame. First, they proposed a new metric, curve saliency, for motion curves that specifies the most important frames of the motion. Second, they detected the final key frames by clustering the computed important frames. The same technology is applied for key frame extraction of video sequences [7,18].

Nowadays, most of studies of algorithm focus on the improvement of algorithm on the results of Matching and Retrieval. In the definition of semantic features, we currently are still in the stage that based on the underlying physical characteristics or underlying semantic features, and haven't designed a motion features representation method with the high semantic information and meet the laws of cognition of people. However, this paper comes up with an

improved algorithm on Matching and Retrieval, and optimizes the rank of Retrieval results, generate thumbnails with semantic meaning by using key frame creatively, which better solves the problem of semantic gap and contributes to the intuitive of the Retrieval results.

The main contributions of this paper are as follows.

1. The use of SK-means with the logic similarity measurement of warping-direction energy of each node to which its parent node corresponds.
2. The proposal of a key frame extraction method based on visualized data analysis for generating thumbnails with good generalization of semantic information. The proposed method overcomes the loss of information of motion trajectory in the similarity determination phase of basic motion type.
3. Democratic decision-making algorithm is adopted to optimize the rank of retrieval results.
4. Unsupervised learning method is utilized to learn the knowledge input by users and then produce more accurate descriptions of semantics to improve the accuracy of the retrieval.

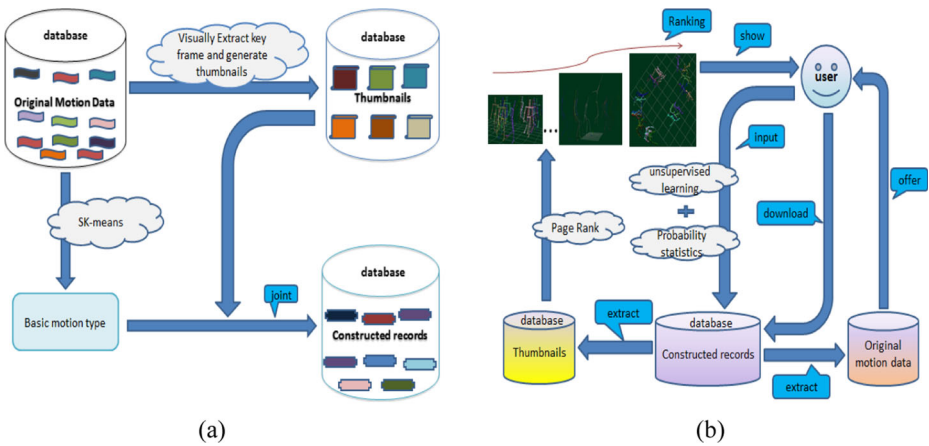
## 2 Framework

The coupling of motions between different joints is very strong, because human motion capture data are specific signals of temporal motion and high-dimensional vectors that indicate body motion capture data. For example, when people walk normally, an association can be clearly observed between the motions of arms and legs. Therefore, hierarchical processing of specific motions can lead to the loss of context, ultimately resulting in the poor matching of motions. At the similarity determination phase of basic motion type, the trajectory information of motion capture data is based on irrelevant details, such as the noise which can be considered unnecessary. For instance, when people walk straight, in circles or backwards, the motion categories belong to one basic type of walk in nature. Therefore, trajectory information is ignored to obtain the best effect of clustering at this phase. Then, we reintroduce the trajectory information in the part of thumbnails that users browse to ensure the integrity of information and improve the accuracy of retrieval. Figure 1 shows the schematic of the method proposed in this paper.

Figure 1a shows the process of clustering all the motion sequences from the motion database and extracting the key frames to generate thumbnails and finally forming structured records. Figure 1b illustrates how the retrieval results are sorted from the structured database depending on the query from the user based on democratic decision making and learning the knowledge from users.

## 3 Generate structured records

An efficient retrieval model requires a good data structure. As shown in Fig. 1, we clustered the motion capture data and extracted the key frames to generate thumbnails and structure them into a record in this paper. By retrieving the record of constructed data only, users can receive the corresponding motion capture data and thumbnails through an efficient algorithm whose time complexity is  $O(1)$ . The structured record consists of `Motion_id`, `Motion_name`, `Motion_basic_type`, `MotionData_path`, `Thumbnail_path`, `Learning_field` and `Probability_statistic`. `Motion_id` represents the unique tag of each record. `Motion_name`

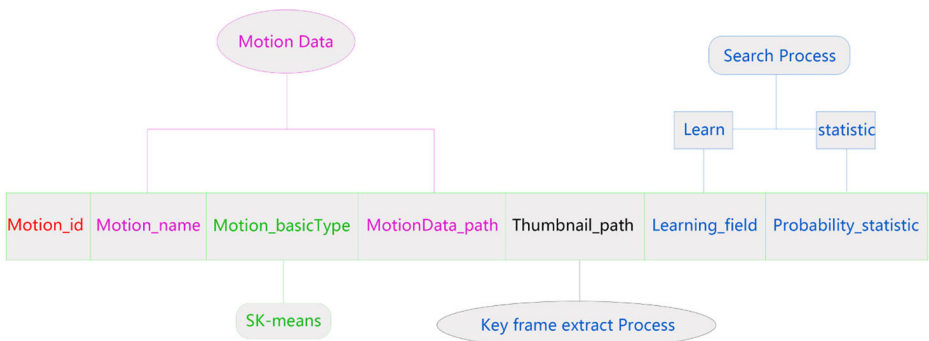


**Fig. 1** Frame work consists of pre-process in (a) and search-process in (b)

represents the name of the motion data. Motion\_basic\_type represents the basic motion type of the motion data. MotionData\_path represents the storage path of the motion data in the Motion database. Thumbnail\_path represents the storage path of the thumbnail generated from the motion data in the Thumbnail database. Learning\_field represents that knowledge library that stores the knowledge learned from users. Probability\_statistic represents the credibility of the basic type to which the motion data belong. Figure 2 shows the organizational structure of the record.

### 3.1 SK-means clustering

K-means clustering algorithm was proposed by Steinhaus in 1955, Lloyd in 1957, Ball and Hall in 1965 and McQueen in 1967 separately in different areas of research. Since then, the K-means algorithm has been widely studied and applied in different subject areas [1] and remains as one of the most popular partition clustering algorithms. Given that various stylized motions are evolved by finite basic types, we firstly chose K motion capture data of basic types as initial cluster centres to perform clustering training. At the same time, we introduced probability and statistics information in the retrieval process to count the retrieval and download the



**Fig. 2** Organization structure of the record

probability of every motion capture data that belong to the same basic type. According to the law of large numbers, motion capture data with a probability of nearly zero in the same type has a high possibility of not belonging to this basic motion type. Therefore, this motion capture data type can be a new basic type as a new cluster center.

In our work, we extended the classic K-means method with the logic similarity measurement of warping-direction energy of each node to which its parent node corresponds. Although traditional K-means can complete clustering well, it needs to repeat the calculation for mean of each changed clustering, thereby reducing the efficiency of the algorithm. SK-means algorithm we proposed puts mean of the  $N$  data with the highest query probability in the same type of data as the clustering centre, and recalculates the mean of clustering only when the probability changes.

**Definition 1:**  $m=[x_1, x_2, x_3, \dots, x_t, \dots, x_f]^T, x_t=[q_{t1}, q_{t2}, q_{t3}, \dots, q_{tj}, \dots, q_{tJ}]$ , where  $m$  denotes the motion capture data and  $x_t$  denotes the angle position represented by angle coordinate (x, y, z) of each joint node relative to the parent node in the  $t$ -th frame.

For a given set  $MS$  containing  $n f \times J$ -dimensional motion capture data, written as  $MS = \{m_1, m_2, m_3, \dots, m_i, \dots, m_n\}$ ,  $m_i \in R^{f \times J}$ , and the quantity named  $K$  of the subset of the data which is going to be generated, SK-means clustering algorithm divides the data into  $k$  partitions and is recorded as  $C = \{c_k, i=1, 2, \dots, K\}$ . Every partition represents a category  $c_k$ , and every  $c_k$  has a category centre named  $\mu_k$ . Warping energy (WE) and direction energy (DE) are combined as judging criteria.

The formula of warping energy is

$$WE = \sum_{t=1}^f \sum_{j=1}^{J-1} \left( we_{tj} = \arccos \left( \frac{q_{tj} \cdot q_{tj+1}}{|q_{tj}| \times |q_{tj+1}|} \right) \right), we_{tj} \in R^{f \times J-1}$$

where  $q$  is a vector represented by angle coordinate (x, y, z) of a joint node of motion capture data  $m$  in Definition 1.

The formula of directory energy is

$$DE = \sum_{t=1}^f \sum_{j=1}^{J-1} \left( de_{tj} = \left| \begin{array}{ccc} \vec{i} & \vec{j} & \vec{k} \\ q_{tjx} & q_{tjy} & q_{tjz} \\ q_{tj+1x} & q_{tj+1y} & q_{tj+1z} \end{array} \right| \right), de_{tj} \in R^{f \times J-1}$$

where  $de_{ij} = q_{ij} \otimes q_{ij+1}$  and  $\otimes$  is a vector cross product operator.

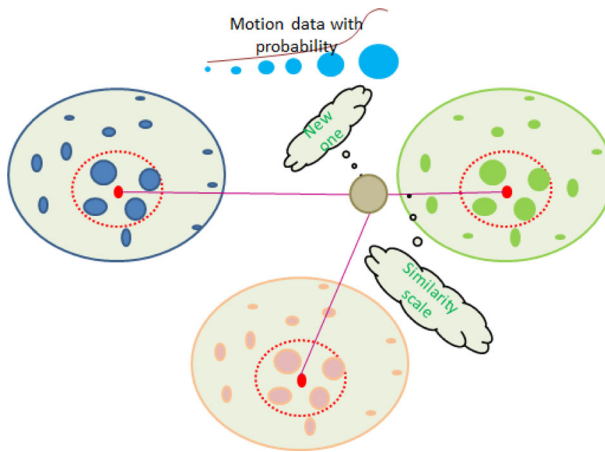
The formula for computing the similarity of each node to the cluster centre in this category is  $LK(c_k) = \sum_{x_i \in c_k} \left\| (WE_{x_i} - WE_{\mu_k}) \times we(DE_{x_i}, DE_{\mu_k}) \right\|^2$  and  $we(DE_{x_i}, DE_{\mu_k})$  has the function to obtain the dot product of  $DE_{x_i}$  and  $DE_{\mu_k}$ .

The objective of clustering is to distribute the inputting data object to the category that the cluster centre to which this data is most similar and  $m_i \in \operatorname{argmin}_{k \in K} (LK(c_k))$ .

Unlike the traditional K-means algorithm, SK-means algorithm only processes the  $k$  records with iteration to obtain the centroid point when the first  $k$  records with the highest probability changes, thereby improving the efficiency of this algorithm. Figure 3 shows the algorithm.

### 3.2 Key frame extraction

The information of human animation is huge, and its dimension is too high. Nonetheless, information visualization technology can analyse high-dimensional datasets; show the patterns



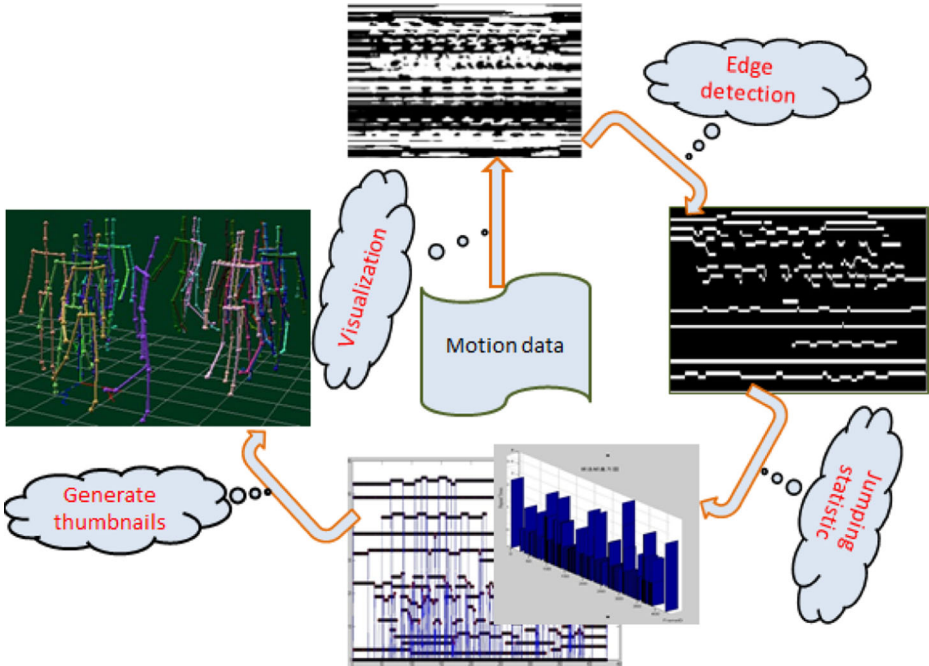
**Fig. 3** Flow of SK-means algorithm

of initial datasets; find outliers in statistics; present the potential knowledge and development trend through the use of graphics, images, virtual reality and other alternatives that people can easily identify and recognize, thus allowing people to make better use of the available information resources. We introduce visual data analysis technique to extract key frame. At the first stage, we mapped the high-dimensional motion capture data to grayscale images according to the line order of the bi-dimensional image. Row order represents the timing sequence of motion; column order represents the rotational component of each joint node. Therefore, this method contains good timing correlation of motion sequence and coupling between different joint motions. At the second stage, we processed the image with edge detection and determined the rotation variable of the main changed joints. At the third stage, we counted the frames to which the jump breakpoint corresponds in the line of edge, and then considered the frames as the set of candidate key frames. At the fourth stage, we extracted the optimal key frame to generate thumbnail from the candidate key frames. Figure 4 shows the process of this algorithm.

### 3.3 Visual data normalization

The human skeleton adopted in this paper consists of 23 joint nodes. We use tree structures to organize each joint node, and the joint node named Hip is the root node of the tree human skeleton. The motion capture data of each frame is based on the angle represented by coordinate  $(x, y, z)$  of the root node in a 3D scene and the rotation partition of each node to which its parent node corresponds. The motion capture data of each frame can be described as a discrete-time vector function  $f(t)=[p_0(t), R_1(t), R_2(t), \dots, R_N(t)]$ , where  $N$  denotes the quantity of joint node,  $f(t)$  represents the motion capture data of the  $t$ -th frame the value of this frame in time dimension can be computed by  $t * F_f$  (Frame frequency), and  $p_0(t)$  represents the position of the root node in the coordinate system.  $R_i(t) \in R^3$  indicates the orientation of the root node in world coordinate system.  $R_i(t) \in R^3$  indicates the rotation of the  $i$ -th node relative to its parent node. Thus, human animation data can be represented using a discrete-time matrix:

$$F = [f(1), f(2), \dots, f(T)]^T (1 \leq t \leq T)$$



**Fig. 4** Thumbnail framework

where  $T$  denotes the total number of frames. Given that the value of the rotation partition of the position of root node and every joint node may exceed 255, we firstly normalized  $F$ :

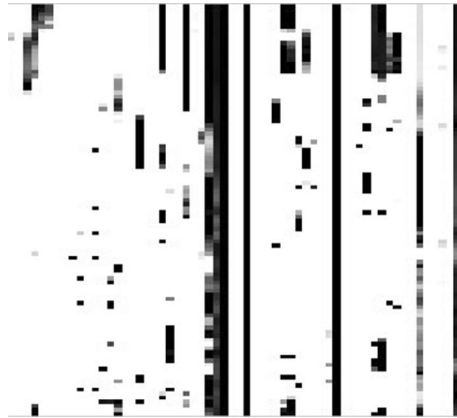
$$\tilde{F} = \left\lfloor \frac{F}{255} \right\rfloor$$

Then, we input the data  $\tilde{F}$  to generate a grayscale image  $I_{\tilde{F}}$ , as shown in Fig. 5.

### 3.3.1 Edge detection

We carefully observed and analysed the motion change image after normalization. The height of the atlas represents the quantity of frame, whereas the width represents the motion capture data defined as  $f(t)$ . By observing the block areas of the atlas longitudinally, we can find that rotation variables of these joints do not change drastically or even do not change over time, and the edge part of each block is the portion of motion mutation. Given that the Canny edge detection algorithm is based on the first derivative of Gaussian function, the Canny operator is symmetric in the edge direction and is anti-symmetric in the direction perpendicular to the edge, thereby indicating that this operator is especially sensitive to the edge in the direction with the most drastic changes. To demonstrate the human motion changes, we processed the motion atlas with edge detection adopting the Canny operator to obtain the most intuitionistic motion edge. Figure 6 shows the detection results of different kinds of operators.





**Fig. 5** Atlas of human motion

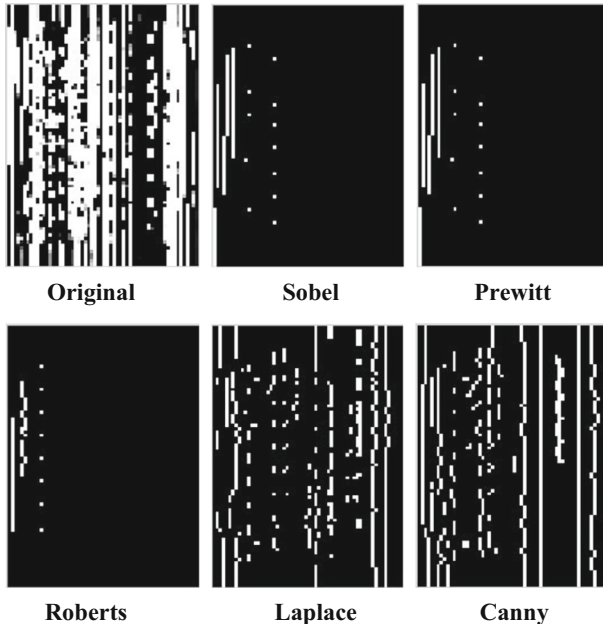
The process of Canny operator edge detection:

Step 1. The image is smoothed with Gaussian filter.

$$\text{Bidimensional Gaussian function: } G(x, y) = \frac{1}{2\pi\delta^2} \exp\left(-\frac{x^2+y^2}{2\delta^2}\right).$$

The first derivative of  $G(x, y)$  in the direction of  $n : G_n = \frac{\partial G}{\partial n} = n \nabla G$ ,

$$n = \begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix}, \nabla G = \begin{bmatrix} \frac{\partial G}{\partial x} \\ \frac{\partial G}{\partial y} \end{bmatrix}, \text{ and } n \text{ is direction vector, } \nabla G \text{ is gradient vector. A}$$



**Fig. 6** Effect diagrams of edge detection with each operator

convolution is executed on the motion change atlas  $F(x,y)$  and  $G_n$ , and at the same time, the direction of  $n$  is changed.  $n$  is the direction of both orthogonality and edge detection when obtaining the max value by  $G_n * F(x,y)$ .

Step 2. The magnitude and direction of the gradient is computed with the finite different of first partial derivative.

$$E_x = \frac{\partial G}{\partial x} * F(x,y), E_y = \frac{\partial G}{\partial y} * F(x,y), A(x,y) = \sqrt{E_x^2 + E_y^2}, \theta = \text{Arc tan} \left( \frac{E_x}{E_y} \right)$$

where  $A(x,y)$  indicates the edge intensity of the image at point  $(x, y)$ , and  $\theta$  is the normal vector of the image at point  $(x, y)$ .

Step 3. The gradient amplitude is processed with non-maxima suppression.

Step 4. The edges are detected and connected by double threshold algorithm.

Figure 7 shows the Canny edge detection result of the motion change atlas.

### 3.3.2 Optimal key frame extraction

Considering the coupling of every joint node of the human skeleton, we counted the number of jump breakpoints of every rotation partition of every joint node in every data frame. We then took the frame as candidate key frames and the number of jump breakpoints as the weight of this frame. Given that the line sequence represents temporal relations, it equips with a good ability of temporal constraint. By counting the edge jump points as the weight of candidate key frame, good information is obtained of the edge postures. The candidate key frame extraction algorithm records the identity of the key frame and the quantity of jump point in this frame with a dynamic bi-dimensional array called *CKeyArrays*, where *cmp* is an auxiliary array,  $T$  denotes the number of frames,  $W$  represents the length of a frame, and  $t$  and  $w$  indicates the loop variable in the line and the row, respectively.

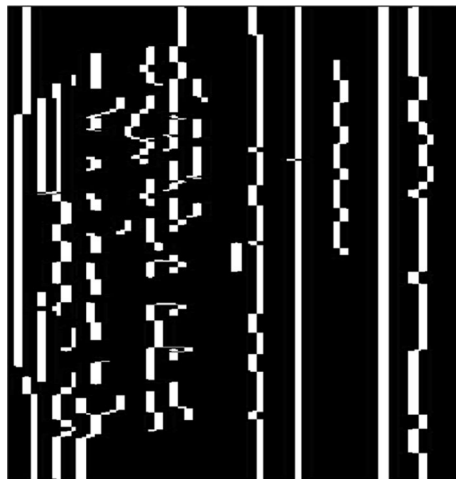


Fig. 7 Main activity atlas

- Step 1. The first frame is taken as the key frame and is added to *CKeyArrays*. Then, variables are initialized:  $t=1$ ,  $w=1$ ,  $cmp=F(2,1)$ .
- Step 2. The algorithm is ended if  $t>T$ ; otherwise, proceed to Step 3.
- Step 3. Proceed to Step 5 if  $w>W$ ; otherwise,  $F(t, w)$  and  $cmp(w)$  are compared. This frame is added to *CKeyArrays*, and the jumping count is added by one if they are different. Otherwise, proceed to Step 4.
- Step 4. If  $w=w+1$ , return to Step 3.
- Step 5. If  $t=t+1$ , return to Step 2;
- CKeyArrays* records all the candidate key frames extracted and jumping counts when the algorithm ends.

Motion capture is equipped with a high sampling rate; human joints have the feature of coupling; and every joint can change simultaneously. Thus, we need to consider the temporal constraint and global constraint simultaneously. Firstly, we clustered the set of candidate key frame on the basis of weight and calculated the number of the candidate frames whose weights are same. Then, according to the expected number of inputted key frames, we extracted the key frames with the highest priority numbers by using an algorithm similar to processes scheduling algorithm.

Input: the expected number of key frames, recorded as  $K_{num}$

- Step 1. Candidate key frames are clustered, and *CKF* is calculated.
- Step 2. If  $K_{num}>C_{num}$ , this type of key frame is added to the optimal key frame named *KeyFrames* and proceed to step 3; otherwise,  $CF_{i:knum}$  is added to *KeyFrames* and proceed to step 4;
- Step 3. If  $i=i+1$ ,  $K_{num}=K_{num} - C_{num}$  and return to step 2.
- Step 4. End.
- KeyFrames* records the optimal key frames of human motion when the algorithm ends. Thumbnails generated by these key frames are characterized by good semantic generalization.

### 3.4 Democratic decision and unsupervised learning

The fundamental principle of Google's "PageRank" is that if many web pages link to a page named A, then A is widely recognized and trusted, and its ranking will be high. For the democratic decision-making of this article, the user downloads certain motion data for most, indicating that the probability of this motion data items belonging to the basic types is highest. Using the N motion data with the highest credibility to calculate the mean of clustering as the clustering centre in the SK-means clustering algorithm not only improve the accuracy of clustering, but also improve the efficiency of the algorithm.

In the process of query, we use unsupervised learning algorithm to learn new knowledge from users, and store every new stylized knowledge item in *learning\_field*. In the query matching process, first do the work of basic movement types match, then match the domain of *learning\_field* to implement stylization query.

## 4 Experiment

To validate the effectiveness and efficiency of the proposed algorithm, we test 450 different motion data clips on ACCAD [[http://accad.osu.edu/research/mocap/mocap\\_data.htm](http://accad.osu.edu/research/mocap/mocap_data.htm)] and CMU [<http://mocap.cs.cmu.edu>] motion database. As following: walk forward 70 clips, run forward 62 clips, walk backward 54 clips, jump 48 clips, skipping 45 clips, picking 36 clips, walk turn change 32 clips, kicking 33 clips, climb 28 clips, box 18 clips, crouch 12 clips, side step 12 clips. There are three steps in our experiment. Firstly, we test SK-means algorithm compared with other state of art algorithm. Secondly, we test the extraction of key frames and then generate thumbnails. Finally, we test the retrieval algorithm compared with other method.

### 4.1 SK-means clustering

Traditional K-means clustering algorithm needs to cluster all the objects iteratively, and the efficiency of this algorithm significantly decreases as the clustering library continues to increase. We improved the traditional K-means algorithm by introducing statistics and recalculating the cluster centre only when the first N records with the highest statistical probability change. As a result, not only is efficiency improved, but the interference of the initial clustering error is reduced. Table 1 shows the compared result of clustering with other algorithm.

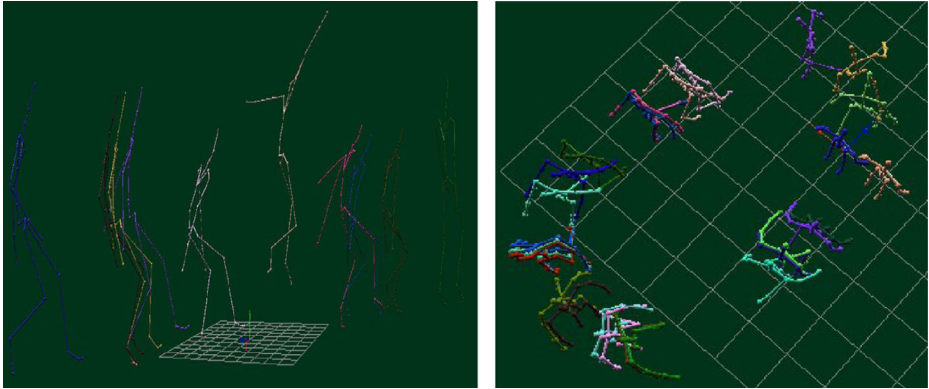
### 4.2 Thumbnails

To achieve the extraction of candidate key frames and then generate thumbnails, we used Visual C++ 6.0 and OpenGL to program the motion editing system, and we use the optimal key frames extraction algorithm to extract the set of key frames after using MATLAB. In this experiment, the original animation of human walking around an entire circle has 429 frames at a rate of 30 frames per second. The performer first walks half a circle, goes straight, and then does the same thing to return to the origin. Firstly, the normalized motion atlas is processed with edge detection using MATLAB. Then, candidate key frame extraction algorithm is used to extract 164 candidate key frames, ensuring that the compression ratio of each reaches 38.2 %. Lastly, optimal key frame extraction algorithm is used to extract 21 key frames, making the compression ratio of each reach 4.89 %. The experiment was operated with a memory of Intel(R) Core(TM)i3 2.27GHz CPU 4GB.

Figure 8 presents the motion capture data of two different types of motion. Thumbnails generated from the set of key frames by the key frame extraction algorithm proposed in this paper effectively generalize the semantic information of these two different types of motion

**Table 1** Clustering accuracies compared with other approaches

Method	Clustering accuracy	Learning method	Total time
Motion string [14]	89.2 %	unsupervised	62 s
Hierarchical Tree [13]	93.6 %	unsupervised	82 s
Proposed method	92.1 %	unsupervised	56 s

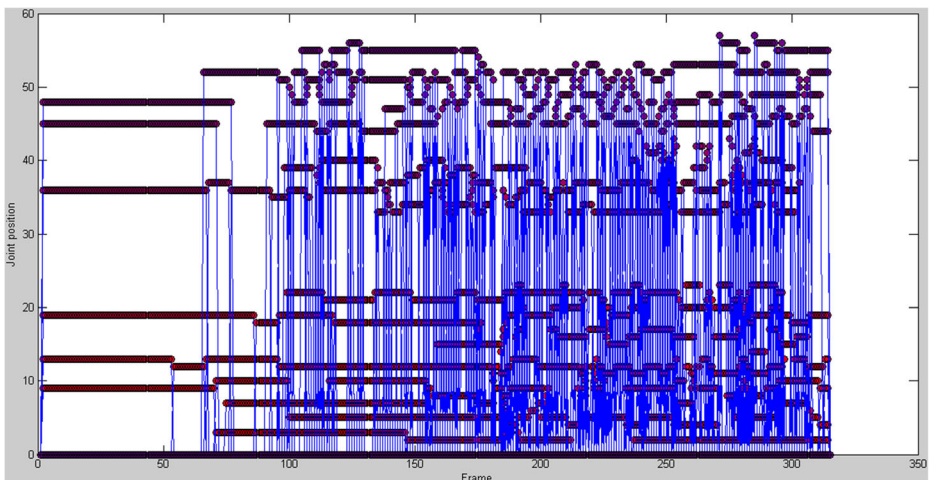


**Fig. 8** Thumbnails of two different types of motion

capture data. To generate a thumbnail with good semantic generalization for motion capture data with both gentle and intense motions, in theory, we need to process the gentle period with dilution sample and process the intense period with dense sample to avoid problems of oversampling and under-sampling. Figure 9 shows that by using the proposed algorithm on data with gentle and intense motions, we can obtain more jump points to process the data well with dense samples in the intense period and obtain fewer jump points to process the data with dilution sample in the gentle period (Fig. 9).

We extracted the key frames of two different types of motion capture data, including complex motions and simple motions, on the ACCAD database and counted the time consumed by this algorithm, the number of frames extracted, and the visual effect. Table 2 shows the statistics.

This experiment indicates that the proposed algorithm obtains good results in terms of clustering motion capture data and to some extent, overcomes the difficulty of manual labelling. In addition, the proposed algorithm selects the set of candidate key frames based



**Fig. 9** Jump points chart

**Table 2** Key frame extraction of motion capture data with different complexities

Motion name	Total frames	Key frames	Algorithm time
RunToCrouch	126	25	5.132 s
RunTurnAround	101	15	3.982 s
RunTurnLeft90	67	10	2.324 s
PutDownBoxToRun	87	8	1.765 s
Skipping	114	20	4.507 s
StandToSkip	143	15	4.987 s
QuickSideStepLeft	90	20	4.304 s
Crouch	180	40	5.038 s
RunBack	76	5	1.756 s
Run	36	5	1.643 s
Walk	142	15	4.649 s
Walk2Hop2Walk	143	15	4.987 s
WalkTrunChange	107	40	5.038 s
WalkTrunLeft45	114	20	4.507 s
WalkToPickUpBox_Working	105	40	5.961 s

on jump points on edge to contain the motion edge postures. Moreover, it samples more during periods with intense motion changes and less during gentle periods to avoid oversampling and under-sampling. Further, it can efficiently generate thumbnails with good visual effect, both for complex combination motions and simple motions. In consideration of the above advantages as well as with the use of democratic decision making and unsupervised learning at the searching phase, the algorithm allows for a more intuitive and accurate retrieval, and data are organized automatically.

### 4.3 Retrieval

User can input the stylized semantic information and keyword of basic motion type into the field of *exactly describe* and *basic type*. Because of all the movement of human body are made of several types of basic movement with some stylized characteristics. For example: rapidly drop down after a sneaky walk, which walk and drop down are the basic type and body movement, and sneaky and rapidly are the stylized semantic description. In this paper, we use 12 kinds of basic movement type data with stylized information for retrieval, and compare it with the existing retrieval method, Table 3 shows the accuracy of retrieval and comparison.

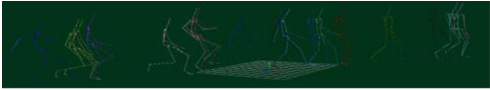
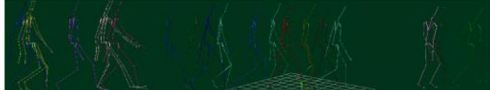




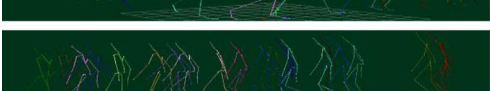
**Table 3** Retrieval accuracy compared with other approaches

Methods	Retrieval accuracy				
	Jumping	Running	Walking	kicking	Average
Motion hierarchy indexing [4]	0.8162	0.8145	0.8123	0.8174	0.8151
Proposed method	0.8313	0.8387	0.8373	0.8392	0.8391

**Query Motion:**

Walk(Basic Type)+Skip(exactly description)

**Retrieved results:**

Rank1		Learning field: skip Probability statistic:1.02%
Rank2		Learning field: walk Probability statistic:2.03%
...		
Rank15		Learning field: pick box Probability statistic:1.14%
Rank16		Learning field: with box Probability statistic:0.89%
...		
Rank68		Learning field: crouch Probability statistic:0.32%
Rank69		Learning field: turn right 90 Probability statistic:0.24%
...		
Rank108		Learning field: backwards Probability statistic:0.02%

**Fig. 10** Example retrieval result of our proposed method

The hierarchical indexing method contains more detail and is more scalable, hence the retrieval could work better in high-recall region. In addition, it segments the motion into a sequence of sub-moves such that the temporal detail is preserved. Our method outperforms the existing methods because the selected features from the warping energy (WE) and direction energy (DE) can effectively abstract the logical meaning of each input motion. Even when a new motion is added to the database, our system can still suggest an initial set of good features. Our proposed method is useful to retrieve different kinds of query motions. Figure 10 illustrates an example of retrieval result using our proposed method.

## 5 Conclusion and future work

Human motion capture data has far exceeded the dimension of human cognize, thus being referred to as high-dimensional motion capture data. Thus, humans find it difficult to analyse such data and understand its inherent laws. Information visualization can analyse a set of this high-dimensional data, show the pattern of the original data and find the outliers in statistic with known methods, including graphics, images and virtual reality. Inspired by this fact, we changed high-dimensional motion capture data into

image data, in consideration of the characteristic that an image's edge changes violently. We generated thumbnails with good semantic generalization after processing human motion capture data with edge detection, such that users can browse the motion type of motion capture data intuitively. With regard to organizing a great deal of motion capture data, we propose a good clustering algorithm which improves the efficiency and achieves good clustering effect. Meanwhile, we refer to democratic decision-making algorithm and unsupervised learning to improve the accuracy and efficiency of motion capture data retrieval.

This experiment considers the basic human motion types that we selected as cluster centres. Given that the basic types that we summarized is not comprehensive enough and that the contents of the retrieval library used in the experiment remains insufficient, we intend to expand the quantity of data in the retrieval library and consequently summarize the basic human motion types more accurately.

**Acknowledgments** This work was supported by National Science Foundation of China(NO.61303142, 60970021,61173096),Natural Science Foundation of Zhejiang Province(NO. Y1110882,Y1110688,R1110679), Higher School Specialized Research Fund for the Doctoral Program.(NO.20113317110001).

## References

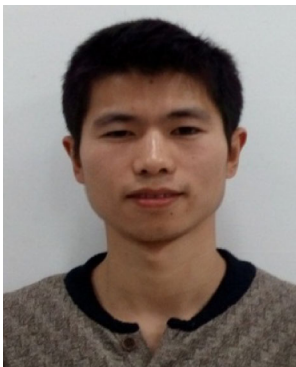
1. Anil KJ (2010) Data clustering: 50 years beyond K-Means. *Pattern Recogn Lett* 651–666
2. Bulut E, Capin T (2007) Key frame extraction from motion capture data by curve saliency, *proc. Comp Anim Soc Agent* 63–67
3. Chao MW, Lin CH, Assa J, et al (2012) Human motion retrieval from hand-drawn sketch., *IEEE Trans Vis Comput Graph* 729–740
4. Deng ZG, Gu Q, Li Q (2009) Perceptually consistent example-based human motion retrieval, *proc. Computer graphics proceedings, annual conference series, ACM S IGGGRAPH*. ACM Press, New York, pp 191–198
5. Ioannidis AI, Chasanis VT, Likas AC (2014) Key-frame extraction using weighted multi-view convex mixture models and spectral clustering, *proc. Pattern Recog 22nd Int Conf IEEE (ICPR)* 3463–3468
6. Lim IS, Thalmann D (2001) Key-posture extraction out of human motion data by curve simplification, *Proc. the 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Istanbul* 1167–1169
7. Lim IS, Thalmann D (2001) Key-posture extraction out of human motion data by curve simplification, *proc. Proc EMBC* 1167–1169
8. Liu F, Zhuang YT, Wu F, et al (2003) 3D motion retrieval with motion index tree. *Comput Vis Image Underst* 265–284
9. Loy G, Sullivan J, Carlsson S (2003) Pose-based clustering in action sequences, *Proc. the 1st IEEE International Workshop on Higher-Level Knowledge in 3D Modeling and Motion Analysis, Nice* 66–72
10. Miura T, Kaiga T, Shibata T, et al (2014) A hybrid approach to key frame extraction from motion capture data using curve simplification and principal component analysis. *IEEE Trans Electr Electron Eng* 697–699
11. Numaguchi N, Nakazawa A, Shiratori T, et al (2011) A puppet interface for retrieval of motion capture data, *Proc. 2011 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*. ACM 157–166
12. Tang JKT, Leung H (2012) Retrieval of logically relevant 3D human motions by adaptive feature selection with graded relevance feedback. *Pattern Recogn Lett* 420–430
13. Wu SG, Wang ZQ, Xia SH (2009) Indexing and retrieval of human motion data by a hierarchical tree, *proc. 16th ACM symposium on virtual reality software and technology*. ACM Press, New York, pp 207–214



14. Wu S, Xia S, Wang Z, et al (2009) Efficient motion data indexing and retrieval with local similarity measure of motion strings. *Vis Comput* 499–508
15. Yang T, Xiao J, Wu F, et al (2006) Extraction of key-frame of motion capture data based on layered curve simplification. *J Comput Aided Des Comput Graph* 1691–1697
16. Zhang Z P (2008) Content-based motion retrieval using vector space model. Massachusetts Institute of Technology. Department of Electrical Engineering and Computer Science
17. Zhang Q, Xue X, Zhou D, et al (2014) Motion key-frames extraction based on amplitude of distance characteristic curve. *Int J Comput Intell Syst* 506–514
18. Zhao L, Qi W, Li S Z, et al (2000) Key-frame extraction and shot retrieval using nearest feature line (NFL), *proc. Proceedings of ACM Workshops on Multimedia*, Los Angeles, CA 217–220



**Xin Wang** is an associate professor at Zhejiang University of Technology. He received the Doctor degree from the Zhejiang University in 2009, he is a member China Computer Society. He is currently actively engaged in computer vision technology and 3D animation technology combining visual animation theory. His research interests include computer animation, character animation, motion capture technology.



**Liangxiu Cheng** is a postgraduate student of Zhejiang University of Technology. His research interests include computer animation, motion capture technology.



**Jiali Jing** is the undergraduate students of Zhejiang University of Technology. Her interests include computer animation



**Herong Zheng** is a professor at Zhejiang University of Technology. He received the Doctor degree from Zhejiang University of Technology in 2009. Since 2009 he is the Zhejiang Province Key Laboratory of intelligent visual media processing technical director of General Office and a professor at the Department of computer science and technology. His research interests include Graphic image, intelligent recognition technology, two-dimensional bar code technology