Efficient Human Motion Retrieval via Temporal Adjacent Bag of Words and Discriminative Neighborhood Preserving Dictionary Learning

Xin Liu, *Member, IEEE*, Gao-Feng He, Shu-Juan Peng, Yiu-ming Cheung, *Senior Member, IEEE*, and Yuan Yan Tang, *Fellow, IEEE*

***Abstract*—Human motion retrieval from motion capture data forms the fundamental basis for computer animation. In this paper, the authors propose an efficient human motion retrieval approach via temporal adjacent bag of words (TA-BoW) and discriminative neighborhood preserving dictionary learning (DNP-DL). The retrieval process includes two phases: offline training and online retrieval. In the first phase, the original skeleton model is first simplified and then pairwise joint distances are computed to characterize each motion frame. Then, a novel motion descriptor, namely TA-BoW, is proposed to discriminatively code the motion appearances, through which the articulated complexity and spatiotemporal dimensionality can be greatly reduced. Subsequently, by considering the neighborhood relationships of intraclass structure and the advantage of Fisher criterion, a DNP-DL method is exploited through which each human action can be discriminatively and sparsely represented by a linear combination of such dictionary atoms. In the second phase, a hierarchical retrieval mechanism is used by incorporating the sparse classification and chi-square ranking, whereby the searching range is significantly reduced. The experimental results show that the proposed human motion retrieval approach performs better than the state-of-the-art competing approaches.**

***Index Terms*—Hierarchical retrieval mechanism, human motion retrieval, neighborhood preserving dictionary, pairwise distance, temporal adjacent bag of words (TA-BoW) .**

Manuscript received February 17, 2017; accepted February 18, 2017. Date of publication March 23, 2017; date of current version November 13, 2017. This work was supported in part by the National Science Foundation of China under Grant 61673185, Grant 61272366, and Grant 61672444, in part by the Science and Technology Research and Development Fund of Shenzhen under Project Code JCYJ20160531194006833, in part by the National Science Foundation of Fujian Province under Grant 2015J01656 and Grant 2017J01112, in part by the Promotion Program for Young and Middle-Aged Teacher in Science and Technology Research under Grant ZQN-PY309 of Huaqiao University, in part by the Faculty Research Grant FRG2/14-15/075, Grant FRG2/15-16/049 of HKBU, in part by the Research Grants of University of Macau under Grant MYRG2015-00050-FST, and in part by the Macau-China Joint Project 008-2014-AMJ. This paper was recommended by Associate Editor J. Civera. *(Corresponding author: Xin Liu.)*

X. Liu, G.-F. He, and S.-J. Peng are with the Department of Computer Science, Huaqiao University, Xiamen 361021, China (e-mail: starxliu@163.com; zziahgf@qq.com; pshujuan@hqu.edu.cn).

Y. M. Cheung is with the Department of Computer Science, Hong Kong Baptist University, Hong Kong, and also with the HKBU Institute of Research and Continuing Education, Shenzhen 518057, China (e-mail: ymc@comp.hkbu.edu.hk).

Y. Y. Tang is with the Department of Computer and Information Science, University of Macau, Macau 999078, China (e-mail: yytang@umac.mo).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/THMS.2017.2675959

I. Introduction

REAL-TIME motion capture (mocap), aimed at gaining a precise representation of the complex human or other object movements, has recently sparked a revolution in the computer animation industry. Consequent to this, human mocap data analysis has, of late, evoked considerable interest by virtue of its attractive applications, ranging from data-driven animation, virtual reality, three-dimensional (3-D) film production, sports biomechanics, athletic training, etc. [1]. For instance, interactive and touchless games became tractable only after using these types of accurate data. With mounting popularity of commercial systems, such as wearable kinematic sensors and multiview cameras, a large amount of precise human mocap data has been increasingly recorded. But, the acquisition of such mocap data is prohibitively expensive, and this emphasizes the growing need for reusing the previously recorded data. Further, a number of research domains, such as human motion synthesis, motion style translation and motion editing, have emerged recently to reuse the precaptured mocap data. For instance, animators often prefer to create new animations by working on prior examples. One of the key issues of reusing mocap data is retrieving a specific motion sequence from a large data repository, which has proven to be an extremely challenging task, because human motion always exhibits highly articulated complexity, in both spatial and temporal domains.

Human mocap data is generally specified by a group of mechanical degrees of freedom in the body, and the motion sequence with a particular semantic meaning can be popularly utilized for character animation. In general, the human motion retrieval problem can be stated as the process of automatic searching for a semantically correlated motion clip from an existing mocap corpus, which comprises mainly two key issues: feature representation and similarity matching; the former aims at characterizing the motion appearance and increasing its separability, while the latter is designed for motion comparison. In general, logically similar motions are not matched numerically. Although different feature representations and matching approaches have been addressed, precise retrieval of a specific human action is still a nontrivial task because of its spatiotemporal and articulated complexity.

To bridge the semantic gap between logical similarity, as perceived by humans, and computable similarity, the authors

propose here an efficient human motion retrieval approach via temporal adjacent bag of words (TA-BoW) and discriminative neighborhood preserving dictionary learning (DNP-DL). The proposed approach improves the state-of-the-art methods by providing the following three contributions.

1) A novel motion descriptor, namely, TA-BoW, is proposed to discriminatively characterize the human motion sequence, through which the articulated complexity and spatiotemporal dimensionality can be greatly reduced.
2) In DL framework, the proposed DNP-DL algorithm aims at not only preserving the neighborhood relationships of intraclass structure, but also at encouraging the discriminability among the interclass variances.
3) A hierarchical retrieval mechanism is exploited to facilitate coarse-to-fine similarity matching. Without complex time-alignment, the proposed approach holds a higher discrimination power to achieve an efficient retrieval task.

The remaining part of this paper is organized as follows. Section II will briefly survey the related work. Section III presents the proposed approach in detail. Section IV introduces the experimental results and makes extensive comparisons. Section V sums up the conclusions drawn for this study.

II. Related Works

In general, human mocap data consists of red–green–blue (RGB) images, skeleton joints, and depth maps [2]. Among these, skeleton joints are generally more productive for computer animation. In the literature, the research areas of human action recognition and retrieval are similar, but not identical. Inherently, action recognition is generally regarded as a process of recognizing the semantic meaning, and action retrieval as a process of retrieving similar motions. Although these two topics are a little different, they often share the same motion representation. In this section, an extensive survey is carried out on motion retrieval works from skeleton joints, followed by presenting some typical motion recognition works.

The representative features that characterize the inherent motion characteristics are crucial to motion analysis. As human movement is of high dimensionality, an intuitive way for characterization is to transform the mocap data into a low-dimensional representation. With this approach, Li *et al.* [3] applied the singular value decomposition (SVD) on multiattribute motion matrices and obtained one representative vector for motion indexing, while Pradhan and Prabhakaran [4] utilized the SVD for indexing the subbody motions in a reduced space. These two approaches demonstrate that similar motion sequences have almost linearly correlated eigenvectors, while different motion sequences have diverse eigenpatterns. Heuristically, Barbic *et al.* [5] utilized the principal component analysis (PCA) to detect inherent motion changes, while Forbes *et al.* [6] presented a weighted PCA to increase the importance of some key joints. Recently, motion representation, characterized by manifold learning [7], eigenjoints [8] and eigenvector [9], was also exploited for mocap data analysis. Although these dimension-reduction methods succeeded in capturing significant motion variances, they often failed in capturing a lot of subtle information concerning the articulated joint movements. As a result, these approaches may degrade the primitive postures and lead to a poor retrieval performance.

To avoid loss of information in the transformed low-dimensional space, some researchers chose to extract representative motion features in a semantical way. For instance, Müller *et al.* [10] defined a class of boolean features and utilized a group of motion templates (for their motion classification [11]. Although this semantically interpretable feature has demonstrated its scalability and efficiency in motion retrieval, the specification of well-defined geometric relationships is very difficult for highly dynamic human motions (e.g., dancing). To overcome this problem, Raptis *et al.* [12] presented an angular skeleton representation to recognize the dance actions, while Vieira *et al.* [13] proposed a group of joint distance matrices for motion classification. Similarly, a group of kinetic interval features [14], a histogram of oriented displacements (HOD) [15] and a local skeletal quad [16] were also studied for motion characterization. In general, these motion features can reduce the motion complexity. However, the temporal dynamics of joint movements have not been adequately investigated, and hence some ambiguous motions may fail in being identified. To tackle this problem, an action graph [17], hidden Markov models (HMM) [18] and spatiotemporal body parts [19] were proposed, which can describe the temporal visual movements within the 3-D joint locations. Evidently, these approaches depend highly on a consistent semantic period, failing which their performance would be degraded to some extent [20]. Until very recently, Kapsouras and Nikolaidis. [21] have been utilizing joint orientation angles and their forward differences to represent an action. But, their approach cannot distinguish a given movement from its reverse, e.g., forward walking from backward walking. In addition, a group of motion strings [22], joint-angle rotations [23], motion keys [24], covariance of 3-D joints [25] and a sequence of the most informative joints [26] have also been exploited in modeling the motion sequences by synchronous consideration of temporal dynamics. Nevertheless, those approaches generally require some prior knowledge to stipulate the key poses or motion keys, which would not be readily accessible to the novice users in the real-world applications.

Recently, sparse representation has been demonstrated to be a powerful tool for data representation [27]. Motivated by this success theory, Zhu *et al.* [28] proposed a sparse decomposition model to encode human motions, while Qi *et al.* [29] sparsely encoded a group of key poses for motion representation. However, those two approaches did not consider the temporal dynamics within motions, because of which the retrieval performances were a bit poor. To overcome this problem, Zhou *et al.* [30] exploited a temporal sparse representation to characterize the motion and utilized spatiotemporal pyramid matching (STPM) to achieve motion retrieval. Since the human motions are mostly natural activities, the semantically similar motions (e.g., walking) may have large variations while some other diverse actions can be very similar to each other. Consequently, such sparse representations may fail to discriminatively represent those motions as well.

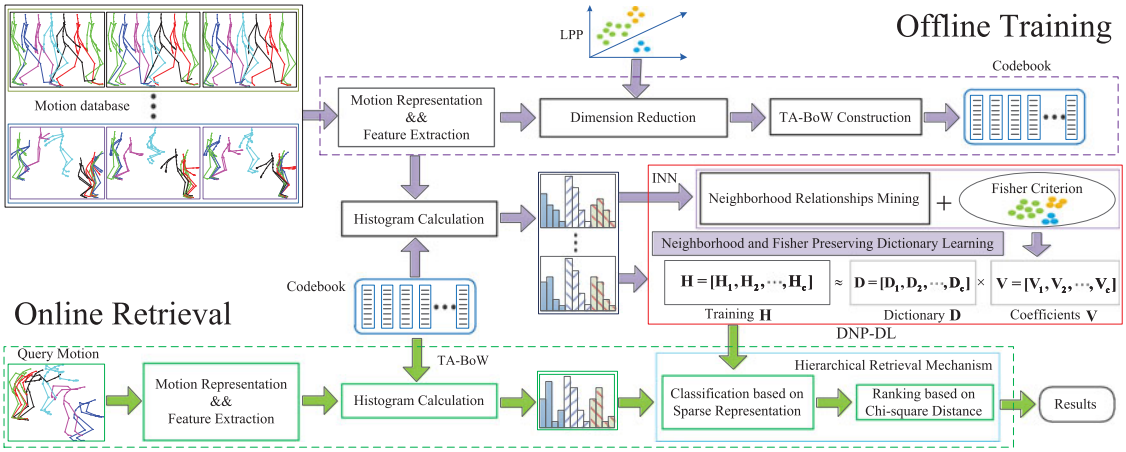


Fig. 1. Schematic pipeline of the proposed human motion retrieval framework.

Similarity matching is another key factor in motion retrieval systems. To measure motion similarity, Adistambha *et al.* [31] utilized dynamic time warping (DTW) for computing motion distance. But, the performance of DTW highly depends on a consistent semantic period; besides, the motion clips with similar semantic meanings are usually not ideally aligned. Some extensions of DTW [32] also suffered similarly from large temporal misalignment. Although the recent isotonic canonical correlation analysis (ISOCCA) [33], canonical time warping (CTW) [34], and correlation-optimized time warping (CoTW) [35] algorithms have delivered a good performance in temporal motion alignment, they are always constrained by computational complexity and exhaustive search for optimum parameters. Histogram-based similarity measurement has recently emerged as another popular matching scheme [36]–[38]. For instance, Barnachon *et al.* [36] employed the Bhattacharyya distance to measure the motion difference between two incremental histograms, whereas Fotiadou *et al.* [38] constructed a pose correspondence matrix to characterize the similarity between two pose histograms. Although those histogram-matching approaches could reduce the impact of motion misalignment, they could not properly compare the temporal dynamics. As a result, matching of some semantically related motions may fail. Therefore, there is still a need to develop a practical and efficient similarity matching algorithm.

III. Proposed Methodology

In general, logically similar motions may not be numerically similar; besides, the articulated complexity of joints often render motion retrieval difficult. To address this issue, the authors present an efficient human motion retrieval approach via TA-BoW and DNP-DL, as shown in Fig. 1. The proposed approach aims not only to discriminatively model the human motion with internal temporal constraint, but also to exploit an efficient approach for motion retrieval. The proposed approach is explained below in detail.

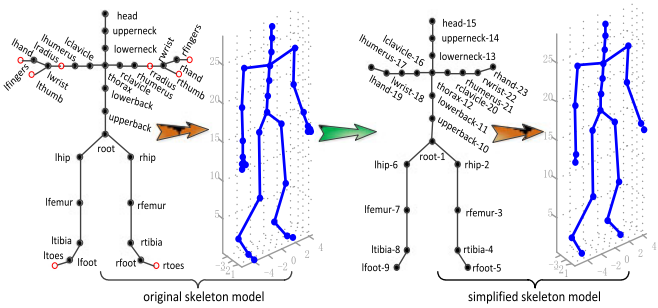


Fig. 2. Selected skeleton joints within HDM05 human mocap data.

A. Motion Representation

In general, the human skeleton model is recorded by a series of articulated joints, which can produce different skeletal poses. In this paper, the typical HDM05 mocap dataset [39] has been selected for illustration, which is often shared by other related datasets for similar representations. HDM05 mocap dataset provides a skeletal model of 31 joints, as shown in Fig. 2. Interestingly, some joints inherently contribute less to the motion analysis, and it is reasonable to simplify the original skeleton model by retaining only 23 joints. The eliminated joints, i.e., the joints rtoes/ltoes, rfingers/lfingers, rthumb/lthumb, rradius/lradius, are marked by the red circle.

It can clearly be seen from Fig. 2 that the simplified skeleton model looks almost the same as the original model. In physical terms, each recorded joint can be further represented by a 3-D position in space. Accordingly, a motion clip M, consisting of n frames, can be described as follows:

$$M = [\mathbf{f_1}, \mathbf{f_2}, \ldots, \mathbf{f_n}] \in R^{3J \times n} \tag{1}$$

where $\mathbf{f_t} = [p_1, \ldots, p_J]^t$ represents one pose at time t, $p_j = (x_j, y_j, z_j)$ is the 3-D position of jth joint in the space, and J denotes the total joint number in the skeleton model.

In practice, joint positions may not well characterize the motion frame discriminatively. Therefore, joint rotations or angles can be employed to characterize articulated movement [12], but it has been found that they have limited capability in expressing

a large variety of motions [30]. For instance, one rotation value, representing the left foot in front of the plan, may be spanned by the right foot. In addition, those types of features may suffer from periodicity of angles. Often, two poses are considered identical, when one of them remains the same even after adding 2π to the other pose vector. As suggested in [13] and [40], the joint pairs are able to reduce the ambiguity between different actions. Inspired by this finding, the pairwise distance $d(i,j)$ between each pair of joints $\{p_i, p_j\}$ has been calculated

$$d(i,j) = \|p_j - p_i\|_2, \quad j > i. \tag{2}$$

It is to be noted that $d(i,j) = d(j,i)$ and M can be represented in terms of a pairwise distance matrix

$$M = [\mathbf{d_1}, \mathbf{d_2}, \ldots, \mathbf{d_n}] \in R^{m \times n} \tag{3}$$

where $\mathbf{d_i}$ is a column vector consisting of all the pairwise distances, and $m = J(J-1)/2$ is the number of all the computed pairwise distances. Often, the motion sequences may contain different kinds of actions and that the pairwise joint distances are sensitive to different performers. To make these motions comparable, the skeleton structure is normalized following the work of [41]. Given the joint p_i and its parent joint p_j, the retargeted position of joint p_i can be calculated

$$p_i^r = p_j + u^i \times b(i,j), u_i = \frac{p_i - p_j}{|p^i - p^j|} \tag{4}$$

where $b(i,j)$ is the initialized bone length between joints p_i and p_j. Evidently, the retargeted joint depends on its parent joint, and the process of normalization of all joints is best begun with the root joint. The skeleton model can thus be well normalized, after which the transferred motions will remain adaptive to different performers.

B. Temporal Adjacent Bag of Words

Intuitively, the pairwise distance matrix can well characterize each motion sequence. However, the whole distance matrix, incorporating all the joint pairs, would significantly increase the computational load because of its high dimensionality. As indicated in [13], the distance matrix with strong correlations is mainly characterized by four lines, and its low dimensional representation would be more effective for motion indexing. In the past, different dimensionality reduction techniques have been developed, e.g., linear PCA [5] and nonlinear Locality Preserving Projections (LPP) [42]. Nonlinear dimensionality reduction techniques could be more suitable to articulated motions [38], which are generally nonlinear. Therefore, LPP algorithm is employed to excavate the inherent structure within the pairwise distance matrix.

Evidently, human motion modeling is a very challenging task because of the ambiguity caused by nonrigid articulations. Recently, BOW model associated with the video descriptors has led to a popular statistical motion framework [43]. However, the traditional BOW model, without temporal information, may suffer from the motion ambiguity and would, therefore, lead to poor matching performance. To the best of the authors' knowledge, the BOW model, with internal temporal constraint, has yet to be exploited for human mocap data analysis. In this section, a TA-BoW is introduced to efficiently characterize the human motion sequence. From a statistical viewpoint, BOW provides a histogram representation by counting the number of codewords existing in the sequence. Often, the histogram, consisting of $\mathbf{w}$ codewords, is created by performing k-means clustering, and the clustering centers can be defined as codewords. Assuming that a motion codebook is learned for each sequence, the motion clip M can be further transformed into the form $\bar{M} = [\bar{\mathbf{d}}_1, \bar{\mathbf{d}}_2, \ldots, \bar{\mathbf{d}}_n]$, where each transformed pose $\bar{\mathbf{d}}_i$ is acquired by replacing the current motion pose $\mathbf{d_i}$ with its nearest codeword in Euclidean space. As a result, the motion clip can be represented concisely by a histogram of codewords: $\mathbf{h}^{\text{BoW}} = [\mathbf{h_1}, \mathbf{h_2}, \ldots, \mathbf{h_w}]$

$$h_i = \sum_{\bar{\mathbf{d}}_j \in \bar{M}} \text{count}\left(L(\bar{\mathbf{d}}_j) = i\right), \quad i = 1, 2, \ldots, \mathbf{w} \tag{5}$$

where $L(\bar{\mathbf{d}}_j)$ returns its nearest codeword, and h_i represents the frequency of each codeword occurring in the sequence.

In essence, the traditional BOW method represents the motion sequence as an orderless collection of local motion features, with no explicit inclusion of temporal correlations. To tackle this problem, a novel TA-BoW, which can improve the distinguishability from traditional BOW, is proposed. The proposed TA-BoW aims not only to count the number of occurrences of each codeword, but also to calculate the number of the codewords that appear consecutively adjacent to the current codeword temporally. By augmenting this temporal relationship, the following matrices are obtained:

1) forward codeword matrix $\mathbf{h^F}$:

$$\mathbf{h^F_{i,t}} = \begin{cases} h_i, & t = i \\ \sum_{\bar{\mathbf{d}}_j \in \bar{M}, j \in F(i)} \text{count}\left(L(\bar{\mathbf{d}}_j) = t\right), & t \neq i \end{cases} \tag{6}$$

2) backward codeword matrix $\mathbf{h^B}$

$$\mathbf{h^B_{t,i}} = \begin{cases} h_i, & t = i \\ \sum_{\bar{\mathbf{d}}_j \in \bar{M}, j \in B(i)} \text{count}\left(L(\bar{\mathbf{d}}_j) = t\right), & t \neq i \end{cases} \tag{7}$$

3) integrated codeword matrix $\mathbf{h^I}$

$$\mathbf{h^I} = \left[\mathbf{h^F}, \mathbf{h^B}\right] \tag{8}$$

where $j \in F(i)$ represents the adjacent neighbors of codeword i in the consecutive forward direction, while $j \in B(i)$ denotes the adjacent neighbors of codeword i in the consecutive backward direction, and $i = 1, 2, \ldots, \mathbf{w}$. A typical example is shown in Fig. 3, in which the motion sequences are mapped into four motion primitives $\{a, b, c, d\}$. The forward and backward codeword matrices are built according to (6) and (7), respectively. It can be seen that $\mathbf{h^F}$ shows the consecutive appearance of the codewords adjacent to the given codeword in the forward direction, and $\mathbf{h^B}$ the consecutive appearance of the codewords adjacent to the given codeword in the backward direction. As a result, the integrated codeword matrix can intrinsically characterize the motion with temporal constraints.

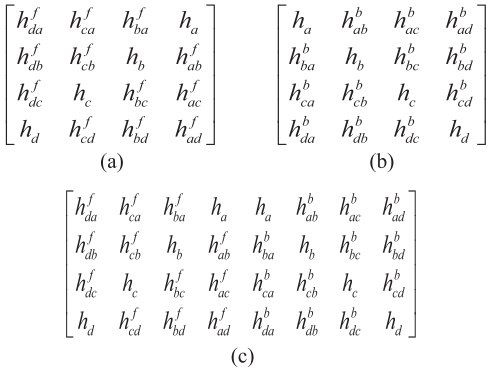


Fig. 3. Examples of the proposed temporal codeword matrices. (a) Forward feature matrix. (b) Backward feature matrix. (c) Integrated feature matrix.

It should be noted that the dimension of histogram has increased from $\mathbf{w}$ to $\mathbf{w} * \mathbf{2w}$, and a histogram with such a very large $\mathbf{w}$ would be computationally expensive. For instance, if the total number $\mathbf{w}$ of codewords is 50, then the dimension of the codeword matrix in (8) will increase to 5000, which is significantly large. In general, temporal relationships exist within the adjacent poses, and the frequency of adjacent codewords is closely related to the current codeword. Inspired by this finding, such TA information is utilized to further simplify the motion histogram. Except for the number of occurrences of each codeword h_i, the maximum number within the forward codeword matrix $\mathbf{h^F}$ and backward codeword matrix $\mathbf{h^B}$ are also counted

$$h_i^{F_m} = \arg\max_t \left(\mathbf{h^F_{i,t}}\right),\ t = 1, 2, \ldots, \mathbf{w} \text{and} \mathbf{t} \neq \mathbf{i} \tag{9}$$

$$h_i^{B_m} = \arg\max_t \left(\mathbf{h^B_{t,i}}\right),\ t = 1, 2, \ldots, \mathbf{w} \text{and} \mathbf{t} \neq \mathbf{i}. \tag{10}$$

By considering the temporal constraint, the numbers of the codewords that appear consecutively adjacent to the current codeword are also computed. Without loss of motion distinguishability, the feature histogram, with internal temporal constraints, can thus be rewritten

$$\mathbf{h}^{\text{TA-BoW}} = \left[h_1^{F_m}, h_1, h_1^{B_m}, \ldots, h_{\mathbf{w}}^{F_m}, h_{\mathbf{w}}, h_{\mathbf{w}}^{B_m}\right]. \tag{11}$$

The feature histogram in (11) is just thrice larger in size than that of the traditional BOW, and the discriminative ability, referred here as TA-BoW model, has greatly improved. Fig. 4 is a typical example of this model, in which eight different motion sequences are mapped into four codewords $\{a, b, c, d\}$. From this figure, it can also be observed that the traditional BOW represents these motion sequences in terms of the same histogram, while the proposed TA-BoW exhibits diverse feature histograms. Evidently, the proposed TA-BoW has more discrimination power and can better characterize these diverse motion sequences.

C. Discriminative Neighborhood Preserving Dictionary Learning

Sparse representation has been demonstrated to be very powerful for classification, and DL is the key issue to the success of that method. In general, the neighboring poses are similar

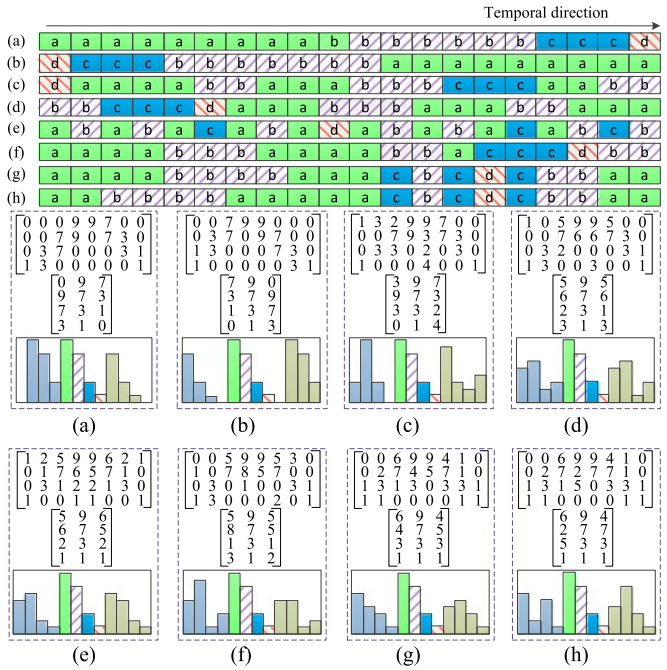


Fig. 4. Illustration of the TA-BoW model. Each sequence is mapped into four codewords $\{a, b, c, d\}$ and transformed into feature matrix by (8), reduced feature matrix and its feature histogram by (11).

because of the high speed of current mocap camera (e.g., 300 fps) and the smoothness of human daily movement. Therefore, the learned dictionary should be able to preserve the same-label neighbors of each training sample, while repelling those belonging to other classes. However, the normal DL methods do not ensure such neighborhood preservation and thus may fail to faithfully depict the intrinsic manifold geometry. In this section, a novel method is proposed to explicitly learn a discriminative dictionary for sparse coding, referred to here as DNP-DL.

1) Neighborhood Relationship Mining: For discriminative learning, it is imperative to preserve the same-label neighbors while inducing a large distance between diverse points of different classes. To achieve this goal, the first step is to find the neighborhood relationships of each sample. In the past, the K-nearest neighbors (KNN) are often selected to search the local geometric structures in training data. However, the fixed neighborhood number may not adaptively reflect the real neighborhood relationships. Also, it cannot guarantee that the identifiability of constructed dictionary basis is discriminative enough. Recently, it has been found that iterative nearest neighbors (INN) [44] algorithm can ensure better similarity grouping adaptively, while keeping the computational similarity with KNN. Heuristically, INN is applied to exploit the neighborhood relationship of each motion clip and simultaneously add this constraint to DL. Given a motion clip $\mathbf{h_i} \in \mathbf{H}$ and its remaining part in the database $\mathbf{H_i^r} = [\mathbf{h_1}, \mathbf{h_2}, \ldots, \mathbf{h_{i-1}}, \mathbf{0}, \mathbf{h_{i+1}}, \ldots, \mathbf{h_N}]$, the optimization objective function of INN can be formulated as

$$\min_{\{c_i\}_{i=1}^{\mathbf{k}}} \mathbf{h_i} - \sum_{i=1}^{\mathbf{k}} c_i \alpha_i, \quad \text{s.t.} \sum_{i=1}^{\mathbf{k}} c_i \approx 1 \tag{12}$$

Algorithm 1: Neighborhood Relationship Mining via INN.

Input: The query motion $\mathbf{h}_\tau \in \mathbf{H}$, $\lambda \in (0,1)$, iterations T, and reconstruction error ε;
Output: Neighborhood relationships $C_\tau = [c_1, c_2, \ldots, c_k]$;
Initialize: $k = 0$, $\acute{\mathbf{q}} = \mathbf{h}_\tau$, $\hat{\mathbf{q}} = 0$, $c_0 = \lambda$;
1. NN search in $\mathbf{H}_\tau^{\mathbf{r}}$, $k = k + 1$:
$\alpha_k = NN\left(\mathbf{H}_\tau^{\mathbf{r}}, \acute{\mathbf{q}}\right)$, $c_k = \frac{c_{k-1}}{(1+\lambda)}$;
2. Approximation and error update:
$\hat{\mathbf{q}} = \hat{\mathbf{q}} + c_k \alpha_k, \quad r = \|\mathbf{h}_\tau - \hat{\mathbf{q}}\|_2$;
3. Adapt the query $\acute{\mathbf{q}} = \acute{\mathbf{q}} + \lambda(\acute{\mathbf{q}} - \alpha_k)$;
4. Repeat steps 1) and 2) until $k > T$ or $r \leq \varepsilon$.

where $\alpha_i \in \mathbf{H_i^r}$, $\mathbf{k}$ represents the number of neighbor elements, and $c_i \in (0,1)$ is the residual weight. It is obvious that the larger the c_i, the more α_i it contributes to the reconstruction of $\mathbf{h_i}$. The main steps of INN algorithm, summarized below under the subhead Algorithm 1, can be well utilized for neighborhood relationship mining. If $C = [C_1, C_2, \ldots, C_N]$ represents the learned neighborhood matrix, the motion relationships can be well exploited through this neighborhood matrix. However, regularizing this neighborhood matrix as a constraint for discriminative DL is still very difficult, because C may be asymmetric. To tackle this problem, the neighborhood matrix is further regularized to be symmetric as $\bar{C} = \{\bar{c}_{ij}\}_{N\times N} = \frac{1}{2}\{c_{ij} + c_{ij}^T\}_{N\times N}$, which can be treated as a weighted neighbor graph with edges.

The discriminative DL encourages the sparse representation of each sample to stay close to its nearest same-label neighbors, and the neighborhood relationship between the sparse coefficients $\{v_i, v_j\}$ can be preserved by minimizing

$$\min \frac{1}{2} \sum_{i,j=1}^{N} \|v_i - v_j\|^2 \bar{c}_{ij}. \tag{13}$$

Alternatively, the objective function in (13) can be relaxed to its equivalent trace form

$$\min \ \psi(V) = \mathrm{tr}\left(V^T\left(\bar{C}^* - \bar{C}\right)V\right) = \mathrm{tr}\left(V^T L V\right) \tag{14}$$

where $\mathrm{tr}(\cdot)$ denotes the trace of a matrix, $\bar{C}^*$ is a diagonal matrix whose entries are the column sum of $\bar{C}$, i.e., $\bar{C}^*_{ii} = \sum_j \bar{c}_{ij}$, and $L = \bar{C}^* - \bar{C}$ is the Laplacian matrix. With this regularization, the nearest same-label neighbors of each motion clip can be well preserved adaptively.

2) Discriminative Regularization in DL: The discriminative DL also encourages that samples from other classes are scarce in the vicinity, and this can be achieved by seeking a Fisher projection to maximize the interclass scatter and minimize the intraclass scatter. To this end, the Fisher discrimination criterion is added as an additional constraint and a DNP-DL approach is proposed to characterize the motion sequence.

If $\mathbf{H} = [\mathbf{H_1}, \mathbf{H_2}, \ldots, \mathbf{H_c}]$ represents the whole training sequence and $\mathbf{H_i}$ all the samples from motion class i, then each subdictionary $\mathbf{D_i}$ of the ith class can be learned separately. As a result, the whole dictionary can be concatenated as $\mathbf{D} = [\mathbf{D_1}, \mathbf{D_2}, \ldots, \mathbf{D_c}]$. Accordingly, the coding coefficient matrix $\mathbf{V}$ of $\mathbf{H}$ over $\mathbf{D}$ can be denoted as $\mathbf{V} = [\mathbf{V_1}, \mathbf{V_2}, \ldots, \mathbf{V_c}]$, and submatrix $\mathbf{V_i}$ contains the coding coefficients over $\mathbf{H_i}$. By imposing both the neighborhood preserving relationships and Fisher discrimination criterion onto the DL, the following optimization problem is obtained

$$J_{\mathbf{D},\mathbf{V}} = \arg\min_{\mathbf{D},\mathbf{V}} \{R(\mathbf{D},\mathbf{V}) + \lambda_1 \|\mathbf{V}\|_1 + \lambda_2 F(\mathbf{V}) + \lambda_3 \psi(\mathbf{V})\} \tag{15}$$

where $R(\mathbf{D},\mathbf{V})$ denotes the reconstruction error that approximates the input data, $\|\mathbf{V}\|_1$ is the regularization term for sparsity, $F(\mathbf{V})$ is the discrimination term regularized by Fisher criterion, $\psi(\mathbf{V})$ is the regularization term of neighborhood relationship preserving, and $\{\lambda_1, \lambda_2, \lambda_3\}$ are the parameters utilized for balancing different items. Further, $R(\mathbf{D},\mathbf{V})$ and $F(\mathbf{V})$ are concretely elaborated as follows:

A. Reconstruction error $R(\mathbf{D},\mathbf{V})$: In sparse representation, the whole dictionary $\mathbf{D}$ is designed to represent the samples of any class $\mathbf{H_i}$; so, $\|\mathbf{H_i} - \mathbf{DV_i}\|_F^2$ has to be as small as possible. Besides, each $\mathbf{H_i}$ should be intensively represented by its corresponding subdictionary $\mathbf{D_i}$ and very sparsely characterized by $\mathbf{D_j}$, $j \neq i$. This implies that minimization of $\|\mathbf{H_i} - \mathbf{D_i V_i^i}\|_F^2$ is essential, and that $\mathbf{V_i^j}$ should have nearly zero coefficients, so that $\sum_{j=1,j\neq i}^{\mathbf{c}} \|\mathbf{D_j V_i^j}\|_F^2$ is small. Hence, $R(\mathbf{D},\mathbf{V})$ should be regularized as

$$R(\mathbf{D},\mathbf{V}) = \sum_{i=1}^{\mathbf{c}} \left(\left\|\mathbf{H_i} - \mathbf{D_i V_i^i}\right\|_F^2 + \|\mathbf{H_i} - \mathbf{DV_i}\|_F^2 + \sum_{j=1,j\neq i}^{\mathbf{c}} \left\|\mathbf{D_j V_i^j}\right\|_F^2 \right). \tag{16}$$

B. Fisher discrimination term $F(\mathbf{V})$: In DL, Fisher criteria can be well applied to regularize coefficient $\mathbf{V}$, so that the samples from different classes are efficiently separated [45]. It aims at minimizing the intraclass scatter $S_W(\mathbf{V})$, while maximizing the interclass scatter $S_B(\mathbf{V})$

$$S_W(\mathbf{V}) = \sum_{i=1}^{\mathbf{c}} \sum_{\mathbf{v_k} \in \mathbf{V_i}} (\mathbf{v_k} - v_i)(\mathbf{v_k} - v_i)^T \tag{17}$$

$$S_B(\mathbf{V}) = \sum_{i=1}^{\mathbf{c}} n_i (v_i - v)(v_i - v)^T \tag{18}$$

where v_i and v are the mean samples of $\mathbf{V_i}$ and $\mathbf{V}$, respectively, and n_i is the number of samples in ith class. Therefore, the discrimination item $F(\mathbf{V})$ can be achieved by

$$F(\mathbf{V}) = \frac{\mathbf{S_W}(\mathbf{V})}{\mathbf{S_B}(\mathbf{V})}. \tag{19}$$

Accordingly, an equivalent form of Fisher criterion that is being imposed onto the DL can be regularized as

$$F(\mathbf{V}) = \mathbf{tr}\left(\mathbf{S_W}(\mathbf{V})\right) - \mathbf{tr}\left(\mathbf{S_B}(\mathbf{V})\right) + \eta \|\mathbf{V}\|_\mathbf{F}^\mathbf{2} \tag{20}$$

where an elastic term $\|\mathbf{V}\|_F^2$, associated with parameter η, is utilized to make the function convex and stable.

C. Proposed DNP-DL model: The proposed DNP-DL model attempts to preserve the neighborhood relationships of intraclass structure, while encouraging discriminability of the interclass variances. The constraints acting on the coding coefficients are formulated as follows:

$$J_{\mathbf{D},\mathbf{V}} = \arg\min_{\mathbf{D},\mathbf{V}} \begin{Bmatrix} \sum_{i=1}^{\mathbf{c}} \left(\|\mathbf{H_i} - \mathbf{D_i}\mathbf{V_i^i}\|_F^2 + \|\mathbf{H_i} - \mathbf{D}\mathbf{V_i}\|_F^2\right. \\ \left. + \sum_{j=1,j\neq i}^{\mathbf{c}} \|\mathbf{D_j}\mathbf{V_i^j}\|_F^2\right) + \lambda_1\|\mathbf{V}\|_{\mathbf{1}} \\ +\lambda_2(tr(S_W(\mathbf{V})) - \mathbf{tr}(\mathbf{S_B}(\mathbf{V}) \\ +\eta\|\mathbf{V}\|_{\mathbf{F}}^{\mathbf{2}}) + \lambda_{\mathbf{3}}\mathbf{tr}\|\mathbf{V^T}\mathbf{L}\mathbf{V}\| \end{Bmatrix}. \tag{21}$$

As a result, the test sample can be well represented by the dictionary atoms of its own class, while dictionary atoms of other classes have little contribution to its reconstruction.

3) Optimization in DNP-DL: The optimization in (21) is not jointly convex to $(\mathbf{D}, \mathbf{V})$, but is individually convex with respect to each of them. Therefore, the solution of (21) can be solved alternatively by optimizing $\mathbf{D}$ and $\mathbf{V}$ [45].

Updating coding coefficient matrix $\mathbf{V}$*:* If the whole dictionary $\mathbf{D}$ is fixed, then the objective function in (21) is reduced to a sparse coding problem

$$J_{\mathbf{V_i}} = \arg\min_{\mathbf{V_i}} \begin{Bmatrix} \left\|\mathbf{H_i} - \mathbf{D_i}\mathbf{V_i^i}\right\|_F^2 + \|\mathbf{H_i} - \mathbf{D}\mathbf{V_i}\|_F^2 \\ + \sum_{j=1,j\neq i}^{\mathbf{c}} \|\mathbf{D_j}\mathbf{V_i^j}\|_F^2 + \lambda_1\|\mathbf{V_i}\|_1 \\ +\lambda_2 F(\mathbf{V_i}) + \lambda_3 tr\left(\mathbf{V_i^T}\mathbf{L}\mathbf{V_i}\right) \end{Bmatrix} \tag{22}$$

where $F(\mathbf{V_i}) = \left\|\mathbf{V_i} - \overline{\mathbf{V_i}}\right\|_F^2 - \sum_{k=1}^{\mathbf{c}} \left\|\overline{\mathbf{V_k}} - V\right\|_F^2 + \eta\|\mathbf{V_i}\|_F^2$, $\overline{\mathbf{V_k}}$ and V are matrices composed of the mean vectors of kth class and all classes, respectively. Then, each $\mathbf{V_i}$ can be updated one by one while all other $\mathbf{V_j}$ $(j \neq i)$s are fixed.

Updating subdictionaries $\mathbf{D_i}$*:* Each $\mathbf{D_i}$ can be updated by fixing coefficient matrix $\mathbf{V}$ and all the other $\mathbf{D_j}$ $(j \neq i)$s. Thus, the objective function in (21) is reduced as

$$J_{\mathbf{D_i}} = \arg\min_{\mathbf{D_i}} \begin{Bmatrix} \left\|\mathbf{H} - \mathbf{D_i}\mathbf{V^i} - \sum_{\mathbf{j}=\mathbf{1},\mathbf{j}\neq\mathbf{i}}^{\mathbf{c}} \mathbf{D_j}\mathbf{V^j}\right\|_F^2 + \\ \left\|\mathbf{H_i} - \mathbf{D_i}\mathbf{V_i^i}\right\|_F^2 + \sum_{j=1,j\neq i}^{\mathbf{c}} \left\|\mathbf{D_i}\mathbf{V_j^i}\right\|_F^2 \end{Bmatrix} \tag{23}$$

where $\mathbf{V^i}$ is the coding coefficient of $\mathbf{H}$ over subdictionary $\mathbf{D_i}$. The quadratic programming problem in (23) can be solved by updating $\mathbf{D_i}$, atom by atom [46]. Once the dictionary $\mathbf{D}$ is initialized, the optimizations of (22) and (23) can be obtained by iteratively repeating the above process until reaching a stopping criterion. The main procedures of optimization are summarized below under Algorithm 2.

4) Complexity of the DNP-DL: In DNP-DL framework, the time complexity of updating coding coefficients is approximately $\sum_i n_i O(p_i q^2)$, where n_i and p_i, respectively, denote the number of training samples and dictionary atoms in the ith class, and q is the feature dimensionality of each training example. Time complexity of updating dictionary atom is approximately $\sum_i p_i O(n_i q)$. As a consequence, the overall time complexity of DNP-DL is $\sum_i n_i O(p_i q^2) + \sum_i p_i O(n_i q)$. Since $n = \sum_i n_i$ and $p = \sum_i p_i$, respectively, represent the total number of training samples and dictionary atoms, the whole time complexity of DNP-DL is acceptable in relation to the traditional sparse coding problem [27].

Algorithm 2: Optimization solution to DNP-DL.

1. Initializing dictionary $\mathbf{D}$: initialize each $\mathbf{D_i}$ as a random vector with unit ℓ_2 norm.
2. Updating the sparse coding coefficients $\mathbf{V}$: fix dictionary $\mathbf{D}$ and calculate $\mathbf{V_i}$, one by one via (22).
3. Updating dictionary $\mathbf{D}$: fix $\mathbf{V}$ and update each $\mathbf{D_i}$, class by class via (23), and atom by atom [46].
4. Checking the values of $J_{\mathbf{D},\mathbf{V}}$ between two iterations; if the difference between them is small enough or if the maximum iteration has reached to its maximum, output $\mathbf{V}$ and $\mathbf{D}$; otherwise, return to step 2, continue.

D. Hierarchical Motion Retrieval Mechanism

In DNP-DL, the learned dictionary will be discriminative enough to sparsely represent the motion sequence. In the past, most of the motion retrieval approaches often chose to compare the query motion clip with all the training samples, which will be quite time-consuming when the training database there is too large. To tackle this problem, a hierarchical retrieval mechanism is exploited to facilitate a coarse-to-fine similarity matching, which involves interclass recognition by sparse classification and intraclass ranking by chi-square measurement. The main steps are summarized as follows.

1) Calculate sparse coding coefficients $\mathbf{v} = [\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_c}]$ of query motion sample $\mathbf{h}_q$ via learned dictionary $\mathbf{D}$

$$\mathbf{v} = \arg\min_{\mathbf{v}} \|\mathbf{h_q} - \mathbf{D}\mathbf{v}\|_{\mathbf{2}}^{\mathbf{2}} + \tau\|\mathbf{v}\|_{\mathbf{1}} \tag{24}$$

where $\mathbf{D}$ is obtained by DNP-DL, $\mathbf{v_i}$ is the coefficient vector over subdictionary $\mathbf{D_i}$, and τ is a scalar constant.

2) Estimate the reconstruction error of the ith class

$$e_i = \|\mathbf{h_q} - \mathbf{D_i}\mathbf{v_i}\|_2^2 + \omega\|\mathbf{v} - \mathbf{v_i}\|_2^2 \tag{25}$$

where v_i is the mean coefficient of ith class, and ω is the weight to balance the contribution of the two terms.

3) Find the label of smallest reconstruction error

$$\text{class}(i) = \arg\min_i \{e_i\},\ i = 1, 2, \ldots, \mathbf{c}. \tag{26}$$

4) Once the motion class c is determined, searching for most of the other similar motion clips can be further ranked by a well-known chi-square distance χ^2. Given two TA-BoW histograms $\mathbf{h_q} = [h_1^q, h_2^q, \cdots, h_{\mathbf{3w}}^q]$ and $\mathbf{h_c} = [h_1^c, h_2^c, \cdots, h_{\mathbf{3w}}^c]$, χ^2 distance can be formulated as

$$\chi^2(\mathbf{h_q}, \mathbf{h_c}) = \frac{1}{2}\sum_{t=1}^{\mathbf{3w}} \frac{(h_t^q - h_t^c)^2}{h_t^q + h_t^c}. \tag{27}$$

TABLE I
DIVIDED ACTION SUBSETS OF MSR-ACTION3D DATABASE

ActionSet1 (AS1)	ActionSet2 (AS2)	ActionSet3 (AS3)
Horizontal arm wave	High arm wave	High throw
Hammer	Hand catch	Forward kick
Forward punch	Draw cross	Side-kick
High throw	Draw tick	Jogging
Hand clap	Draw circle	Tennis swing
Bend	Two-hand wave	Tennis serve
Tennis serve	Side-boxing	Golf swing
Pick-up and throw	Forward kick	Pick-up and throw

IV. EXPERIMENTAL RESULTS

For experimental evaluation, two publicly available datasets are selected: MSR-Action3D [17] and HDM05 library [39]. There are two main differences between these datasets: 1) The MSR-Action3D dataset, acquired by a depth camera of Kinect sensor, often incorporates more noise than the HDM05 dataset, acquired by a multicamera motion capturing system; 2) The frame rate of HDM05 dataset is much higher (120 fps) than that of the MSR-Action3D dataset (15 fps).

A. Experimental Setup and Parameter Tuning

MSR-Action3D dataset is selected to evaluate the proposed motion descriptor on action recognition task because of its limited size. In contrast to this, HDM05 library provides a large number of motion sequences, which will be used extensively to validate the retrieval performance. All the experiments were conducted on an Intel Core$^{\mathrm{TM}}$ i3 3.30 GHz processor with 8-GB memory, implementing the coding language with MATLAB. In LPP algorithm, the reduced dimension was automatically obtained by preserving 99% energy, and the neighboring points were empirically set at 30 and the heat kernel function at 3.7. In general, the vocabulary size $\mathbf{w}$ often balances the tradeoff between discrimination and complexity. Following [38], several sizes, $\mathbf{w} = \{\mathbf{10}, \mathbf{30}, \mathbf{50}, \mathbf{70}, \mathbf{90}, \mathbf{110}, \mathbf{150}\}$, were tested and the best performance was selected. Similar to the parameters suggested in the literature [44], [45], several parameters in INN algorithm were set as $\lambda = 0.5$, $\varepsilon = 10e - 5$, and those in DNP-DL as $\eta = 1, \lambda_1 = 0.005, \lambda_2 = 0.05$ and $\lambda_3 = 0.05$. In addition, the parameters in sparse coding were chosen as $\tau = 0.001$ and $\omega = 0.05$, empirically.

B. Datasets From MSR-Action3D Database

The public MSR-Action3D database [17] consists of ten subjects, performing 20 actions, with up to three repetitions. This led to a total of 567 sequences, and only skeleton joints were employed in the experiments.

For a fair comparison, the same experimental settings as those employed in work [40], were utilized, and the data divided into three action sets (i.e., AS1, AS2, AS3), as shown in Table I. For each action subset, cross-subject test [17] was conducted. That is, half of the subjects (1, 3, 5, 7, 9) were used for training and the other half for testing. Similarly, 20 skeleton joints [47] were selected to shape the posture, and the pairwise joint distances computed to characterize the human movements. Since the size of MSR-Action3D database is limited, it is improper to train a DNP-DL in each action subset. Therefore, SVM, associated with histogram intersection kernel (H), was utilized to perform action recognition, and the proposed motion descriptor was compared with some typical skeleton-based representations, i.e., a bag of 3-D points [17], histogram of 3-D joints [18], eigenjoints [8], skeletal quads [16], covariance of 3-D Joints (Cov3DJ) [25], histograms of part sets [19], HOD (16 bins) with two-level temporal pyramid (TP) [15], pose-based TP [37], Hanklet-based HMM [20], and BoW with SVM (H) [38].

TABLE II
RECOGNITION ACCURACY, TESTED ON MSR-ACTION3D DATASET

Method (cross-subject test)	Recognition accuracy			
	AS1	AS2	AS3	Overall
A bag of 3-D points (Li *et al.* [17])	72.9	71.9	79.2	74.67
Histogram of 3-D Joints (Xia *et al.* [18])	87.98	85.48	63.46	78.97
Eigenjoints (Yang *et al.* [8])	74.5	76.1	96.1	82.33
Skeletal Quads (Evangelidis *et al.* [16])	88.39	86.61	94.59	89.86
Cov3DJ descriptor (Hussein *et al.* [25])	88.04	89.29	94.29	90.53
Histograms of part sets (Wang *et al.* [19])	–	–	–	90.22
HOD+2-level TP (Gowayyed *et al.* [15])	92.39	90.18	91.43	91.26
Pose-based TP (Eweiwi *et al.* [37])	–	–	–	90.1
BOW+SVM (H) (Fotiadou *et al.* [38])	87.03	81.02	86.08	84.71
Hanklet-based HMM (Presti *et al.* [20])	–	–	–	89
TA-BoW+SVM (H) (Ours)	92.36	90.68	91.06	91.37

The recognition accuracies obtained on different representations were shown in Table II. It is to be noted that some subset results were not reported in [19], [20] and, [37], and most of the recognition accuracies obtained by the previous motion representations [8], [16]–[18], [38] are less than 90%. This is probably because actions in AS1 and AS2 subsets share almost similar movements, while those in AS3 subset are complex, but distinct. Evidently, similar motions of diverse semantics are difficult to identify; for example, "Hammer" tends to be confused with "Forward Punch" in an AS1 subset. In addition, the 3-D skeleton joints, acquired from depth maps, often incorporate the view variance, besides being very noisy. Under such circumstances, the sampling scheme of method [17], which depends on the same view, may fail to identify some similar motions, while the Eigenjoints [8] and skeletal quads [16] are very sensitive to noise. Specifically, Xia *et al.* [18] and Presti *et al.* [20] selected the discrete HMM to model temporal motion evolutions, which may not fully characterize the motion temporality; besides, its recognition performances is rather poor. Another plausible reason is that the complex actions adversely affect HMM identification when the number of training samples is small. Although the histograms of part sets [19], covariance of 3-D joints [25], pose-based TP [37], and the BOW method [38] could well characterize the motion sequence, those descriptors without internal temporal regularization may fail to identify some complex motions, and hence their performances are not competitive in practice.

Comparatively speaking, the motion descriptor proposed here has achieved promising results in all the action subsets and simultaneously obtained the best overall recognition accuracy.

TABLE III
DIVIDED SUBSETS IN HDM05 DATASET

Subsets	Training set	Testing set
A	Two-third of the samples in each class	Remaining part
B	Half of the samples in each class	Remaining part
C	One-third of the samples in each class	Remaining part

For instance, the recognition accuracy is higher than 90% in all the subsets. That is, the proposed TA-BoW is discriminative enough to identify diverse human actions, including those with subtle differences, and is more robust against noise. Also, the proposed motion descriptor has yielded a very competitive performance with method [15]. This supremacy can be attributed to the physical meaning of the descriptor illustrated in Section III-A, which contains significant spatiotemporal information about the sequence. In contrast to this, the approach in [15] utilized three orthogonal cartesian planes to describe the 3-D trajectories of each body joint, and applied a TP approach to capture the temporal evolution of motions. Therefore, the TA-BoW proposed here is inherently more discriminative to encode both spatial and temporal information within the motion sequences. Further, as suggested in [37], [48], and [49], a total of 20 actions were employed, considering all the 252 combinations by choosing half of the actors for training and the other half for testing. This setting is generally considered more challenging than the former setting because it involves more action classes. Fortunately, the average accuracy obtained by the proposed motion descriptor is 74.7±3.1 (mean value and standard deviation), while that of the BOW descriptor [38] is 61.8 ± 3.9. This indicates that the proposed TA-BoW, with internal temporal constraint, has higher discriminatory power, and its competing performance is demonstrated by the experiments.

C. Datasets From HDM05 Database

The popular HDM05 MoCap dataset consists of 130 motion classes with multiple trials, and five different actors enrolled in each class. For evaluation, 339 motion clips were manually collected from 10 different human actions. In total, more than 82 298 frames were scattered into ten categories: Clapping, Elbow to Knee, Hopping, JumpJack, Kicking, Punching, Skiering, Squatting, TurnLeft, Walking. As described in Table III, the collected motion sequences were divided into three subsets and then various experiments conducted.

Earlier, most dimension reduction methods have often been failing to exploit the motion temporality (e.g., SVD [4], PCA [5], and manifold learning [7]), while some typical motion alignment methods have inherently been constrained by computational complexity, requiring large runtime (e.g., CTW [34] and CoTW [35]). Therefore, it is very difficult to perform a fair and meaningful comparison with those approaches. To evaluate motion retrieval performance, five competing approaches were selected for comparison, i.e., DTW [31], derivative DTW (DDTW) [32], ISOCCA [33], BOW [38], and STPM [30]. The main goal of DTW, DDTW, and ISOCCA approaches is to search for a globally optimal path that can map the domain of the query motion sequence onto the indexing sequences. Specifically, the motion dimensionality conducted by these three approaches was first reduced to 15, using PCA, while keeping 99% of the energy. The BOW method first builds a group of BOW-based histogram and then utilizes SVM classifier for motion classification, whereby similar motions can be retrieved by ranking the motion sequences in the corresponding motion class. The recent STPM method first introduces a temporal sparse representation to represent the motion sequence, and then applies a spatial TP matching scheme to perform similarity measurement. In this approach, the gap value of frame index is set at 10 and the level of total pyramid at 3. Meanwhile, other parameters suggested by the exiting works are selected in all experiments.

TABLE IV
RECOGNITION RATES OBTAINED BY DIFFERENT CLASSIFIERS, EACH TABLE CELL SHOWS THE MEAN VALUE AND STANDARD DEVIATION

		Classifiers (%)				
Subsets	Descriptor	KNN+E	KNN+C	SVM+L	SVM+H	OUR
A	BOW	81.1 ± 2.5	86.6 ± 2.7	89.3 ± 2.3	86.6 ± 2.7	89.3 ± 2.3
	TA-BoW	87.5 ± 2.2	89.3 ± 2.3	94.6 ± 2.0	97.3 ± 1.7	97.3 ± 1.7
B	BOW	74.4 ± 4.2	82.1 ± 3.1	87.5 ± 2.9	87.5 ± 2.9	87.5 ± 2.9
	TA-BoW	80.9 ± 2.7	85.7 ± 2.4	97.0 ± 2.1	97.0 ± 2.1	97.0 ± 2.1
C	BOW	78.5 ± 3.3	80.6 ± 3.7	86.3 ± 2.9	76.2 ± 3.3	88.5 ± 2.7
	TA-BoW	86.3 ± 2.9	84.1 ± 2.9	88.5 ± 2.7	89.9 ± 2.5	93.4 ± 2.3

1) Performance Analysis: First, an extensive comparison was attempted between the traditional BOW representation and the proposed TA-BoW descriptor under different classifiers. Specifically, Euclidean (E) and Cosine (C) distances were selected for KNN classifier, and linear kernel (L) and histogram intersection kernel (H) for SVM classifier. The mean recognition rates obtained from five runs are shown in Table IV. Under the same classifier, it can be found that the mean recognition rates obtained by BOW representation are all less than 90%. This descriptor provided little information about temporal information, consequent to which some ambiguous motions might not have been well characterized and a certain ratio of motion clips might have been misidentified. In contrast, the proposed TA-BoW descriptor has achieved a higher recognition rate under different classifiers. For instance, the mean recognition rate of Subset B, tested with the TA-BoW descriptor and "SVM+L" classifier, reached up to 97%, which is significantly higher than that achieved by the BOW counterpart. It indicates that the proposed TA-BoW descriptor, incorporating temporal constraint, serves better for complex motion representation. Also, the proposed DNP-NL classifier has achieved a higher and more stable recognition rate (i.e., less standard deviations) under the same representations.

The proposed approach is further compared with the existing methods by computing the confusion matrix, which is a specific table layout that is utilized to visualize classification performance. Typical results obtained by performing the test on subset B are shown in Fig. 5. It can be observed that the DTW and DDTW approaches yielded lower accuracies, and some motion sequences were falsely recognized as other semantics,

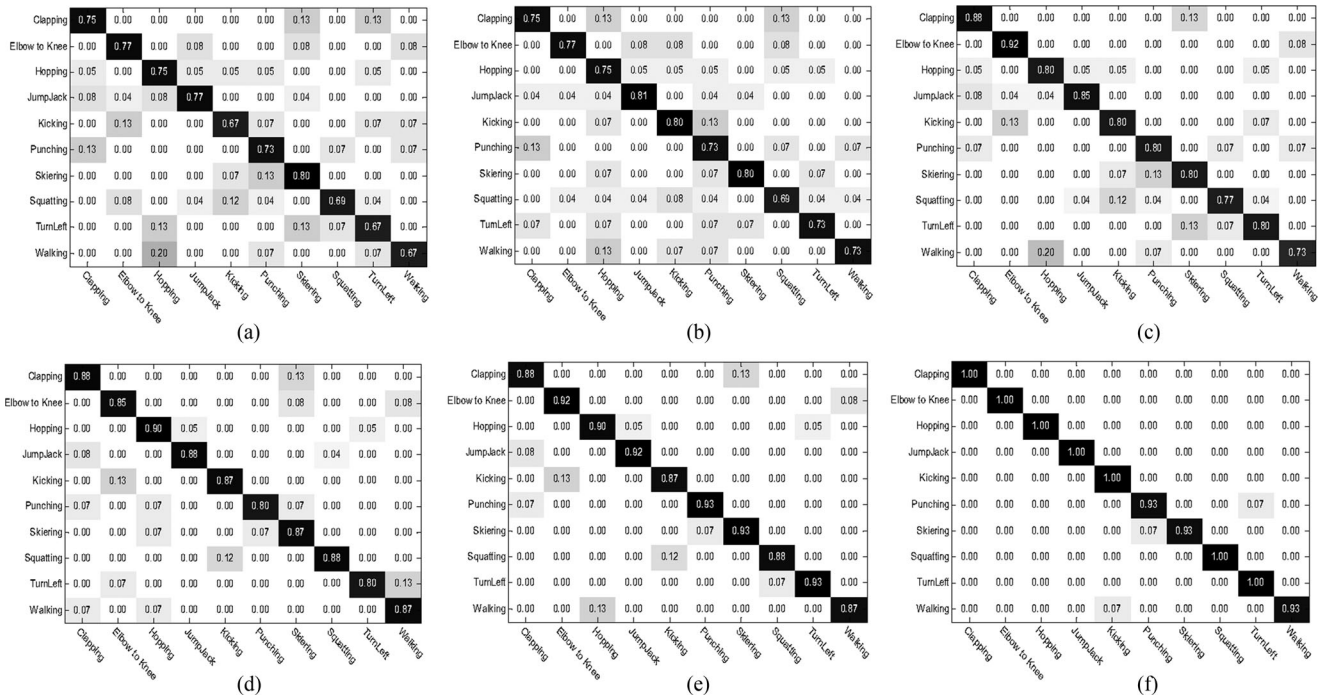


Fig. 5. Confusion matrices obtained by different approaches and tested within subset B. (a) DTW. (b) DDTW. (c) ISOCCA. (d) BoW. (e) STPM. (f) OURS.

i.e., *Walking* and *Kicking*. These two approaches were used to compute the similarity between two motion clips by aligning their time dimensions. The concerned motion clips would be generally identified as being similar if their motion sequences commenced with the same poses. But, human motions often show large spatiotemporal variations between different executions of the same action, and thus the motion clips with similar semantic meanings are usually not aligned temporally. Therefore, logically similar motions may not strictly be temporally similar or start with the same poses. For example, two walking motions, one with stepping the right leg first and the other with stepping the left leg first, would be confused, in two different classes. ISOCCA approach similarly performed similarity matching by linearly mapping two sequences into a common subspace, through which the property of nondecreasing monotonicity can be preserved in time. Although its classification performance has been improved to some degree, such an approach depends highly on a consistent periodic or aperiodic motion period. Otherwise, some logically similar motions may be confused, e.g., *Squatting* and *Walking*.

Instead of time-alignment, BOW simplified the motion representation, in terms of the histogram, and was able to adapt to the slightly misaligned motions. But, such a method represented the motion sequence as an orderless collection of primitive poses, which often failed to identify some complex motion sequences, e.g., "Elbow to Knee." By considering the temporal relationship within the motion sequence, the STPM approach could efficiently identify some complex motion sequences. Unfortunately, even this approach may fail to distinguish between some confusable motions. For instance, the classification rates of "Clapping" and "Kicking" sequences are lower than 90%. The main reasons for failure are twofold: First, the STPM approach utilized the normalized joints to directly represent the motion sequence, which may not well characterize the inherent motion semantics. The articulated complexity within the motions may therefore result in identification confusion; Second, the dictionaries learned directly from concatenated motion frames, often induce sparse ambiguity, which degrades its classification accuracy because of its limited discrimination power.

TABLE V
MEAN CLASSIFICATION RATES OBTAINED BY DIFFERENT METHODS, EACH TABLE CELL SHOWS THE MEAN VALUE AND STANDARD DEVIATION

	Methods (%)					
Subsets	DTW	DDTW	ISOCCA	BOW	STPM_TSR	OURS
A	72.3 ± 3.9	79.6 ± 3.8	83.6 ± 3.6	86.6 ± 2.7	91.3 ± 2.1	97.3 ± 1.7
B	61.6 ± 4.5	75.6 ± 4.1	80.9 ± 3.9	87.5 ± 2.9	90.5 ± 2.4	97.0 ± 2.1
C	62.5 ± 4.7	69.7 ± 4.5	72.8 ± 4.3	76.2 ± 3.3	85.1 ± 2.9	93.4 ± 2.3

In contrast to this, the proposed approach has achieved, notwithstanding the diversity in motion semantics, the best classification performance and correct identification of most of the motion clips, e.g., "Clapping" and "Kicking" sequences. Further, instances of identification confusion are significantly fewer than those of the other five competing approaches. For instance, "Hopping" was slightly confused with *Kicking*, and only 7% of the "Walking" sequences were falsely classified.

Next, the mean classification rates within all the tested motions were computed and the results presented in Table V. It can be seen that DTW, DDTW, and ISOCCA methods delivered low classification rates, especially when the training samples were very small. For instance, the mean classification rates, obtained by these three approaches, are all less than 75% for Subset C.

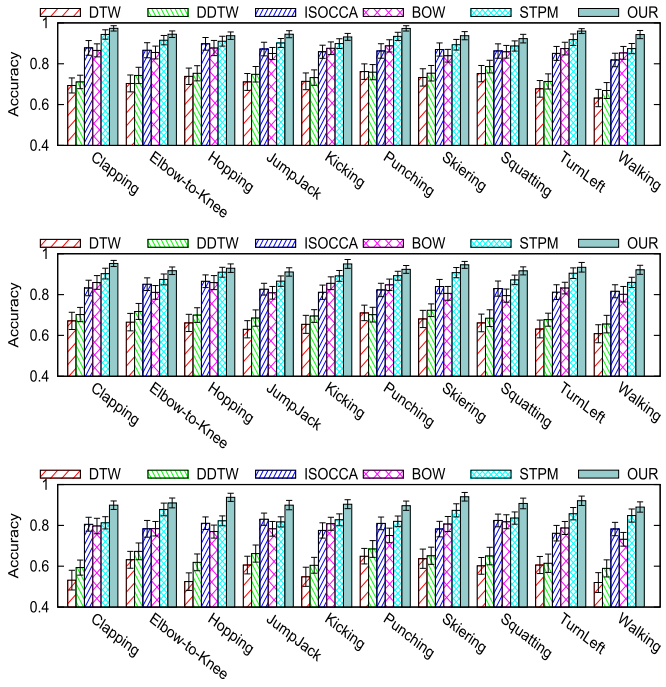


Fig. 6. Average precisions with standard deviation bars, obtain by using different approaches: Top: Subset A; Middle: Subset B; and Bottom: Subset C.

As regards the BOW and STPM approaches, they could, no doubt, recognize most of the motion clips, but some of the identified results deviated a little from the ground truth because of articulated complexity and motion ambiguity. Comparatively speaking, the proposed approach could more successfully recognize different motion semantics appropriately, and the mean classification rates are very competitive in practice. Besides, more stable performances, in terms of low standard deviation values, were achieved. For example, 97.3% of all the testing motion clips of Subset A were correctly identified, and the standard deviation was only around 1.7%. The main advantages of this high classification rate, achieved by the proposed approach, are twofold: First, the proposed TA-BoW exploits the internal temporal relationships among the codewords in such a way that the spatiotemporal property within the motion is well characterized. As a consequence, the whole motion sequence is well represented and the articulated complexity is greatly reduced; Second, the dictionaries learned from DNP-DL can well preserve the same-label neighbors of each training sample, while repelling those belonging to the other classes. As a result, the motion clip can be discriminatively characterized by a more compact set of dictionary atoms.

The retrieval accuracy was further evaluated by calculating precision and recall. For present experiment, the motion sequences, selected from the top two identified categories, were ranked. The average precisions with standard deviation bars, and the average precision-recall curves are shown in Figs. 6 and 7, respectively. From these figures, it can be seen that the proposed approach has achieved the best retrieval performance, in terms of higher average precisions and better precision-recalls. This has become possible for three reasons:

1) The proposed TA-BoW model aims at not only counting the number of codeword occurrences, but also at calculating the number of the codeword that appears consecutively adjacent to the current codeword temporally. As a result, the histogram obtained would be more interpretable, intuitive, and semantically valid for motion representation;
2) The reconstruction error and sparsity inducing penalty of DL are minimized alongside a neighborhood relationship preserving item. Accordingly, each tested motion clip can be sparsely represented by a linear combination of such discriminative dictionary atoms.
3) The presented hierarchical retrieval mechanism, incorporating sparse classification scheme and chi-square ranking, can facilitate coarse-to-fine similarity matching, excluding thereby some motion clips that are likely to be confused with other semantics. Therefore, the proposed approach is particularly suitable for complex motion retrieval in real-world applications.

2) Discussion and Analysis: It is worth noting that, in the clustering process, the vocabulary size of the codebook $\mathbf{w}$ is controlled by the number of key frame clusters, and that a tradeoff is required between discriminability and generalizability to arrive at an appropriate size of vocabulary. A compact codebook with very few entries will have limited discrimination power, while a large codebook size may induce complexity and ambiguity sparsely. To the best of authors' knowledge, there is, as yet, no consensus on what should be considered the right vocabulary size. Several vocabulary sizes, varying from 10 to 150, were tested by the present authors, and the experimental results and fitting curves, illustrating the impact of $\mathbf{w}$ on recognition accuracy are shown in Fig. 8. From this figure, it can be seen that, with increase in vocabulary size, the recognition accuracy first increased moderately, then peaked at a point, and finally decreased thereafter. This suggests that, when the vocabulary size is small, the codewords will not be comprehensive, but when the vocabulary size becomes very large, sparse ambiguity would result.

Next, the influence of different parameter values in (22) on mean recognition accuracy was evaluated. In Section III-C, it has been clearly shown that neighborhood relationship-preserving item encourages the sparse representation of each sample to stay close to its nearest same-label neighbors, while Fisher discrimination minimizes the intraclass scatter and maximizes the interclass scatter. Evidently, these two items are so correlated that it is reasonable to assign them the same weights in (22). Therefore, the scope of evaluation was confined to assessing the influence of the values between λ_1 and pair (λ_2, λ_3). Following the values suggested in [45], the mean recognition accuracies were tested with different balancing values $\lambda_1 = \{0.001, 0.002, 0.005, 0.007, 0.01\}$ and $(\lambda_2, \lambda_3) = \{0.01, 0.02, 0.05, 0.07, 0.1\}$, and the results are shown in Fig. 9. From this figure, it can be seen that different settings of the balancing values in (22) do affect the mean recognition accuracy, but not much differently. Experimentally, the suggested values often delivered better performances.

Further, the storage space for N motion clips can be expressed as $S = \sum_{i=1}^{N} n_i \times 3J$, where n_i denotes the total frame number of the ith clip. In contrast to that the proposed approach represents each motion clip as a TA-BoW vector and the storage space becomes $\widehat{S} = N \times 3\mathbf{w}$, where $\mathbf{w}$ denotes the number of

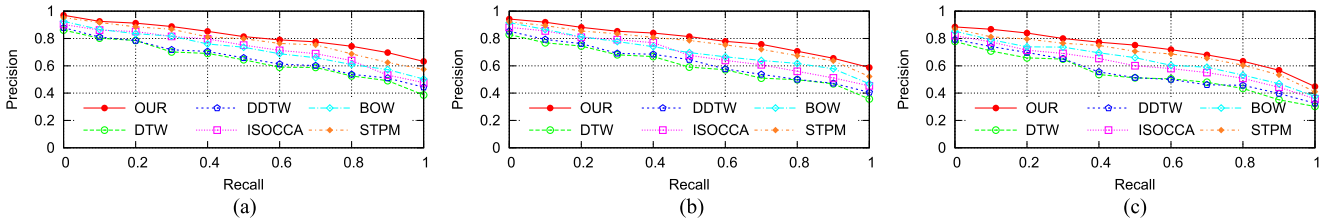


Fig. 7. Average precision-recall curves conducted on different subsets and obtained by different approaches. (a) Subset A. (b) Subset B. (c) Subset C.

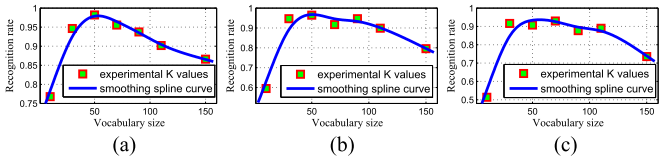


Fig. 8. Influence of vocabulary size on recognition accuracy. (a) Subset A. (b) Subset B. (c) Subset C.

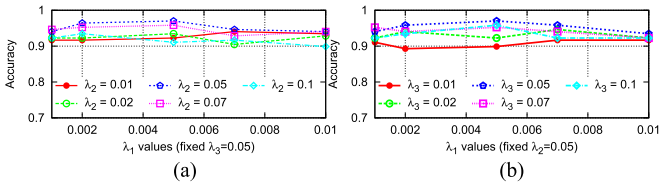


Fig. 9. Influence of $\{\lambda_1, \lambda_2, \lambda_3\}$ on mean recognition accuracy. (a) Influence of λ_1 and λ_2. (b) Influence of λ_1 and λ_3.

codewords. Since the total number of motion frames is significant greater than the number of motion clips, and the vocabulary size is comparable to the joint number, the proposed TA-BoW model can greatly reduce the storage space in comparison with that of the frame level case.

Moreover, although the time-series matching methods, e.g., DTW, DDTW, and ISOCCA, do not need the training phase, the computational costs of these approaches are significantly higher and all their matching times exceed 5 h. This is mainly because, these approaches require dynamic programming algorithm to temporally align the motion sequences and to compare each motion clip, one by one, within the whole database. All this evidently requires more runtime. Since the dimensionality of the proposed TA-BoW is higher than that of BOW, and the proposed DNP-DL method shares more discrimination items in the learning phase, the computation time of the proposed method would be much more than that of other methods. Fortunately, the runtime of the proposed method is acceptable, in that its execution time is around 19.5 m, in contrast to 18.1 and 35.5 m, respectively, of BOW and STPM. Except for the simplest BOW method, which does not require DL process, the execution time of the proposed method is less than that of the STPM method. The main advantage of the proposed method is that its hierarchical retrieval mechanism facilitates coarse-to-fine similarity matching, which serves to significantly reduce the searching ranges. Although a discriminative DL involves more computational load, it needs no complex time-series matching process. Consequently, the computational load of the proposed approach is comparable to that of the competing methods, and the proposed retrieval scheme would be more suitable for indexing large motion databases.

V. Conclusion

This paper presented a novel framework that allows for flexible and efficient human motion retrieval from mocap data. In our approach, the original skeleton model is first simplified and then a novel TA bag of words is used to characterize the motion appearances, through which the articulated complexity can be greatly reduced. In addition, an improved DNP-DL framework is presented. The framework sparsely represents the tested motion, in which the reconstruction error and sparsity inducing penalty are minimized alongside a neighborhood relationship preserving item and a discriminative item. Moreover, a hierarchical retrieval mechanism is addressed to facilitate coarse-to-fine similarity matching. Without complex time-alignment, the proposed retrieval approach can perform well on different kinds of motion semantics, and the experimental results have demonstrated its outstanding performance.

Further research is warranted along the present lines of work in order to solve several problems. For example, if new motions of other semantics are incorporated into the training database, then the proposed method will have to learn the dictionaries again, which would be time-consuming. Therefore, it would be necessary to extend the algorithm, so that it can handle the new motion data adaptively. In addition, questions like how to precisely determine the vocabulary size of a codebook, and how to efficiently fuse with other types of motion descriptors, e.g., HOD, are yet to be solved.

References

[1] L. Kovar and M. Gleicher, "Automated extraction and parameterization of motions in large data sets," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 559–568, 2004.

[2] L. L. Presti and M. La Cascia, "3D skeleton-based human action classification: A survey," *Pattern Recognit.*, vol. 53, pp. 130–147, 2016.

[3] C. Li, S. Q. Zheng, and B. Prabhakaran, "Segmentation and recognition of motion streams by similarity search," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 3, no. 3, 2007, Art. no. 16.

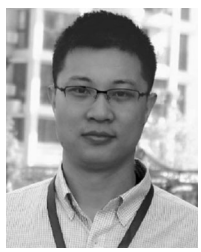
[4] G. N. Pradhan and B. Prabhakaran, "Indexing 3-D human motion repositories for content-based retrieval," *IEEE Trans. Inf. Technol. Biomed.*, vol. 13, no. 5, pp. 802–809, Sep. 2009.

[5] J. Barbic, A. Safonova, J.-Y. Pan, C. Faloutsos, J. K. Hodgins, and N. S. Pollard, "Segmenting motion capture data into distinct behaviors," in *Proc. Graph. Interface*, 2004, pp. 185–194.

[6] K. Forbes and E. Fiume, "An efficient search algorithm for motion data using weighted PCA," in *Proc. ACM SIGGRAPH/Eurograph. Symp. Comput. Animation*, 2005, pp. 67–76.

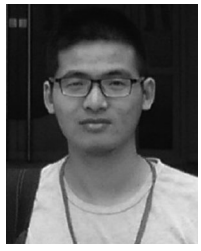
[7] X. Guo, Q. Zhang, R. Liu, D. Zhou, and J. Dong, "3D human motion retrieval based on ISOMAP dimension reduction," in *Proc. Int. Conf. Artif. Intell. Comput. Intell.*, 2011, pp. 159–169.

[8] X. Yang and Y. L. Tian, "Eigenjoints-based action recognition using naive-Bayes-nearest-neighbor," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. Workshops*, 2012, pp. 14–19.
[9] P. Wang, R. W. Lau, Z. Pan, J. Wang, and H. Song, "An eigen-based motion retrieval method for real-time animation," *Comput. Graph.*, vol. 38, pp. 255–267, 2014.
[10] M. Müller, T. Röder, and M. Clausen, "Efficient content-based retrieval of motion capture data," *ACM Trans. Graph.*, vol. 24, no. 3, pp. 677–685, 2005.
[11] M. Müller and T. Röder, "Motion templates for automatic classification and retrieval of motion capture data," in *Proc. ACM SIGGRAPH/Eurograph. Symp. Comput. Animat.*, 2006, pp. 137–146.
[12] M. Raptis, D. Kirovski, and H. Hoppe, "Real-time classification of dance gestures from skeleton animation," in *Proc. ACM SIGGRAPH/Eurograph. Symp. Comput. Animat.*, 2011, pp. 147–156.
[13] A. W. Vieira, T. Lewiner, W. R. Schwartz, and M. Campos, "Distance matrices as invariant features for classifying MoCap data," in *Proc. IEEE Int. Conf. Pattern Recognit.*, 2012, pp. 2934–2937.
[14] T. Huang, H. Liu, and G. Ding, "Motion retrieval based on kinetic features in large motion database," in *Proc. ACM Int. Conf. Multimodal Interact.*, 2012, pp. 209–216.
[15] M. A. Gowayyed, M. Torki, M. E. Hussein, and M. El-Saban, "Histogram of oriented displacements (HOD): Describing trajectories of human joints for action recognition," in *Proc. IEEE Int. Joint Conf. Artif. Intell.*, 2013, pp. 1351–1357.
[16] G. Evangelidis, G. Singh, and R. Horaud, "Skeletal quads: Human action recognition using joint quadruples," in *Proc. IEEE Int. Conf. Pattern Recognit.*, 2014, pp. 4513–4518.
[17] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3dD points," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. Workshops*, 2010, pp. 9–14.
[18] L. Xia, C.-C. Chen, and J. Aggarwal, "View invariant human action recognition using histograms of 3D joints," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. Workshops*, 2012, pp. 20–27.
[19] C. Wang, Y. Wang, and A. L. Yuille, "An approach to pose-based action recognition," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 915–922.
[20] L. L. Presti, M. La Cascia, S. Sclaroff, and O. Camps, "Gesture modeling by Hanklet-based hidden Markov model," in *Proc. Asian Conf. Comput. Vis.*, 2014, pp. 529–546.
[21] I. Kapsouras and N. Nikolaidis, "Action recognition on motion capture data using a dynemes and forward differences representation," *J. Visual Commun. Image Represent.*, vol. 25, no. 6, pp. 1432–1445, 2014.
[22] S. Wu, S. Xia, Z. Wang, and C. Li, "Efficient motion data indexing and retrieval with local similarity measure of motion strings," *Visual Comput.*, vol. 25, nos. 5–7, pp. 499–508, 2009.
[23] J. Sedmidubsky, J. Valcik, and P. Zezula, "A key-pose similarity algorithm for motion data retrieval," in *Proc. IEEE Int. Conf. Adv. Concepts Intell. Vis. Syst.*, 2013, pp. 669–681.
[24] M. Kapadia, I. K. Chiang, T. Thomas, and N. I. Badler, "Efficient motion retrieval in large motion databases," in *Proc. ACM SIGGRAPH Symp. Interact. 3D Graph. Games*, 2013, pp. 19–28.
[25] M. E. Hussein, M. Torki, M. A. Gowayyed, and M. El-Saban, "Human action recognition using a temporal hierarchy of covariance descriptors on 3D joint locations," in *Proc. IEEE Int. Joint Conf. Artif. Intell.*, 2013, pp. 2466–2472.
[26] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy, "Sequence of the most informative joints (SMIJ): A new representation for human skeletal action recognition," *J. Visual Commun. Image Represent.*, vol. 25, no. 1, pp. 24–38, 2014.
[27] M. Yang, L. Zhang, X. Feng, and D. Zhang, "Sparse representation based fisher discrimination dictionary learning for image classification," *Int. J. Comput. Vision*, vol. 109, no. 3, pp. 209–232, 2014.
[28] M. Zhu, H. Sun, and Z. Deng, "Quaternion space sparse decomposition for motion compression and retrieval," in *Proc. ACM SIGGRAPH/Eurograph. Symp. Comput. Animat.*, 2012, pp. 183–192.
[29] T. Qi, Y. Feng, J. Xiao, Y. Zhuang, X. Yang, and J. Zhang, "A semantic feature for human motion retrieval," *Comput. Animat. Virtual Worlds*, vol. 24, nos. 3/4, pp. 399–407, 2013.
[30] L. Zhou, Z. Lu, H. Leung, and L. Shang, "Spatial temporal pyramid matching using temporal sparse representation for human motion retrieval," *Visual Comput.*, vol. 30, no. 6, pp. 845–854, 2014.
[31] K. Adistambha, C. H. Ritz, and I. S. Burnett, "Motion classification using dynamic time warping," in *Proc. IEEE Workshop Multimedia Signal Process.*, 2008, pp. 622–627.
[32] E. J. Keogh and M. J. Pazzani, "Derivative dynamic time warping," in *Proc. SIAM Int. Conf. Data Mining*, 2001, pp. 1–11.
[33] S. Shariat and V. Pavlovic, "Isotonic CCA for sequence alignment and activity recognition," in *Proc. Int. Conf. Comput. Vis.*, 2011, pp. 2572–2578.
[34] F. Zhou and F. Torre, "Canonical time warping for alignment of human behavior," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 2286–2294.
[35] S. A. Etemad and A. Arya, "Correlation-optimized time warping for motion," *Visual Comput.*, vol. 31, no. 12, pp. 1569–1586, 2015.
[36] M. Barnachon, S. Bouakaz, B. Boufama, and E. Guillou, "Human actions recognition from streamed motion capture," in *Proc. IEEE Int. Conf. Pattern Recognit.*, 2012, pp. 3807–3810.
[37] A. Eweiwi, M. S. Cheema, C. Bauckhage, and J. Gall, "Efficient pose-based action recognition," in *Proc. Asian Conf. Comput. Vis.*, 2014, pp. 428–443.
[38] E. Fotiadou and N. Nikolaidis, "Activity-based methods for person recognition in motion capture sequences," *Pattern Recognit. Lett.*, vol. 49, pp. 48–54, 2014.
[39] M. Müller, T. Röder, M. Clausen, B. Eberhardt, B. Krüger, and A. Weber, "Documentation MoCap database HDM05," Univ. Bonn, Bonn, Germany, Tech. Rep. CG-2007-2, June 2007.
[40] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 1290–1297.
[41] H. Shum and E. S. Ho, "Real-time physical modelling of character movements with microsoft kinect," in *Proc. ACM Symp. Virtual Reality Softw. Technol.*, 2012, pp. 17–24.
[42] X. He and P. Niyogi, "Locality preserving projections," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 16, 2004, pp. 153–161.
[43] R. Poppe, "A survey on vision-based human action recognition," *Image Vision Comput.*, vol. 28, no. 6, pp. 976–990, 2010.
[44] R. Timofte and L. Van Gool, "Iterative nearest neighbors," *Pattern Recognit.*, vol. 48, no. 1, pp. 60–72, 2015.
[45] M. Yang, L. Zhang, X. Feng, and D. Zhang, "Sparse representation based fisher discrimination dictionary learning for image classification," *Int. J. Comput. Vision*, vol. 109, no. 3, pp. 209–232, 2014.
[46] M. Yang, L. Zhang, J. Yang, and D. Zhang, "Metaface learning for sparse representation based face recognition," in *Proc. IEEE Int. Conf. Image Process.*, 2010, pp. 1601–1604.
[47] J. Shotton *et al.*, "Real-time human pose recognition in parts from single depth images," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 1297–1304.
[48] C. Chen, R. Jafari, and N. Kehtarnavaz, "Action recognition from depth sequences using depth motion maps-based local binary patterns," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2015, pp. 1092–1099.
[49] A. W. Vieira, E. R. Nascimento, G. L. Oliveira, Z. Liu, and M. F. Campos, "On the improvement of human action recognition from depth map sequences using space–time occupancy patterns," *Pattern Recognit. Lett.*, vol. 36, pp. 221–227, 2014.



Xin Liu (M'08) received the Ph.D. degree in computer science from Hong Kong Baptist University, Hong Kong, in 2013.

He was a Postdoctoral Fellow at the University of Macau, Macau, China. He is currently an Associate Professor in the Department of Computer Science and Technology, Huaqiao University, Xiamen, China. He has published and coauthored more than 30 papers. His research interests include computer vision, pattern recognition, machine learning, and medical image analysis.



Gao-Feng He received the M.Sc. degree in computer science from Huaqiao University, Xiamen, China, in 2016.

He is currently a Senior Algorithm Engineer with Deepdraw Intelligence Company Limited, Hangzhou, China. His research interests include motion analysis, pattern recognition, data mining, and computer graphics.

Mr. He received the China National Scholarship for graduate students in 2015.

Shu-Juan Peng received the B.E. degree from Hubei University, Wuhan, China, in 2004, and the Ph.D. degree in computer science and technology from Wuhan University, Wuhan, China, in 2009.

She is currently an Associate Professor in the Department of Computer Science and Technology, Huaqiao University, Xiamen, China. Her research interests include computer graphics, pattern recognition, and machine learning.

Yiu-ming Cheung (SM'06) received the Ph.D. degree in computer science and engineering from the Chinese University of Hong Kong, Hong Kong, in 2000.

He is currently a Full Professor in the Department of Computer Science, Hong Kong Baptist University, Hong Kong. His research interests include machine learning, pattern recognition, and visual computing.

Prof. Cheung is a Senior Member of the Association for Computing Machinery. He is the Founding Chairman of the Computational Intelligence Chapter of the IEEE Hong Kong Section. He is also an Associate Editor of the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, *Pattern Recognition*, *Knowledge and Information Systems*, and the *International Journal of Pattern Recognition and Artificial Intelligence*. He is an IET/IEE Fellow and BCS Fellow.

Yuan Yan Tang (F'04) received the Ph.D. degree in computer science from Concordia University, Montreal, QC, Canada.

He is currently a Chair Professor in the Faculty of Science and Technology, University of Macau, Macau, China, and a Professor, an Adjunct Professor, and an Honorary Professor with several institutions, including Chongqing University, Chongqing, China, Concordia University, and Hong Kong Baptist University, Hong Kong. He has authored or coauthored more than 370 technical papers published in journals and conference proceedings, and more than 20 monographs, books, and book chapters on electrical engineering and computer science. His research interests include wavelet theory and applications, pattern recognition, and artificial intelligence.

Prof. Tang is the Founder and the Editor-in-Chief of the *International Journal of Wavelets, Multiresolution, and Information Processing*, and an Associate Editor of several international journals related to pattern recognition and artificial intelligence. He is the Founder and the Chair of the Pattern Recognition Committee of the IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS, and the Founder and the General Chair of the International Conferences on Wavelets Analysis and Pattern Recognition. He is a Fellow of the IAPR.