Contents lists available at SciVerse ScienceDirect

J. Vis. Commun. Image R.

journal homepage: www.elsevier.com/locate/jvci

Sequence of the most informative joints (SMIJ): A new representation for human skeletal action recognition

Ferda Ofli [a,*], Rizwan Chaudhry [b], Gregorij Kurillo [a], René Vidal [b], Ruzena Bajcsy [a]

[a] Tele-immersion Lab, University of California – Berkeley, Berkeley, CA, USA
[b] Center for Imaging Sciences, Johns Hopkins University, Baltimore, MD, USA

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Human action representation
Human action recognition
Informative joints
Bag-of-words
Linear dynamical systems
Normalized edit distance
Cross-database generalization
Berkeley MHAD
HDM05
MSR Action3D

ABSTRACT

Much of the existing work on action recognition combines simple features with complex classifiers or models to represent an action. Parameters of such models usually do not have any physical meaning nor do they provide any qualitative insight relating the action to the actual motion of the body or its parts. In this paper, we propose a new representation of human actions called sequence of the most informative joints (SMIJ), which is extremely easy to interpret. At each time instant, we automatically select a few skeletal joints that are deemed to be the most informative for performing the current action based on highly interpretable measures such as the mean or variance of joint angle trajectories. We then represent the action as a sequence of these most informative joints. Experiments on multiple databases show that the SMIJ representation is discriminative for human action recognition and performs better than several state-of-the-art algorithms.

1. Introduction

Human motion analysis has remained as one of the most important areas of research in computer vision. Over the last few decades, a large number of methods have been proposed for human motion analysis (see the surveys by Moeslund et al. [1,2] and Turaga et al. [3] and most recently by Aggarwal and Ryoo [4] for a comprehensive analysis). In general all methods use a parametric representation of human motion and develop algorithms for comparing and classifying different instances of human activities under these representations.

One of the most common and intuitive methods for representation of human motion is a temporal sequence of approximate human skeletal configurations. The skeletal configurations represent hierarchically arranged joint kinematics with body segments reduced to straight lines. In the past, extracting accurate skeletal configurations from monocular videos was a difficult and unreliable process, especially for arbitrary human poses. Motion capture systems on the other hand could provide very accurate skeletal configurations of human actions based on active or passive markers positioned on the body; however, the data acquisition was limited to controlled indoor environments. Methods for human motion analysis that relied heavily on accurate skeletal data, therefore, became less popular over the years as compared to the image feature-based activity recognition methods. In the latter, spatio-temporal interest points are extracted from monocular videos and the recognition is based on learned statistics on large datasets [5–7]. Recently, with the release of several low-cost and relatively accurate 3D capturing systems, such as the Microsoft Kinect, real-time 3D data collection and skeleton extraction have become much easier and more practical for the applications of natural human computer interaction, gesture recognition and animation, thus reviving interest in the skeleton-based action representation.

Existing skeleton-based methods for human action recognition are primarily focused on modeling the dynamics of either the full skeleton or a combination of body segments. To represent the dynamics of normalized 3D positions of joints or joint angle configurations, most of the methods use linear dynamical systems (LDS) or non-linear dynamical systems (NLDS), e.g., in [8–10], or hidden Markov models (HMM), see, e.g., the earlier work by Yamato et al. [11] and a review of several others in [12]. Recently Taylor et al. [13,14] proposed using conditional restricted Boltzman machines (CRBM) to model the temporal evolution of human actions. While these methods have been very successful for both human activity synthesis and recognition, their representation of human motion is in general not easy to interpret in connection to the physical and qualitative properties of the human motion. For example, the parameters obtained from the LDS modeling of the skeletal joint trajectories will likely describe positions and velocities of the

* Corresponding author.
E-mail addresses: fofli@eecs.berkeley.edu (F. Ofli), rizwanch@cis.jhu.edu (R. Chaudhry), gregorij@eecs.berkeley.edu (G. Kurillo), rvidal@cis.jhu.edu (R. Vidal), bajcsy@eecs.berkeley.edu (R. Bajcsy).

1047-3203/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jvcir.2013.04.007

individual joints, which do not directly convey any information about the changes in the skeletal configuration of the human body as the action is performed.

When humans perform an action, we can observe that each individual performs the same action with a different style, generating dissimilar joint trajectories; however, all individuals activate the same set of joints contributing to the overall movement, roughly in the same order. In our approach we take advantage of this observation to capture invariances in human skeletal motion for a given action. Given an action, we propose to identify the most informative joints in a particular temporal window by finding the relative informativeness of all the joints in that window. We can quantify the informativeness of a joint using, for example, the entropy of its joint angle time series. In the case of a Gaussian random variable, its entropy is proportional to the logarithm of its variance. Therefore, the joint that has the highest variance of motion as captured by the change in the joint angle can be defined as the most informative, assuming the joint angle data are independent and identically distributed (i.i.d.) samples from a one-dimensional Gaussian. Such a notion of informativeness is very intuitive and interpretable. During performance of an action, we can observe that different joints are activated at different times with various degree. Therefore, the ordered sequence of informative joints in a full skeletal motion implicitly encodes the temporal dynamics of the motion.

Based on these properties, we recently proposed in [16] the *sequence of the most informative joints (SMIJ)* as a new representation for human motion based on the temporal ordering of joints that are deemed to be the most informative for performing an action. In [16], we briefly compared the performance of the SMIJ representation to other action representations, based on the histograms of motion words, as well as the methods that explicitly model the dynamics of the skeletal motion. In this paper, we provide a more comprehensive description of the SMIJ representation and other feature representations and further evaluate their quality using action classification as our performance test. In addition, we propose a different metric for comparison of SMIJ features, based on normalized edit distance [17], which outperforms the normalized Levenshtein distance, applied in our previous work. Furthermore, we show that our simple yet highly intuitive and interpretable representation performs much better than standard methods for the task of action recognition from skeletal motion data.

2. Sequence of the most informative joints (SMIJ)

The human body is an articulated system that can be represented by a hierarchy of joints that are connected with bones, forming a *skeleton*. Different joint configurations produce different skeletal poses and a time series of these poses yields the skeletal motion. An action can thus simply be described as a collection of time series of 3D positions (i.e., 3D trajectories) of the joints in the skeleton hierarchy. This representation, however, lacks important properties such as invariance with respect to the choice of the reference coordinate system and scale of the human.

A better description is obtained by computing the joint angles between any two connected limbs and using the time series of joint angles as the skeletal motion data. Let $\mathbf{a}^i$ denote the joint angle time series of joint i, i.e., $\mathbf{a}^i = \{a^i_t\}_{t=1}^{t=T}$ where T is the number of frames in an action sequence. An action sequence can then be seen as a collection of such time-series data from different joints, i.e., $\mathbf{A} = [\mathbf{a}^1 \mathbf{a}^2 \dots \mathbf{a}^J]$, where J is the number of joints in the skeleton hierarchy. Hence, $\mathbf{A}$ is the $T \times J$ matrix of joint angle time series representing an action sequence.

Common modeling methods such as LDS or HMM model the evolution of the time series of joint angles. However, instead of directly using the original joint angle time-series data $\mathbf{A}$, one can also extract various types of features from $\mathbf{A}$ that in general reduce the size of data yet preserve the information that is discriminative of each action. For the sake of generality, we denote this operation with the mapping function $\mathcal{O}$ in the remainder of this paper unless an explicit specification is necessary. Here $\mathcal{O}(\mathbf{a}) : \mathbb{R}^{|\mathbf{a}|} \rightarrow \mathbb{R}$ is a function that maps a time series of scalar values to a single scalar value. Furthermore, one can extract such features either across the entire action sequence (i.e., global features) or across smaller segments of the time-series data (i.e., local features). The former case describes an action sequence with its global statistics, whereas the latter case emphasizes more the local temporal statistics of an action sequence. Examples include the *mean* or *variance* of joint angle time series, or the *maximum angular velocity* of each joint as observed over the entire action sequence or inside a small temporal window.

2.1. Motivation of the proposed representation

In this paper we approach the action recognition with the following hypothesis: Different actions require humans to engage different joints of the skeleton at different intensity (energy) levels at different times. Hence, the ordering of joints based on their level of engagement across time should reveal significant information about the underlying dynamics, i.e., the invariant temporal structure of the action itself.

In order to visualize this phenomenon, let us consider the labeled joint angle configuration shown in Fig. 1(a), and perform a simple analysis on the Berkeley multimodal human action database (Berkeley MHAD), (see Section 4.1 for details about the datasets). The analysis is based on the following steps: (i) partition the joint angle time series of an action sequence into a number of congruent temporal segments, (ii) compute the *variance* of the joint angle time series of each joint over each temporal segment (note that the mapping function $\mathcal{O}$ is defined to be the *variance* operator in this particular case), (iii) rank-order the joints within each segment based on their variance in descending order, (iv) repeat the first three steps to get the orderings of joints for all the action sequences in the dataset. Below we investigate the resulting set of joint orderings for different actions in the Berkeley MHAD.

Fig. 1(b) shows the distribution of the most informative, i.e., the top-ranking, joint for different actions across all repetitions from all subjects in the Berkeley MHAD. Specifically, each entry in the figure shows the percentage of the time that the trajectory of a given joint within a segment has the highest variance. Notice that simple actions, such as *punch* and *wave one*, engage only a small number of joints, while more complex actions, such as *sit-stand*, engage several joints in different parts of the body. Nevertheless, the set of the most informative joints are different for different actions. Joint 10 (*RElbow*) is the most informative joint 47% of the time, followed by joint 9 (*RArm*) 35% of the time in the *wave one* action. Both joints 10 (*RElbow*) and 13 (*LElbow*) are the most informative joints more than 40% of the time in the *punch* action. On the other hand, almost half of the joints appear as the most informative at some point in the actions *sit-stand*, *sit down* and *stand up*; however, the differences across the sets of engaged joints in each of these three actions are still noticeable. For instance, joint 19 (*LKnee*) is engaged more in the *sit-stand* action than in the *sit down* and *stand up* actions.

Fig. 2 shows the stacked histogram distributions of the 6 most informative joints for four different actions taken from the Berkeley MHAD. Even though the overall set of the most informative joints looks similar for the actions *jump* and *jumping jacks*, there are significant differences particularly in the distribution of joints at different rankings for different actions. Specifically, joints 15 (*RKnee*) and 19 (*LKnee*) appear more than 60% of the time as either the 1st- or 2nd-ranking joint for the *jump* action whereas this ratio is between 40% and 50% for the *jumping jacks* action. Furthermore,

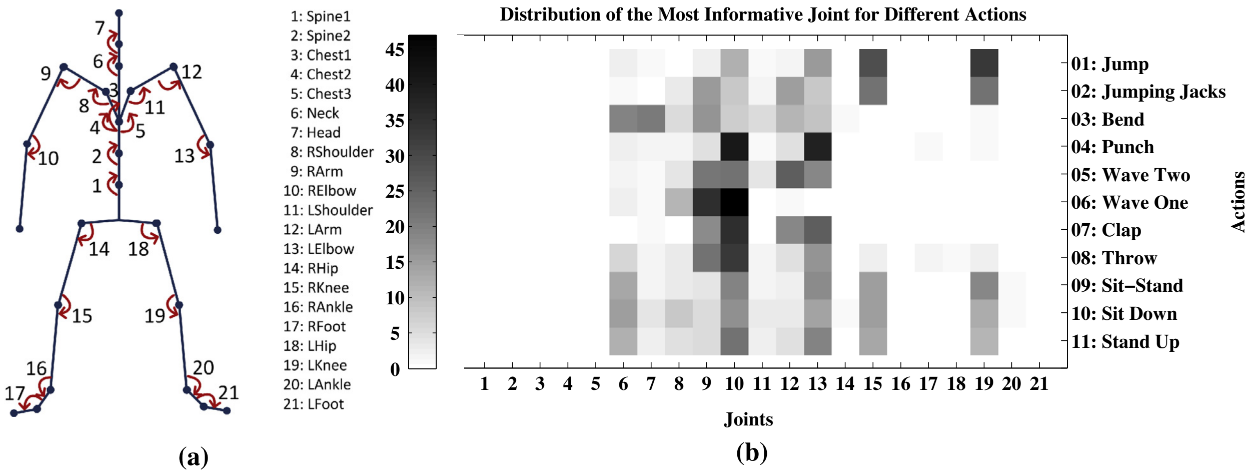

Fig. 1. (a) The structure of the skeleton used in the Berkeley MHAD and the corresponding set of 21 joint angles. (b) Distribution of the most informative joint for different actions in the Berkeley MHAD. Each entry corresponds to the percentage of the time that a given joint is deemed to be the most informative for a given action (darker means higher percentage). Some actions, such as *punch* and *wave one*, are represented only with a few number of joints whereas other actions, such as *sit-stand*, *sit down* and *stand up*, require many more joints.

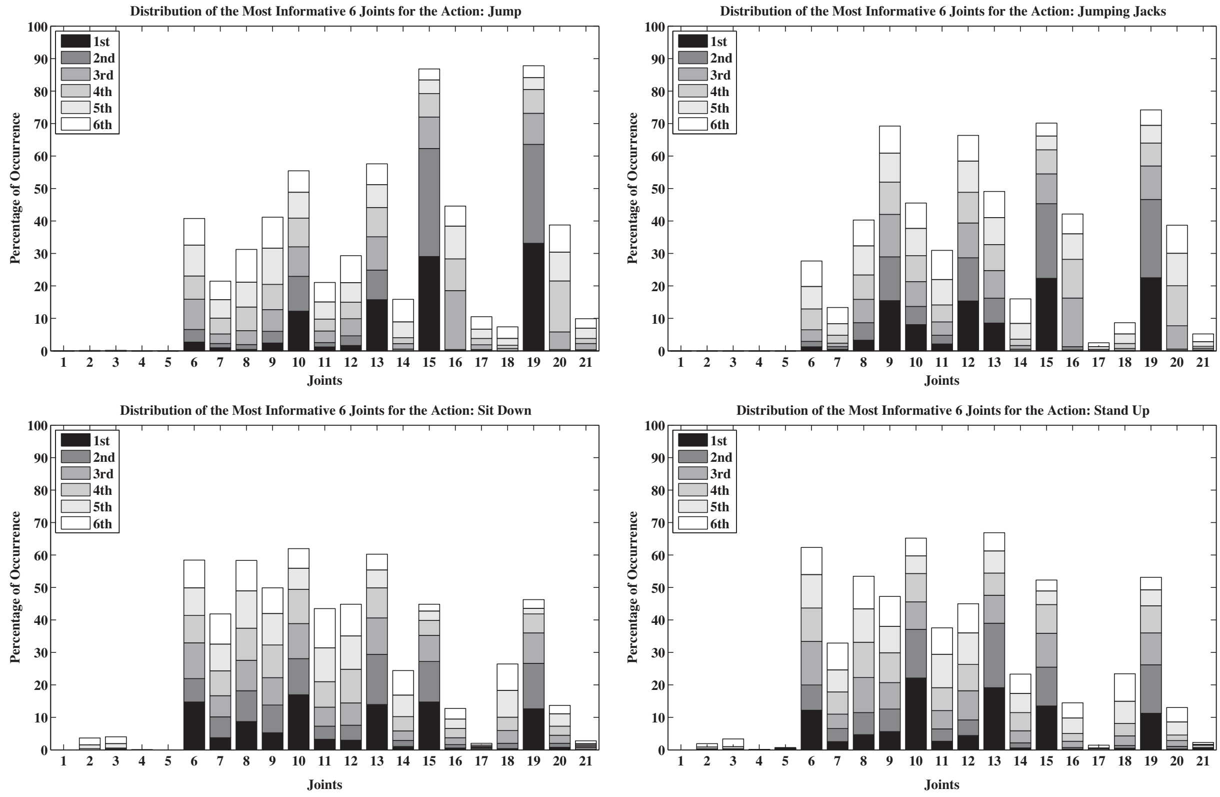

Fig. 2. Stacked histogram distribution of the most informative, i.e., the top-ranking, 6 joints for four actions selected from the Berkeley MHAD.

joints 10 (*RElbow*) and 13 (*LElbow*) tend to rank in the top three at least 30% of the time for the *jump* action whereas joints 9 (*RArm*) and 12 (*LArm*) tend to rank in the top three for the *jumping jacks* action. On the contrary, for the actions *sit down* and *stand up*, both the overall set of the most informative joints and the distribution of joints at different rankings are very similar. We further examine the temporal orderings of the most informative joints to demonstrate how the proposed representation can distinguish between actions with similar histogram distributions.

Fig. 3 shows the temporal orderings of the most informative joint across the first seven subjects in the Berkeley MHAD for the actions *sit down* and *stand up*. The joints are color-coded according to the color bar displayed on the right side, while similar color codes are used for symmetric joints to improve the clarity of the visualization. Furthermore, the action recordings from different subjects are sorted with respect to their lengths to better emphasize the similarities between the temporal orderings of the most informative joint for different action recordings. The plots suggest

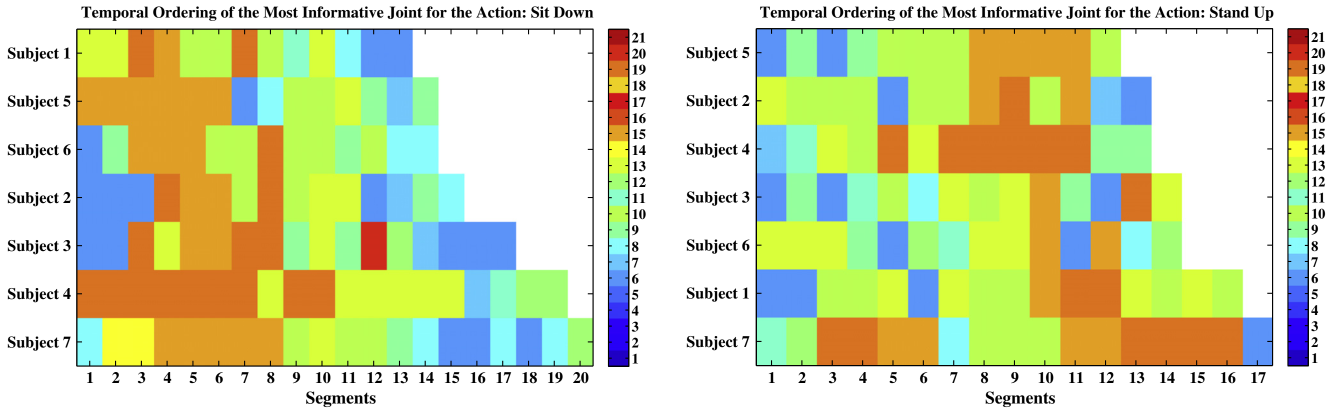

Fig. 3. Temporal orderings of the most informative joint for the actions *sit down* and *stand up* across the first seven subjects in the Berkeley MHAD. Each row of a plot represents the sequence of the most informative joint extracted from the first action recording of the corresponding subject.

that even though the overall set of the most informative joints look very similar for the actions *sit down* and *stand up*, the temporal ordering of the most informative joint can be used to distinguish these two actions successfully. Specifically, the *orange*-colored joints (i.e., joints 15 (*RKnee*) and 19 (*LKnee*)) are engaged the most at the beginning of the *sit down* action as opposed to being engaged the most towards the end of the *stand up* action. Conversely, the *blue*/*green*-colored joints (i.e., joints 9 (*RArm*), 10 (*RElbow*), 12 (*LArm*) and 13 (*LElbow*)) are engaged the most interchangeably at the end of the *sit down* action as opposed to being engaged the most at the beginning of the *stand up* action. The observed temporal ordering corresponds to the nature of the action as the subjects used their arms for support when first getting out of the chair and conversely when sitting down.

In summary, our initial analysis suggests that different sets of joints, as well as their temporal ordering, reveal discriminative information about the underlying structure of the action. This is precisely the main motive to propose sequences of the top N most informative joints as a new feature representation for human skeletal action recognition. Hence, the new feature representation, which we refer to as *sequence of the most informative joints* (SMIJ), has two main components: (i) the set of the most informative joints in each time segment, and (ii) the temporal ordering of the set of the most informative joints over all of the time segments.

2.2. Segmentation

To extract the SMIJ representation from the joint angle time-series data, we first need to partition the action sequence into a number of, say N_s, temporal segments. Let $\mathbf{a}^i_k = \{a^i_k\}_{k=t^k_s,\ldots,t^k_e}$ be a segment of $\mathbf{a}^i$ where t^k_s and t^k_e denote the start and the end frames for the segment k, respectively. Then, the joint angle time-series data of joint i, $\mathbf{a}^i$, can be written as a collection of temporal motion segments, $\mathbf{a}^i = \{\mathbf{a}^i_k\}_{k=1,\ldots,N_s}$.

In general the length of the motion segments should be such that significant information about the underlying dynamics of the movement is contained within the segment. We hypothesize that the size and the type of partitioning will influence the discriminative properties of the proposed SMIJ representation since the atomic motion units will be captured differently depending on the size and the number of the segments. The temporal segmentation of human motion into such atomic motion units is, however, still an open research problem. Some of the top-down approaches in this domain rely on creating models of atomic motion units a priori from expert knowledge and training data [18,19]. Some of the bottom-up segmentation techniques, on the other hand, utilize the principle component analysis and data compression theory [20–22]. The problem of temporal segmentation of human activities is, however, beyond the scope of this study. In this paper, we thus examine two different elementary methods to partition the action sequences:

(a) *Segmentation with Fixed Number of Segments*. In this approach, we divide the time-series data associated with one sample action sequence into a fixed number of congruent segments of variable size. The resulting feature representation thus have the *same size* for all action sequences. However, this method yields shorter temporal segments for shorter action sequences and longer temporal segments for longer action sequences, which can be interpreted as a temporal normalization of the actions. Note, that this is the same segmentation approach used in our previous work on SMIJ representation [16].

(b) *Segmentation with Fixed Temporal Window*. In the second approach, we divide the time-series data associated with one sample action sequence into a variable number of congruent segments of a given fixed size. This alternative approach is novel for the SMIJ representation and leads to uniform temporal analysis of all action sequences. Unlike the first approach, where the feature representations are of the same length for all the actions, this approach yields different number of segments for different action sequences. Hence, the resulting feature representation has *different size* for different action sequences.

The *fixed temporal window* based segmentation is prevalent in several domains such as speech/audio processing, or signal processing in general, since it is more intuitive and generalizable. We likewise hypothesize that the *fixed temporal window* approach will yield more discriminative SMIJ representations of different actions than the *fixed number of segments* approach. We analyze the results of these two partitioning approaches in Section 4. Examination of more advanced temporal segmentation methods remains to be our future work.

2.3. Measure of information

Once we partition the action sequence into congruent temporal segments, we apply the mapping function $\mathcal{O}$ to each segment and write the action sequence as a collection of features, $\mathbf{F} = \{\mathbf{f}_k\}_{k=1,\ldots,N_s}$, where

$$\mathbf{f}_k = \left[\mathcal{O}\left(\mathbf{a}^1_k\right)\mathcal{O}\left(\mathbf{a}^2_k\right)\cdots\mathcal{O}\left(\mathbf{a}^J_k\right)\right]. \tag{1}$$

The feature function, $\mathcal{O}(\mathbf{a}_k^i)$, provides a measure of informativeness of the joint i in the temporal segment k. In information theory, one measure of information of a signal is the *entropy* [23], which is defined as

$$h(X) = -\int_{\mathcal{X}} f(x)\log f(x)dx, \tag{2}$$

where X is a continuous random variable with probability density function f whose support is a set $\mathcal{X}$. For a Gaussian distribution with variance σ^2, the entropy can be calculated as

$$h(X) = \frac{1}{2}\left(\log 2\pi\sigma^2 + 1\right). \tag{3}$$

Therefore, the entropy of a Gaussian random variable is proportional to the logarithm of its variance [15]. Assuming that the joint angle time-series data $\mathbf{a}^i$ are i.i.d. samples from a one-dimensional Gaussian distribution, we can measure the informativeness of $\mathbf{a}^i$ in each temporal segment by computing the corresponding variance in each segment. Based on this assumption, we choose the mapping function $\mathcal{O}$ to be the *variance* operator in the remainder of this paper. A more sophisticated information theoretic measure of informativeness that also considers signal noise remains to be a future work.

2.4. Ordering

After the temporal segmentation and mapping of each joint-angle time series, we rank-order all the joints in $\mathbf{f}_k$ within each temporal segment k based on their informativeness, i.e., $\mathcal{O}(\mathbf{a}_k^i)$, and define SMIJ features as

$$\begin{aligned} &\mathbf{S} = (s_{kn})_{k=1,\ldots,N_s;n=1,\ldots,N}, \\ &s_{kn} = \text{idof}(\text{sort}(\mathbf{f}_k), n), \end{aligned} \tag{4}$$

where the sort operator *sorts* the joints based on their local $\mathcal{O}$ score in descending order, the idof $(\cdot, n)$ operator returns the *id* (or equivalently, the *label*) of a joint that ranks nth in the joint ordering, and N specifies the number of top-ranking joints included in the representation, resulting in a $N_s \times N$-dimensional feature descriptor. In other words, the SMIJ features represent an action sequence by encoding the set of N most informative joints at a specific time instant (by rank-ordering and keeping the top-ranking N joints) as well as the temporal evolution of the set of the most informative joints throughout the action sequence (by preserving the temporal order of the top-ranking N joints). Fig. 4 shows the most informative 6 joints at the key frames of two actions selected from the HDM05 dataset (see Section 4.1 for details about the datasets).

We acknowledge that using the proposed representation of an action as a sequence of rank-ordered joints significantly reduces the amount of information contained in the original time series data. We argue, however, that the remaining information is discriminative enough to distinguish different actions. In our experiments, we show that such an extreme abstraction in the feature representation yields satisfactory results for action recognition. A more detailed feature representation will become necessary when the set of the most informative joints and their orderings are very similar for two different actions. In such case, we can enrich the representation by retaining not only the ranking of the most informative joints, but also some information about the motion of the most informative joint, such as the direction of the motion of the most informative joint, within each temporal segment. We leave this avenue as a future research direction.

2.5. Metrics for comparing SMIJ

Each SMIJ is defined over a fixed alphabet – the joint labels. Therefore, comparison of the SMIJ features from two different sequences $\mathbf{S}^i$ and $\mathbf{S}^j$ is equivalent to comparison of strings. The distance metric as a measure of similarity between two strings with finite sequence of symbols is often defined using edit-distance functions, which consist of counting the minimum number of edit operations needed to transform one string into the other. The edit operations include *insertion*, *deletion*, *substitution* of a single character, or *transposition* of two adjacent characters. These four edit operations were first introduced by Damerau [24] who applied them to automatic detection and correction of spelling errors. Subsequent to Damerau's work, Levenshtein introduced in [25] the corresponding edit distance to deal with multiple edit operations, such as deletion, insertion, and reversals, but excluded *transposition*. His distance metric is known as the Levenshtein distance. Wagner and Fischer first proposed a dynamic programming algorithm for calculating the Levenshtein distance in [26] and then extended the set of allowable edit operations to include *transposition* of two adjacent characters in [27].

Interestingly, none of the aforementioned algorithms considered *normalization* of the distance metric that would appropriately rate the weight of the (edit) errors with respect to the length of the sequences (strings) that are compared. Even though the normalization may not be crucial for comparing strings of the same length, it becomes critical for comparing strings of different lengths, as pointed out by Marzal and Vidal in [17]. In their seminal work, Marzal and Vidal proposed an algorithm called the normalized edit distance (NED) based on finding the minimum of $W(P)/L(P)$ (not only the minimum of $W(P)$), where P is an editing path between $\mathbf{S}^i$ and $\mathbf{S}^j, W(P)$ is the sum of the weights of the elementary edit operations of P, and $L(P)$ is the number of these operations (length of P). They also emphasized that this minimization cannot be carried out by first minimizing $W(P)$ and then normalizing it by the length of the obtained path $L(P)$, which they refer to as *post-normalization*.

In our original work on the activity recognition using SMIJ features [16], we used the *post-normalized* Levenshtein distance to compare the sequences of features. All the action sequences were partitioned into *fixed number of segments*, generating feature vectors of the same length. The *post-normalized* Levenshtein distance was thus able to properly describe the level of similarity between the sequences. In this paper, we extend the temporal segmentation by including segmentation with the *fixed temporal window* which generates sets of sequences of various lengths. The Levenshtein distance does not properly account for the length variations between the sequences since it only seeks to find the minimum number of edit operations, in other words, minimizes $W(P)$ only. Instead, we apply a more sophisticated distance metric, i.e., the normalized edit distance [17], which considers variable lengths of feature vectors and a proper normalization of the distance metric.

Using the normalized edit distance, we define the distance between two different SMIJ representations $\mathbf{S}^i$ and $\mathbf{S}^j$ as follows:

$$D_S\left(\mathbf{S}^i, \mathbf{S}^j\right) = \sum_{n=1}^{N} \text{NED}(s_{*,n}^i, s_{*,n}^j), \tag{5}$$

where $s_{*,n}$ refers to the nth column of $\mathbf{S}$, i.e., the sequence of nth-ranking joints, and N is the number of the most informative joints included in the SMIJ representation.

3. Alternative feature representations

In this section, we briefly describe three alternative feature representations against which we compare the results of the proposed SMIJ representation. We consider two standard methods, linear dynamical system parameters (LDSP) with a linear system identification approach and histogram of motion words (HMW) with a

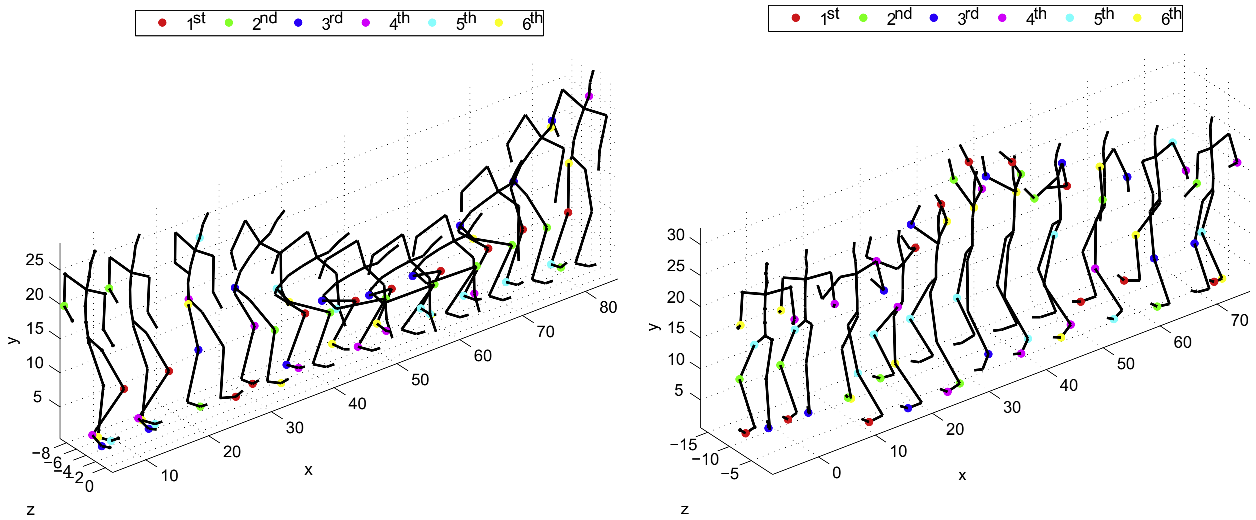

Fig. 4. The most informative 6 joints are highlighted along the key frames of two different actions from HDM05 dataset (*deposit floor* on the left and *throw basketball* on the right).

bag-of-words model. In addition we compare our results with the histogram of the most informative joints (HMIJ) which was also proposed originally in [16] as an alternative to SMIJ. We believe that the three alternative feature representations, i.e., HMIJ, HMW and LDSP, adopted from a wide range of popular approaches, allow us to demonstrate the power of the proposed SMIJ features in terms of discriminability and interpretability for human action recognition.

3.1. Histogram of motion words (HMW)

Histogram of motion words is based on the popular bag-of-words method [28] for visual categorization. In this approach, we first cluster the set of all $\mathbf{f}_k$s for a given action sequence into K clusters (i.e., motion words) using K-means or K-medoids. Next, we count for each sequence the number of motion words by assigning each $\mathbf{f}_k$ to its closest motion word. After l_1-normalization, we obtain the histogram-of-motion words (HMW) representation, which captures the overall distribution of motion words in the form of a histogram for each sequence. Since the HMW ignores temporal relations between smaller action units, a sequence with scrambled $\mathbf{f}_k$ will yield the same HMW representation as the original sequence.

As the distance metric for comparing HMW features, we use χ^2 distance defined as follows:

$$D_{\chi^2}\left(\mathrm{H}^i,\mathrm{H}^j\right) = \frac{1}{2}\sum_{k=1}^{K}\frac{\left(h_k^i - h_k^j\right)^2}{\left(h_k^i + h_k^j\right)}, \tag{6}$$

where $\mathrm{H}^i = (h_1^i,\ldots,h_K^i)$ and $\mathrm{H}^j = (h_1^j,\ldots,h_K^j)$ are two l_1-normalized histograms and K is the number of bins in the histograms, or equivalently, the number of clusters that has to be decided a priori. For all the experiments we chose $K = 20$ (which will be discussed further in Section 4.2). Note, that since the final clustering result depends on the initial condition, the final recognition rate can change based on the motion words computed during the clustering stage. We therefore perform 20 runs for each clustering experiments and compute the corresponding HMW representations for each action sequence. We provide the mean and standard deviation of the classification rates achieved over these 20 runs using methods such as 1-nearest neighbor (1-NN) with χ^2 distance on histograms and support vector machine (SVM) with a Gaussian kernel using χ^2 distance on histograms.

3.2. Linear dynamical system parameters (LDSP)

As mentioned earlier, one of the most common techniques to analyze human motion data is based on modeling the motion with a linear dynamical system over the entire sequence (e.g., see [8,29–31] for more details) and using the LDS parameters (LDSP) as an alternative feature representation. A general LDS is defined by the following set of equations:

$$\mathbf{y}_t = \boldsymbol{\mu} + C\mathbf{x}_t + \mathbf{w}_t, \tag{7}$$

$$\mathbf{x}_t = A\mathbf{x}_{t-1} + B\mathbf{v}_t, \tag{8}$$

where $\mathbf{y}_t \in \mathbb{R}^p$ is the output of the LDS at time t and is linearly related to the hidden state, $\mathbf{x}_t \in \mathbb{R}^m$. Furthermore, $\mathbf{x}_t$ depends linearly on only the previous state $\mathbf{x}_{t-1}$. $\boldsymbol{\mu} \in \mathbb{R}^p$ is the mean output and $\mathbf{v}_t \in \mathbb{R}^{m_v}$ is a zero-mean, unit variance i.i.d. Gaussian process that drives the state process $\mathbf{x}_t$. Similarly, $\mathbf{w}_t \in \mathbb{R}^p \sim \mathcal{N}(0,\sigma^2 I)$ is a zero mean uncorrelated output noise process. The joint-angle time-series is hence modeled as the output of an LDS and can therefore be represented by the tuple $(A, C, B, \boldsymbol{\mu}, \sigma^2, \mathbf{x}_0)$, where $A \in \mathbb{R}^{m\times m}$ is the dynamics matrix, $C \in \mathbb{R}^{p\times m}$ is the observation matrix, $B \in \mathbb{R}^{m\times m_v}(m_v \leqslant m)$ is the input-to-state matrix, σ^2 is the identical covariance of each output dimension, and $\mathbf{x}_0$ is the initial state of the system.

Given a feature time-series, these parameters can be computed using *system identification* for which several methods exist, e.g., N4SID [32] and EM [33]. We choose to use the sub-optimal but very fast method by Doretto et al. [34] to identify the system parameters for the joint-angle time-series of a given action sequence.

Once these parameters are identified for each of the action sequences, various metrics can be used to define the similarity between these LDSs. In particular, three major types of metrics are (i) geometric distances based either on subspace angles between the *observability* subspaces of the LDSs [35] or on an *alignment* distance between two LDS parameters under a group action [36], (ii) algebraic metrics such as the Binet-Cauchy kernels [37], and (iii) information theoretic metrics such as the KL-divergence [38]. We use the geometric distance known as the Martin distance [35] as the metric between dynamical systems for classification based on LDSP using methods such as 1-NN and SVM.

The LDSP representation, as opposed to the aforementioned SMIJ and HMW representations, does not require segmentation of the joint angle trajectories. Therefore, it captures only the global temporal information about an action while ignoring the details on the local levels.

3.3. Histograms of the most informative joints (HMIJ)

Finally, we propose an alternative feature representation, which is based on a similar idea to SMIJ, but disregards the temporal information. Instead of stacking the most informative N joints from all temporal segments into a matrix of symbols, while keeping the temporal order of the joints intact, we create histograms separately for the 1st-ranking joints, 2nd-ranking joints, and so on, from all temporal segments. The histograms are then concatenated as a feature descriptor, called histograms of the most informative joints (HMIJ), to represent each action sequence in the following form:

$$\mathrm{HMIJ} = \left(\mathrm{hist}\left(\{\mathrm{idof}(\mathrm{sort}(\mathbf{f}_k), n)\}_{k=1,\ldots,N_s}\right)\right)_{n=1,\ldots,N}. \qquad (9)$$

Here the hist operator creates a J-bin l_1-normalized histogram from the input joint sequence, resulting in JN-dimensional feature descriptor. Since HMIJ is a histogram-based representation, we use the χ^2 distance given in (6) to compute the distance between HMIJ features for classification based on 1-NN and SVM with a Gaussian kernel.

It is important to note that the HMIJ feature representation ignores the temporal order of the most informative N joints, and hence, it will be useful for evaluating the importance of preserving the temporal ordering in the feature representation, which is preserved by the SMIJ feature representation.

4. Experiments

In this section we compare our proposed feature representation SMIJ (described in Section 2) against the baseline feature representations (explained in Section 3) on the datasets outlined in Section 4.1 using action recognition as a test, and provide experimental results in Section 4.2.

4.1. Datasets

We evaluate the performance of each feature representation described above on three different human action datasets of 3D skeleton data. Two of the datasets were obtained using a high quality motion capture system, while the third one contains skeleton data obtained from a single-viewpoint depth sensor. Each dataset has almost completely distinct set of actions with different frame rates, different skeleton extraction method, and hence, skeleton data of various dynamic properties and data fidelity. The goal of including such diverse input data was to examine how discriminative the proposed SMIJ method is with respect to the varying properties of these datasets. The diversity is relevant in the first set of experiments where we aim to evaluate the performance of the feature representations on a wide range of actions. For the second set of experiments, where we evaluate the action recognition across datasets, we select a small subset of actions that are shared between the first two datasets.

4.1.1. Berkeley multimodal human action database (Berkeley MHAD)

We recently collected a dataset that contains 11 actions performed by 12 subjects using an active optical motion capture system (PhaseSpace Inc, San Leandro, CA) [39]. The motion data was recorded with 43 active LED markers at 480 Hz. For each subject we collected 5 repetitions of each action, yielding a total of 659 action sequences (after excluding one erroneous sequence). We then extracted the skeleton data by post-processing the 3D optical motion capture data. The actions lengths vary from 773 to 14,565 frames (corresponding to approximately 1.6–30.3 s). The set of actions consisted of *jump, jumping jacks, bend, punch, wave one hand, wave two hands, clap, throw, sit down, stand up,* and *sit down/stand up*.

4.1.2. Motion capture database HDM05

From the popular HDM05 database [40] we arbitrarily selected 16 actions performed by 5 subjects. In this dataset, subjects performed each action with various number of repetitions, resulting in 393 action sequences in total. The motion capture data, which was captured with the frequency of 120 Hz, also includes the corresponding skeleton data. The duration of the action sequences ranges from 56 to 901 frames (corresponding to approximately 0.5–7.5 s). The set of actions consisted of *deposit floor, elbow to knee, grab high, hop both legs, jog, kick forward, lie down floor, rotate both arms backward, sneak, squat, throw basketball, jump, jumping jacks, throw, sit down,* and *stand up*.

4.1.3. MSR Action3D database

Finally, we also evaluated the action recognition performance on the MSR Action3D dataset [41] consisting of the skeleton data obtained from a depth sensor similar to the Microsoft Kinect with 15 Hz. Due to missing or corrupted skeleton data in some of the available action sequences, we selected a subset of 17 actions performed by 8 subjects, with 3 repetitions of each action. The subset consisted of 379 action sequences in total, with the duration of the sequences ranging from 14 to 76 frames (corresponding to approximately 1–5 s). The set of actions included *high arm wave, horizontal arm wave, hammer, hand catch, forward punch, high throw, draw x, draw tick, draw circle, hand clap, two hand wave, side-boxing, forward kick, side kick, jogging, tennis swing,* and *tennis serve*.

4.1.4. Database standardization

Before proceeding with the action recognition experiments on the aforementioned datasets, we need to consider some important factors that can potentially influence the recognition results. The three databases have different acquisition frame rates, which affect the maximal number of temporal segments that can be extracted from a sequence. Another important factor to be considered is the number of repetitions of an action unit in a particular sample. The Berkeley MHAD database for instance, contains five repetitions of an action unit in each sample sequence, e.g., the person jumps or claps five times, whereas in the other two datasets, the subject performs only one repetition of similar action unit.

For the purpose of consistent and objective comparison of the classification performance, we standardize the datasets with respect to the frame rate and the action unit segmentation. The action sequences of the Berkeley MHAD dataset with multiple repetitions of the same action were split into individual sequences with only one repetition, thus extending the dataset by several shorter action sequences. Furthermore, we downsampled the extended Berkeley MHAD dataset from 480 fps to 120 fps and upsampled the MSR Action3D dataset from 15 fps to 120 fps to match the frame rate for all the datasets, using lowpass decimation and interpolation, respectively.

4.2. Action recognition results

In this section we examine the quality of different feature representations by evaluating their classification performance using well-established methods such as 1-nearest neighbor (1-NN) and support vector machine (SVM) with the corresponding distance metrics introduced earlier in the respective subsections. For SVM based classification, we follow one-vs-one classification scheme and use Gaussian kernel $K(\bullet,*) = e^{-\gamma D^2(\bullet,\star)}$ with an appropriate distance function $D(\bullet,*)$ depending on the feature type listed above.[1] As for the SVM hyperparameters, we set the regularization parameter C to 1 and the Gaussian kernel function parameter γ to the inverse of the mean value of the distances between all training sequences as suggested in [42]. In order to determine the number of clusters (K) for the HMW representation, we performed preliminary classification experiments for different values of K, and observed that the performance saturates for $K > 20$. In addition, we aim to match the dimensions of the histograms in the HMW representation to the dimensions of the histograms in the HMIJ representation.[2] Since the histograms in the HMIJ representation can be 20- to 22-dimensional,[3] we chose $K = 20$ for the HMW representation.

4.2.1. Action recognition on the same database

In the first set of experiments, we performed action recognition on each of the aforementioned datasets separately. We used roughly 60% of the data for training and the remainder for testing. Specifically, we used 7 subjects (384 action sequences) for training and 5 subjects (275 action sequences) for testing on the Berkeley MHAD database, 3 subjects (216 action sequences) for training and 2 subjects (177 action sequences) for testing on the HDM05 database, and finally, 5 subjects (226 action sequences) for training and 3 subjects (153 action sequences) for testing on the MSR Action3D database.

Fig. 5 shows the classification results for the first three feature types (i.e., SMIJ, HMIJ and HMW) on three different datasets for a range of values for the *fixed number of segments* approach. Specifically, the plots in different columns correspond to different feature types, i.e., SMIJ, HMIJ and HMW, from left to right, respectively. The plots in different rows show recognition results from different datasets, i.e., Berkeley MHAD, HDM05 and MSR Action3D, from top to bottom, respectively. In all plots, the vertical axis represents the classification performance and the horizontal axis represents the number of segments, ranging from 5 to 50 with a step size of 5. Different colors in the SMIJ and HMIJ plots represent different number of the most informative joints (N) included in the feature representation. On the other hand, different colors in the HMW plots represent different clustering methods (i.e., K-means and K-medoids) used to obtain final feature representation. The solid lines in the plots show SVM-based classification results whereas the dotted lines show NN-based classification results. The HMW plots show the mean and standard deviation of the classification rates achieved over 20 runs as explained in Section 3.1. Arranged identically to Fig. 5, Fig. 6 shows the classification results for a range of different window sizes for the *fixed temporal window* approach.

We observe in Fig. 5 that as we increase the number of segments in the *fixed number of segments* approach, the classification performance first improves and then saturates when the number of segments is sufficiently large, especially for Berkeley MHAD and HDM05 datasets. On the contrary, we see the opposite trend in Fig. 6. That is, the classification performance in general tends to decrease as we increase the temporal window size in the *fixed temporal window* approach. These two observations are consistent since increasing the number of segments corresponds to decreasing the segment size, and vice versa. Nevertheless, determining a proper window size is important. A very large window size results in poorer time resolution that yields over-smoothed statistical information about the underlying time-series data whereas a really short window size results in unreliable (i.e., noisy) statistical information that degrades the quality of the representation. Therefore, a window size that matches (or is reasonably proportional to) the duration of atomic action units under consideration should be sought for the best performance of the feature representation. For the set of databases we examine in this paper, our experimental results suggest that a window size of 40 or 50 ms in general yields the optimal performance for all segmentation based feature representations, i.e., SMIJ, HMIJ and HMW, (which is further discussed in the remainder of this section).

Next, we can observe in Figs. 5 and 6 that the recognition results improve when using more than the single most informative joint ($N > 1$). The blue lines (both solid and dotted) in the SMIJ and HMIJ plots indicate lower classification rates for $N = 1$ with respect to the other different colored lines which represent the classification results for $N > 1$. Another observation common to the plots in Figs. 5 and 6 is that SVM-based classification results (solid lines) are in general better than the NN-based classification results (dotted lines) in all plots. Note also that the overall performance of HMIJ is usually worse than that of SMIJ. Such performance is expected since HMIJ does not capture the temporal ordering of the sequence of the ordered joints and therefore loses discriminability. For HMW, we see that classification results based on K-medoids clustering outperforms those based on K-means clustering.

Table 1(a) summarizes the best classification performances attained by different feature types on different datasets using different classification methods, all extracted from Fig. 5. Similarly, Table 1(b) shows the best classification performances extracted from Fig. 6. Note that the LDSP results are identical in both tables since LDSP features do not depend on partitioning of the skeleton data.[4] The pair of numbers in parenthesis noted together with the classification rates for SMIJ and HMIJ indicate the number of the most informative joints (N) and the number of segments (N_s), respectively, at which the corresponding classification results are achieved. Similarly, the number in parenthesis provided together with the classification rate for HMW indicate the number of segments at which the corresponding classification result is achieved. The best classification performance is obtained for different values of N for different number of segments (or for different length temporal windows) for different datasets, as shown in Tables 1(a) and 1(b). However, the best classification rates are achieved mostly when using 50 segments for the *fixed number of segments* approach or 40-ms windows for the *fixed temporal window* approach. This observation is also in accordance with our previous discussion about determining the temporal window size. On the other hand, there is a risk of over-fitting for the SMIJ representation since the number of classification parameters increases as N (or, the number of segments N_s) increases, while the amount of training data remains the same.

In general, the best classification results are obtained by the SMIJ representation for Berkeley MHAD and HDM05 datasets and by the HMIJ representation for the MSR Action3D dataset for both of the aforementioned segmentation methods. The SMIJ features perform worse than the other reference features in the MSR Action3D dataset due to its low frame rate at the capture time and more noisy skeleton data. More importantly, the classification results obtained by the *fixed temporal window* approach are in general better than the ones obtained by the *fixed number of*

[1] For the sake of exactness, we note that we do not compute the square of the distance function $D_{\chi^2}(\mathrm{H}^i, \mathrm{H}^j)$ when we compute the corresponding kernel since the χ^2 distance function already returns squared-distances by definition.

[2] Recall from Section 3.3 that the dimension of the histograms in the HMIJ representation depends on the number of joint angles in the associated skeleton structure.

[3] There are 20, 21 and 22 joint angles in the associated skeleton structures of the MSR Action3D, Berkeley MHAD and HDM05 databases, respectively.

[4] Recall from Section 3 that they are, on the contrary, obtained by LDS modeling of the entire joint angle time-series data as a whole.

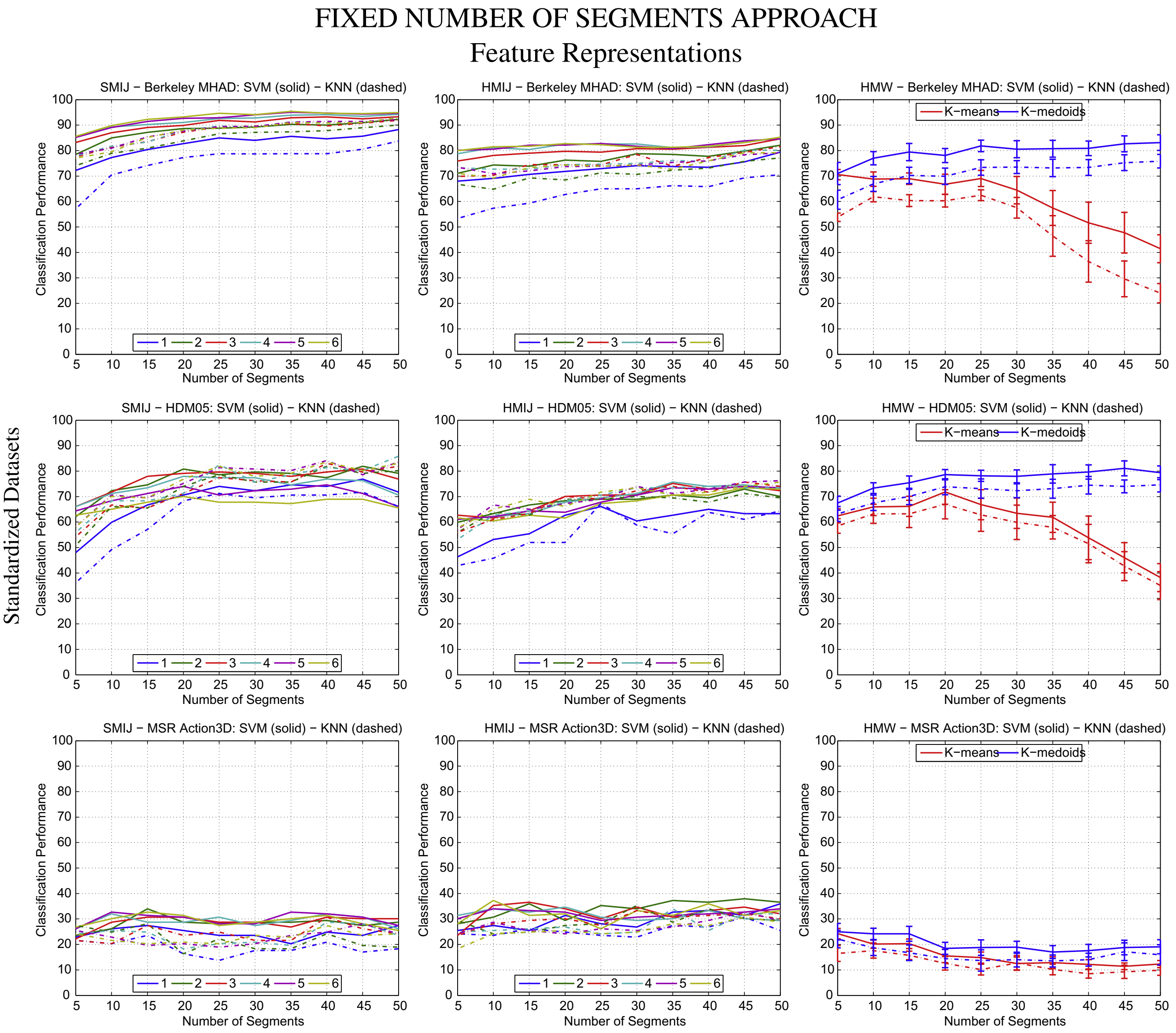

Fig. 5. Classification results for the *fixed number of segments* approach for data segmentation. The plots in different columns correspond to different feature representations, i.e., SMIJ, HMIJ and HMW, from left to right, respectively. The plots in different rows are based on different datasets, i.e., Berkeley MHAD, HDM05 and MSR Action3D, from top to bottom, respectively. In all plots, the vertical axis is classification performance and the horizontal axis is the number of segments, ranging from 5 to 50 with a step size of 5. For the SMIJ and HMIJ plots, different colors represent different number of the most informative joints (N) included in the representation. For the HMW plots, different colors represent different clustering methods (i.e., K-means and K-medoids). For all plots, the solid lines demonstrate the SVM-based classification results whereas the dotted lines demonstrate the NN-based classification results. Note that the HMW plots show the mean and standard deviation of the classification rates achieved over 20 runs as explained in Section 3.1.

segments approach. This observation confirms our hypothesis stated in Section 2.2 and renders the SMIJ representation based on the *fixed temporal window* segmentation method more flexible and easily applicable in practice.

There are still two critical factors that remain to be addressed among different databases in the future. One of the factors is the properties of the underlying acquisition systems used to collect the action data. Specifically, the Berkeley MHAD and HDM05 databases are obtained using accurate, high-frame rate motion capture systems from which it is possible to extract clean and smooth skeleton data. On the contrary, the MSR Action3D database is obtained by an early version of the Microsoft Kinect device, and therefore, the skeleton data extracted from the depth data is not as accurate or smooth as the skeleton data obtained from a motion capture system, and has low frame rate. Thus, the classification performance of any feature representation suffers from high noise existing in the skeleton data.

The other critical factor that remains to be addressed among different databases is the set of actions they contain, which eventually impacts the classification performance. In order to further investigate the reasons behind the poor performance of the SMIJ feature representation on the MSR Action3D dataset in addition to the noise effects, we examine the confusion matrix of the SVM-based classification which yielded the best performance as 33.33% using the SMIJ features on the standardized MSR Action3D dataset (with $N = 6$ and 200-ms window), presented in Table 1(b). The confusion matrix in Table 2 shows 0% recognition rate for 8 of the actions, out of 17, almost all of which are based on a basic single arm motion such as *high arm wave*, *horizontal arm wave*, *forward punch*, *tennis swing*, and such. The most

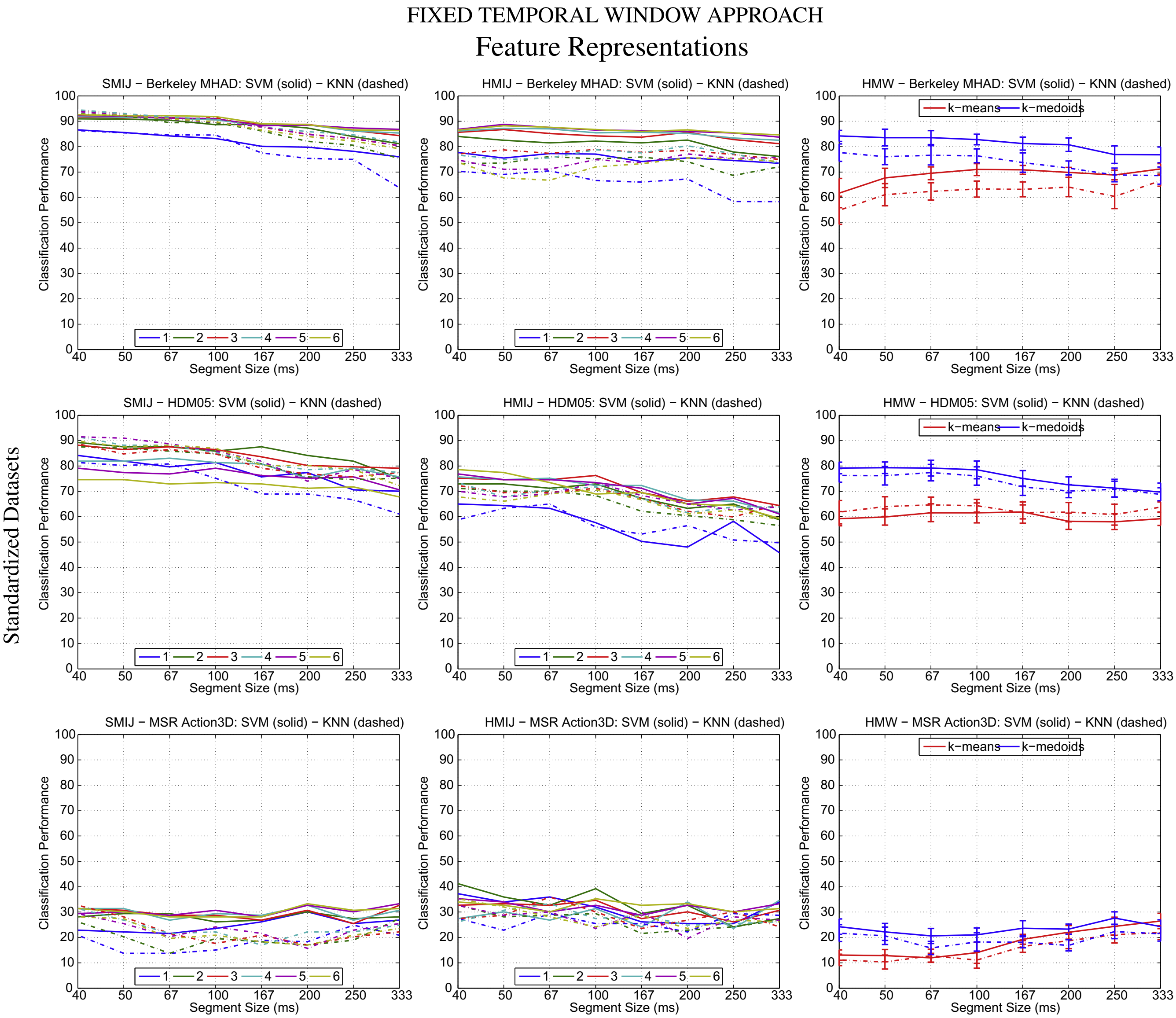

Fig. 6. Classification results for the *fixed temporal window* approach for data segmentation. The organization of the figure is the same as Fig. 5, except that the horizontal axis now represents temporal window size in milliseconds, taking on values from the set {40, 50, 67, 100, 167, 200, 250, 333}.

informative joints in all of these actions are the *elbow* and the *shoulder* of the corresponding arm. The proposed feature representation SMIJ is however a coarse representation of the action based on simple measures calculated from a set of spherical joint angles extracted from the skeleton structure. For instance, the SMIJ representation of waving your arm up in the air will be very similar to the SMIJ representation of swinging your arm along your body. Due to this reason, the first seven actions in the MSR Action3D dataset are classified as one of the more dominant actions such as *draw tick* or *draw circle*. If we exclude the 0% classification rates when we compute the overall classification performance of the SMIJ features, we see that the average classification rate is actually around 63%, which is almost double the initial classification rate.

For the sake of completeness, we also include the confusion matrices of the SVM-based classification which yielded the best performance using the SMIJ features for the standardized Berkeley MHAD database as 92.58% (with $N = 6$ and 40-ms temporal window), and for the HDM05 database as 89.27% (with $N = 2$ and 40-ms temporal window), in Tables 3 and 4, respectively. For the HDM05 database, the actions *grab high*, *hop both legs*, *throw basketball* and *throw* are the least distinguishable actions for the SMIJ features since the set of the most informative joints and the order of their activations along time are roughly the same. Similarly, for the Berkeley MHAD database, the most similar actions with respect to the SMIJ features are the *punch*, *wave two hands* and *throw* actions because these three actions depend on the motion of the two arms.

4.2.2. Action recognition across databases

In our second set of experiments, we tested the performance of the aforementioned feature representations in a cross-database recognition scenario, where a classifier is trained on one dataset and tested on another. Cross-database validation represents a challenging task that requires examining generalization of the proposed feature representations across different conditions of the data acquisition. Cross-database experimentation is often ignored by the community due to several open research questions. Recently, Torralba and Efros analyzed in [43] several examples of popular object recognition datasets and showed that training on specific data collections creates biased results which limit the per-

Table 1
The best action classification results for different feature representations obtained by different data segmentation approaches for the standardized datasets. The pair of numbers in parenthesis for SMIJ and HMIJ indicate the number of the most informative joints and the number of segments for the *fixed number of segments* approach, respectively, whereas they indicate the number of the most informative joints and the window size for the *fixed temporal window* approach, respectively. Similarly, the numbers in parenthesis for HMW indicate the number of segments for the *fixed number of segments* approach and the window size for the *fixed temporal window* approach. Numbers in bold highlight the maximum classification rates achieved in each column of the tables.

	Berkeley MHAD		HDM05		MSR Action3D	
	1-NN	SVM	1-NN	SVM	1-NN	SVM
(a) Fixed number of segments approach						
SMIJ	**93.10**	**95.37**	**85.88**	**81.92**	30.72	33.99
	(4/50)	(6/35)	(4/50)	(2/45)	(3/40)	(2/15)
HMIJ	81.57	82.57	76.27	75.71	**35.29**	**37.91**
	(3/50)	(6/50)	(5/50)	(4/35)	(2/30)	(2/45)
HMW	75.82	83.11	74.60	81.10	22.09	25.00
	(50)	(50)	(50)	(45)	(5)	(5)
LDSP	84.86	92.79	67.80	70.62	28.76	33.33
(b) Fixed temporal window approach						
SMIJ	**94.54**	92.58	**91.53**	**89.27**	32.68	33.33
	(4/40)	(6/40)	(4/40)	(2/40)	(3/40)	(6/200)
HMIJ	80.33	88.77	73.45	78.53	**35.95**	**41.18**
	(4/200)	(5/50)	(4/40)	(6/40)	(3/67)	(2/40)
HMW	77.66	84.23	77.37	79.32	22.32	27.65
	(40)	(40)	(67)	(50)	(250)	(250)
LDSP	84.86	**92.79**	67.80	70.62	28.76	33.33

formance of the object detection algorithms developed and evaluated against such datasets. Their cross-dataset object recognition experiments showed that many of the proposed algorithms do not generalize beyond the given dataset that they were initially demonstrated on. By including cross-database evaluation/experiments, we aim to show that the proposed SMIJ representation indeed captures the invariances in the human skeletal motion, and therefore, shows more resilience to the “dataset bias” with better cross-dataset generalization characteristics among other feature representations.

To pursue this task, we first determined a set of actions that are common in two of the three datasets we examined in this paper, namely the Berkeley MHAD and HDM05 datasets. We found five actions that were performed similarly: *jump, jumping jacks, throw, sit down,* and *stand up*. To accommodate for the differences in the skeleton structure between the two datasets, we determined a set of 16 joints that are common to both, as shown in Fig. 7.

The plots in the top row of Fig. 8 highlight the most informative 3 joints along the key frames of the *sit down* action taken from the Berkeley MHAD (left) and HDM05 (right). More detailed examination reveals that, except for the *preparation* phase at the beginning of the action, the most informative 3 joints between the two datasets match most of the time. Specifically, as the subject gets support from his arms to keep his balance before starting the action, the *arm* and *elbow* joints become the most informative. Then, the subject starts the sitting action by bending his knees and leaning backward towards the chair, rendering the *knee* joints the most informative. Finally, the subject brings his body into a balance position by moving his head backwards and putting his arms on his lap, making the *neck* as well as the *arm* and *elbow* joints the most informative.

The plots in the bottom row of Fig. 8 show the stacked histogram distributions of the most informative 3 joints for the same action. Despite some subtle differences, the distribution of the most informative 3 joints for the *sit down* action in both datasets show strong similarities. Specifically, joints 6 (*RElbow*), 8 (*LElbow*) rank around 40% of the time in the top three for both datasets. Similarly, joints 3 (*Neck*), 5 (*RArm*), 7 (*LArm*), 10 (*RKnee*) and 14 (*LKnee*) appear at least 20% of the time in the top three. Even though some of these joints reach the level of occurrence as high as 30% for the

Table 2
Confusion matrix for SVM classification of the MSR Action3D when $N = 6$ and the *fixed temporal window* length is 200 ms.

Actions	*high arm wave*	*horizontal arm wave*	*hammer*	*hand catch*	*forward punch*	*high throw*	*draw x*	*draw tick*	*draw circle*	*hand clap*	*two hand wave*	*side-boxing*	*forward kick*	*side kick*	*jogging*	*tennis swing*	*tennis serve*
1	0.0							11.11	44.44			22.22				22.22	
2		0.00						22.22	55.56							22.22	
3			0.00					22.22	44.44			22.22				11.11	
4				0.00				22.22	22.22			44.44					11.11
5					0.00			33.33	22.22			44.44					
6						0.00		44.44	11.11			11.11				33.33	
7							0.00	22.22	44.44							33.33	
8								66.67	22.22			11.11					
9								11.11	66.67							22.22	
10										77.78	22.22						
11										33.33	33.33						33.33
12												55.56					44.44
13													100.00				
14													100.00	0.00			
15															88.89		11.11
16								55.56	11.11			11.11				22.22	
17									22.22			11.11				11.11	55.56
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17

Please cite this article in press as: F. Ofli et al., Sequence of the most informative joints (SMIJ): A new representation for human skeletal action recognition, J. Vis. Commun. (2013), http://dx.doi.org/10.1016/j.jvcir.2013.04.007

Table 3
Confusion matrix for SVM classification of the Berkeley MHAD when $N = 6$ and the *fixed temporal window* length is 40 ms.

Actions	*jump*	*jumping jacks*	*bend*	*punch*	*wave one hand*	*wave two hands*	*clapping*	*throw*	*sit down*	*stand up*	*sit down/stand up*
1	99.23	0.77									
2		100.00									
3			100.00								
4				81.60	1.60			16.80			
5			10.40	3.20	84.80			1.60			
6			0.80			80.80		18.40			
7							100.00				
8			4.00					96.00			
9									100.00		
10									8.00	92.00	
11		4.00	4.00								92.00
	1	2	3	4	5	6	7	8	9	10	11

Table 4
Confusion matrix for SVM classification of the HDM05 when $N = 2$ and the *fixed temporal window* length is 40 ms.

Actions	*deposit floor*	*elbow to knee*	*grab high*	*hop both legs*	*jog*	*kick forward*	*lie down floor*	*rotate both arms backward*	*sneak*	*squat*	*throw basketball*	*jump*	*jumping jacks*	*throw*	*sit down*	*stand up*
1	75.00									25.00						
2		100.00														
3			100.00													
4	16.67		33.33	50.00												
5					100.00											
6						100.00										
7	10.00						80.00		10.00							
8								100.00								
9									100.00							
10										100.00						
11			50.00								50.00					
12												100.00				
13													100.00			
14	16.67		66.67											16.67		
15			20.00												80.00	
16	10.00															90.00
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16

Berkeley MHAD dataset, they are somewhat suppressed in the HDM05 dataset. There are also some interesting differences between these two distribution plots. For example, joint 4 (*Head*) almost never appears in the top three ranking in the HDM05 dataset whereas it appears around 20% of the time in the top three in the Berkeley MHAD dataset. On the other hand, joints 12 (*RFoot*) and 16 (*LFoot*) almost never appear in the top three in the Berkeley MHAD dataset even though they rank around 20% of the time in the top three in the HDM05 dataset. These differences suggest that the underlying skeleton structures in the two datasets have different levels of sensitivity and noise for different joints that affect the set of the most informative joints and their temporal orderings. Even the instructions given to the subjects on how to perform an action lead to *style* variations in the performed action, resulting in differences in the set of the most informative joints or in their temporal orderings.

Finally, we demonstrate the performance of the aforementioned features in an action classification setting in which the classifiers are trained on one dataset and tested on another dataset for the set of common actions mentioned above. For this purpose, we used the same 7 Berkeley MHAD training subjects from the first set of experiments to model the action classifiers and all 5 HDM05 subjects to test the trained classifiers. In order to compare the cross-database classification results with the same-database classification results, we also considered the train/test splits used in the first set of experiments for the Berkeley MHAD and HDM05. To keep the discussion concise, we fix the number of the most informative joints included in the SMIJ and HMIJ representations to $N = 4$, and employ the *fixed temporal window* approach with 167-ms window. The same window size is also used for the HMW features. Note that we choose N and the temporal window size with the average action classification results in Fig. 6 not to favor any of the feature representations.

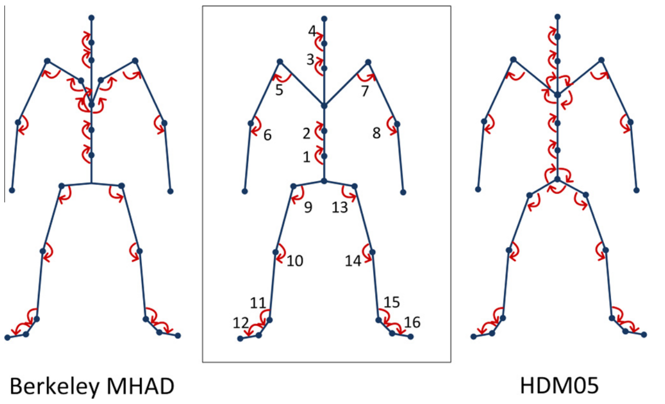

Fig. 7. The set of joint angles that are common in skeletal structures of both Berkeley MHAD and HDM05 databases are shown on the skeletal figure in the middle.

Table 5 shows the classification results for the cross-database generalization experiments. Our first observation is that, as expected, all feature representations perform better when trained and tested on the same dataset (i.e., 7 MHAD/5 MHAD and 3

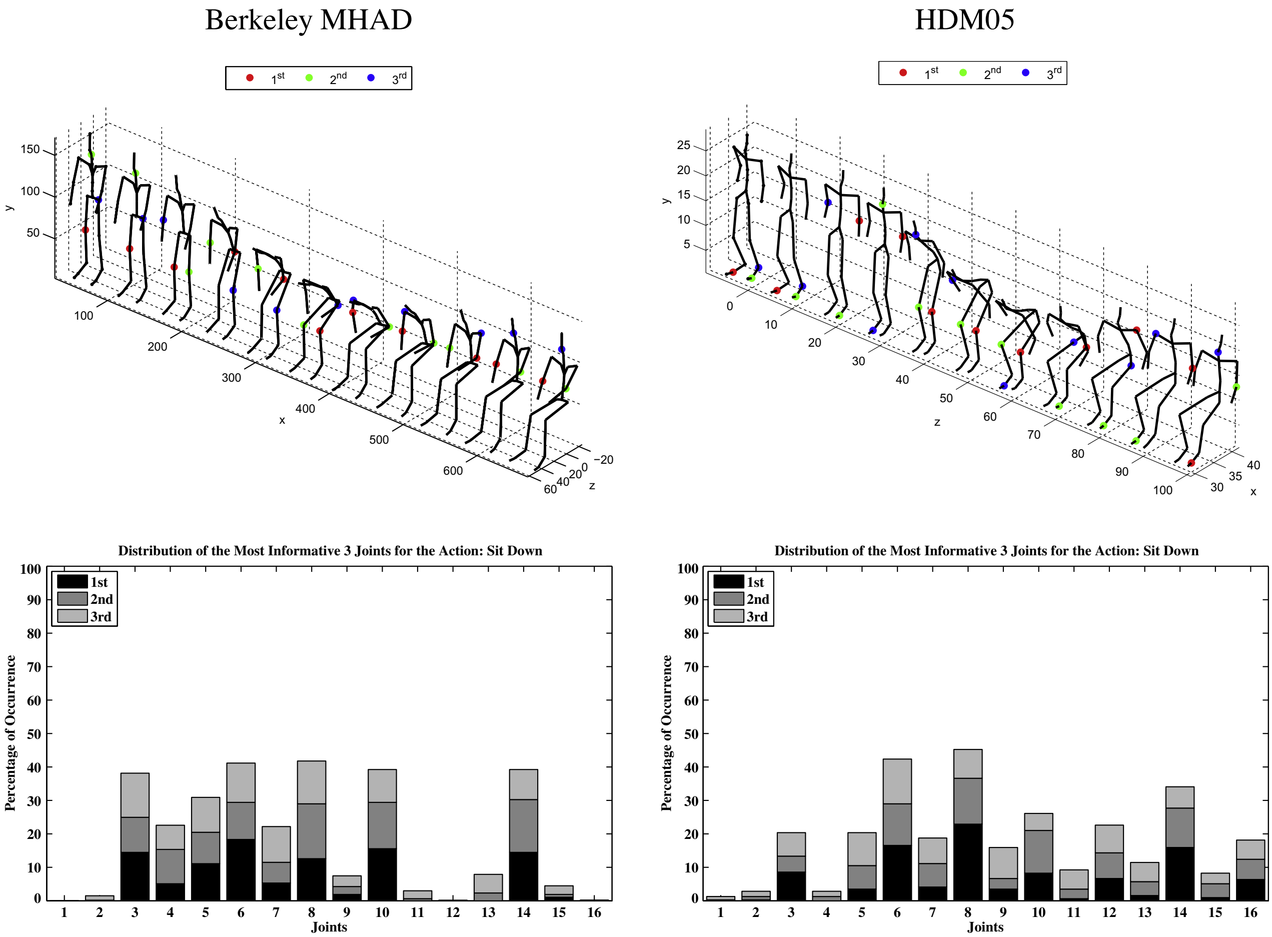

Fig. 8. The top row highlights the most informative 3 joints along the key frames of the *sit down* action taken from the Berkeley MHAD (left) and HDM05 (right). The bottom row shows the corresponding stacked histogram distributions of the most informative 3 joints of the same action in different databases.

Table 5
Classification performances of different feature representations in different train/test scenarios for the set of common actions in the Berkeley MHAD and HDM05 (i.e., *jump, jumping jacks, throw, sit down,* and *stand up*). Numbers in bold highlight the maximm classificaion rates achieved in each column.

	7 MHAD/ 5 HDM05		7 MHAD/ 5 MHAD		3 HDM05/ 2 HDM05	
	1-NN	SVM	1-NN	SVM	1-NN	SVM
SMIJ	**85.21**	**88.03**	**93.10**	**95.86**	**89.71**	**98.53**
HMIJ	70.42	70.42	73.79	83.45	77.94	83.82
HMW	68.31	71.87	75.97	80.44	86.54	89.71
LDSP	76.06	85.92	88.28	95.17	86.76	94.12

HDM05/2 HDM05). For cross-database generalization, Table 5 shows that the SMIJ representation yields the best classification results for the cross-database action recognition experiment among other feature representations by achieving 85.21% for NN-based classification and 88.03% for the SVM-based classification. To evaluate the change in the classification performance of a feature representation between training and testing on the same and different datasets, we compute their percentage drops. The resulting classification percentage drops are presented in Table 6. The average percentage drop is the smallest (with 8.08%) for the SMIJ representation. Finally, Table 7 shows the confusion matrix of the SVM-based classification using the SMIJ features for the cross-database action classification experiment. We see that classification results for all actions except *throw* are higher than 85%, reaching 95% for the *stand up* and 100% for the *jumping jacks* actions. The classification result for the *throw* action is the lowest due to *style* differences in this particular action, possibly due to different instructions given to the subjects on how to perform the action (i.e., to throw an object farther or closer) between the two databases. In summary, the proposed SMIJ representation outperforms other feature representa-

Table 6
Percentage drops for different features in cross-database action recognition experiments are shown for 7 MHAD/ 5 HDM05 vs. 7 MHAD/ 5 MHAD (labeled as 7 M/ 5H vs. 7 M/ 5 M) as well as 7 MHAD/ 5 HDM05 vs. 3 HDM05/ 2 HDM05 (labeled as 7 M/ 5H vs. 3H/ 2H). Average percentage drop for a particular feature representation is computed as the mean of percentage drops of the corresponding four comparisons.

	7 M/ 5H vs. 7 M/ 5 M		7 M/ 5H vs. 3H/ 2H		Average % drop
	1-NN	SVM	1-NN	SVM	
SMIJ	8.47	**8.17**	**5.02**	10.66	**8.08**
HMIJ	**4.57**	15.61	9.65	15.99	11.45
HMW	10.08	10.65	21.07	19.89	15.42
LDSP	13.84	9.72	12.33	**8.71**	11.15

Please cite this article in press as: F. Ofli et al., Sequence of the most informative joints (SMIJ): A new representation for human skeletal action recognition, J. Vis. Commun. (2013), http://dx.doi.org/10.1016/j.jvcir.2013.04.007

Table 7
Confusion matrix for SVM classification performance of the SMIJ features for the cross-database generalization experiment.

Actions	*jump*	*jumping jacks*	*throw*	*sit down*	*stand up*
1	86.11	13.89			
2		100.00			
3			42.86	35.71	21.43
4		10.00		85.00	5.00
5		5.00			95.00
	1	2	3	4	5

tions in the challenging cross-database action recognition experiment.

5. Conclusions

We have proposed a very intuitive and qualitatively interpretable skeletal motion feature representation, called sequence of the most informative joints (SMIJ). Unlike most feature representations used for human motion analysis, which rely on sets of parameters that have no physical meaning, the SMIJ representation has a very specific practical interpretation, i.e., the ordering of the joints by their informativeness and their temporal evolution for a given action. More specifically, in the SMIJ representation, a given action sequence is divided into a number of temporal segments. Within each segment, the joints that are deemed to be the most informative are selected. The sequence of such most informative joints is then used to represent an action.

In this paper, we extended our original work, [16], to provide more detailed description of the proposed feature representation and provided a comprehensive analysis of the recognition performance of different feature representations based on their action classification performance. We showed that the intuitive and qualitatively interpretable feature representation, SMIJ, performs better than the other reference feature representations (i.e., HMIJ, HMW and LDSP) in action recognition tasks on three different datasets in two different experimental settings. In addition, we demonstrated the power of the SMIJ feature representation in a cross-database experiment which resulted in relatively high recognition rates.

One of the limitations of the SMIJ representation that remains to be addressed is its insensitivity to discriminate different planar motions around the same joint. The joint angles are computed between two connected body segments in 3D spherical coordinates, thus capturing only a coarse representation of the body configuration. This limitation is apparent especially in the MSR Action3D dataset as shown from our experiments. The SMIJ representation could be further extended by using an alternative joint angle description (e.g., Euler angles, exponential maps) and choosing the measure of informativeness, $\mathcal{O}$, accordingly.

Acknowledgments

This work is supported in part by the European Research Council grant VideoWorld as well as the grants NSF 0941362, NSF 0941463, NSF 0941382 and ONR N000141310116.

References

[1] T.B. Moeslund, E. Granum, A survey of computer vision-based human motion capture, Computer Vision and Image Understanding (CVIU) 81 (2001) 231–268.

[2] T.B. Moeslund, A. Hilton, V. Krüger, A survey of advances in vision-based human motion capture and analysis, Computer Vision and Image Understanding (CVIU) 104 (2006) 90–126.

[3] P. Turaga, R. Chellappa, V.S. Subrahmanian, O. Udrea, Machine recognition of human activities: a survey, IEEE Transactions on Circuits and Systems for Video Technology 18 (2008) 1473–1488.

[4] J. Aggarwal, M. Ryoo, Human activity analysis: a review, ACM Computing Surveys 43 (2011) 16:1–16:43.

[5] P. Dollár, V. Rabaud, G. Cottrell, S. Belongie, Behavior recognition via sparse spatio-temporal features, in: Proceedings of IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS), 2005.

[6] I. Laptev, On space-time interest points, International Journal of Computer Vision (IJCV) 64 (2–3) (2005) 107–123.

[7] I. Laptev, M. Marszalek, C. Schmid, B. Rozenfeld, Learning realistic human actions from movies, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008.

[8] A. Bissacco, A. Chiuso, Y. Ma, S. Soatto, Recognition of human gaits, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2001, vol. 2, pp. 52–58.

[9] A. Bissacco, A. Chiuso, S. Soatto, Classification and recognition of dynamical models: the role of phase, independent components, kernels and optimal transport, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 29 (2007) 1958–1972.

[10] S. Ali, A. Basharat, M. Shah, Chaotic invariants for human action recognition, in: Proceedings of IEEE International Conference on Computer Vision (ICCV), 2007.

[11] J. Yamato, J. Ohya, K. Ishii, Recognizing human action in time-sequential images using hidden Markov model, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1992.

[12] J.K. Aggarwal, Q. Cai, Human motion analysis: a review, Computer Vision and Image Understanding (CVIU) 73 (1999) 90–102.

[13] G.W. Taylor, G.E. Hinton, S. Roweis, Modeling human motion using binary latent variables, in: Proceedings of Neural Information Processing Systems (NIPS), 2007.

[14] G.W. Taylor, G.E. Hinton, Factored conditional restricted Boltzmann machines for modeling motion style, in: Proceedings of International Conference on Machine Learning (ICML), 2009.

[15] T. Cover, j. Thomas, Elements of Information Theory, Wiley, 2006.

[16] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, R. Bajcsy, Sequence of the most informative joints (SMIJ): a new representation for human skeletal action recognition, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2012, pp. 8–13.

[17] A. Marzal, E. Vidal, Computation of normalized edit distance and applications, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 15 (1993) 926–932.

[18] P. Beaudoin, S. Coros, M. van de Panne, P. Poulin, Motion-motif graphs, in: Proceedings of the 2008 ACM SIGGRAPH/Eurographics Symposium on Computer Animation (SCA), 2008, pp. 117–126.

[19] M. Müller, A. Baak, H.-P. Seidel, Efficient and robust annotation of motion capture data, in: Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation (SCA), 2009, pp. 17–26.

[20] J. Barbič, A. Safonova, J.-Y. Pan, C. Faloutsos, J.K. Hodgins, N.S. Pollard, Segmenting motion capture data into distinct behaviors, in: Proceedings of Graphics Interface (GI), 2004, pp. 185–194.

[21] A. López-Méndez, J. Gall, J. Casas, L.V. Gool, Metric learning from poses for temporal clustering of human motion, in: Proceedings of British Machine Vision Conference (BMVC), 2012, pp. 49.1–49.12.

[22] F. Zhou, F.D. la Torre, J.K. Hodgins, Hierarchical aligned cluster analysis for temporal clustering of human motion, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 99 (2012) 1–15.

[23] C. Shannon, W. Weaver, The Mathematical Theory of Communication, University of Illinois Press, 1949. vol. 1.

[24] F.J. Damerau, A technique for computer detection and correction of spelling errors, Communications of the ACM 7 (1964) 171–176.

[25] V.I. Levenshtein, Binary codes capable of correcting deletions, insertions and reversals, Soviet Physics Doklady 10 (1966) 707–710.

[26] R.A. Wagner, M.J. Fischer, The string-to-string correction problem, Journal of the ACM 21 (1974) 168–173.

[27] R.A. Wagner, R. Lowrance, An extension of the string-to-string correction problem, Journal of the ACM 22 (1975) 177–183.

[28] C. Dance, J. Willamowski, L. Fan, C. Bray, G. Csurka, Visual categorization with bags of keypoints, in: Proceedings of European Conference on Computer Vision (ECCV), 2004.

[29] H.E. Cetingul, R. Chaudhry, R. Vidal, A system theoretic approach to synthesis and classification of lip articulation, in: Proceeding of International Workshop on Dynamic Vision, 2007.

[30] R. Chaudhry, A. Ravichandran, G. Hager, R. Vidal, Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009, pp. 1932–1939.

[31] R. Vidal, A. Chiuso, S. Soatto, Application of hybrid system identification in computer vision, in: Proceedings of the European Control Conference, 2007, pp. 27–34.

[32] P.V. Overschee, B.D. Moor, N4SID: Subspace algorithms for the identification of combined deterministic-stochastic systems, Automatica, Special Issue in Statistical Signal Processing and Control 30 (1994) 75–93.

[33] R. Shumway, D. Stoffer, An approach to time series smoothing and forecasting using the EM algorithm, Journal of Time Series Analysis 3 (1982) 253–264.

[34] G. Doretto, A. Chiuso, Y. Wu, S. Soatto, Dynamic textures, International Journal of Computer Vision (IJCV) 51 (2003) 91–109.

[35] K.D. Cock, B.D. Moor, Subspace angles and distances between ARMA models, System and Control Letters 46 (2002) 265–270.

[36] B. Afsari, R. Chaudhry, A. Ravichandran, R. Vidal, Group action induced distances for averaging and clustering linear dynamical systems with applications to the analysis of dynamic visual scenes, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012.

[37] S. Vishwanathan, A. Smola, R. Vidal, Binet–Cauchy kernels on dynamical systems and its application to the analysis of dynamic scenes, International Journal of Computer Vision (IJCV) 73 (2007) 95–119.

[38] A. Chan, N. Vasconcelos, Probabilistic kernels for the classification of auto-regressive visual processes, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2005, vol. 1, pp. 846–851.

[39] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, R. Bajcsy, Berkeley MHAD: A comprehensive multimodal human action database, in: Proceedings of IEEE Workshop on Applications of Computer Vision (WACV), 2013.

[40] M. Müller, T. Röder, M. Clausen, B. Eberhardt, B. Krüger, A. Weber, Documentation Mocap Database HDM05, Technical Report CG-2007-2, Universität Bonn, 2007.

[41] W. Li, Z. Zhang, Z. Liu, Action recognition based on a bag of 3D points, in: Proceedings of Computer Vision and Pattern Recognition Workshops (CVPRW), 2010, pp. 9–14.

[42] J. Zhang, M. Marszalek, S. Lazebnik, C. Schmid, Local features and kernels for classification of texture and object categories: a comprehensive study, International Journal of Computer Vision (IJCV) 73 (2007) 213–238.

[43] A. Torralba, A. Efros, Unbiased look at dataset bias, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011, pp. 1521–1528.

Please cite this article in press as: F. Ofli et al., Sequence of the most informative joints (SMIJ): A new representation for human skeletal action recognition, J. Vis. Commun. (2013), http://dx.doi.org/10.1016/j.jvcir.2013.04.007