Sequence of the Most Informative Joints (SMIJ): A New Representation for Human Skeletal Action Recognition

Ferda Ofli[1], Rizwan Chaudhry[2], Gregorij Kurillo[1], René Vidal[2] and Ruzena Bajcsy[1]
[1]Tele-immersion Lab, University of California, Berkeley
[2]Center for Imaging Sciences, Johns Hopkins University

Abstract

Much of the existing work on action recognition combines simple features (e.g., joint angle trajectories, optical flow, spatio-temporal video features) with somewhat complex classifiers or dynamical models (e.g., kernel SVMs, HMMs, LDSs, deep belief networks). Although successful, these approaches represent an action with a set of parameters that usually do not have any physical meaning. As a consequence, such approaches do not provide any qualitative insight that relates an action to the actual motion of the body or its parts. For example, it is not necessarily the case that clapping can be correlated to hand motion or that walking can be correlated to a specific combination of motions from the feet, arms and body. 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. The selection of joints is based on highly interpretable measures such as the mean or variance of joint angles, maximum angular velocity of joints, etc. We then represent an action as a sequence of these most informative joints. Our experiments on multiple databases show that the proposed representation is very discriminative for the task of 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*. [12, 13] and most recently by Aggarwal and Ryoo [1] for a comprehensive analysis). In general all methods use a mathematical representation of human motion and develop algorithms for comparing and classifying different instances of human activities under these representations. A common and intuitive method to represent human motion is to use a sequence of approximate human skeletal configurations. 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 provide very accurate skeletal configurations of human actions but are limited to only laboratory settings. Therefore, methods that relied heavily on accurate skeletal data slowly fell out of favor and most state-of-the-art activity recognition methods extract spatio-temporal interest points from monocular videos and learn their statistics [7, 8, 9]. Recently, with the release of several low-cost and relatively accurate 3D capturing systems, such as the Microsoft Kinect, 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 skeleton-based action representation.

Skeleton-based approaches for human activities were primarily focused on modeling the dynamics of either the full skeleton or a combination of limb parts. Majority of the methods use Linear Dynamical Systems (LDS) or Non-Linear Dynamical Systems (NLDS), *e.g*. in [4, 5, 3] or Hidden Markov Models (HMM), see *e.g*. the earlier work by Yamato *et al*. [17] and a review of several others in [2], to represent the dynamics of normalized 3D positions of joints or joint angle configurations. Recently Taylor *et al*. in [16, 15] 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, they represent human motion with a set of dynamics/observation parameters that, in general, do not have a qualitatively interpretable property.

A key observation that we make is that even though humans perform the same action differently, while generating dissimilar joint trajectories, the same set of joints, roughly in the same order, are activated during the performance of these actions. In our approach we take advantage of this observation to capture invariances in human skeletal motion in a given action. We propose finding the relative *informative-*

978-1-4673-1612-5/12/$31.00 ©2012 IEEE

ness of all the joints in a temporal window during an action. A joint is the most *informative* in a particular temporal window if, for example, it has the highest variance of motion as captured by the change in the joint angle. Such a notion of informativeness is very intuitive and interpretable. Furthermore, the ordered sequence of informative joints in a full skeletal motion implicitly models the temporal dynamics of the motion. In this paper, we therefore propose a new representation for human motion based on the sequence of the most informative joints. We compare the performance of this representation to several holistic action representations, based on the histograms of motion words, as well as the methods that explicitly model the dynamics of the skeletal motion. We will 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 the *skeleton*. Different joint configurations produce different skeletal poses and a time series of these poses yields the skeletal motion. An action can be simply 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 view- and scale-invariance.

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.*, $\mathbf{a}^i = \{a_t^i\}_{t=1}^{t=T}$ 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.*, $A = \left[\mathbf{a}^1\ \mathbf{a}^2\ \ldots\ \mathbf{a}^J\right]$, where J is the number of joints in the skeleton hierarchy. Hence, A is the $T \times J$ matrix of joint angle time series representing an action sequence.

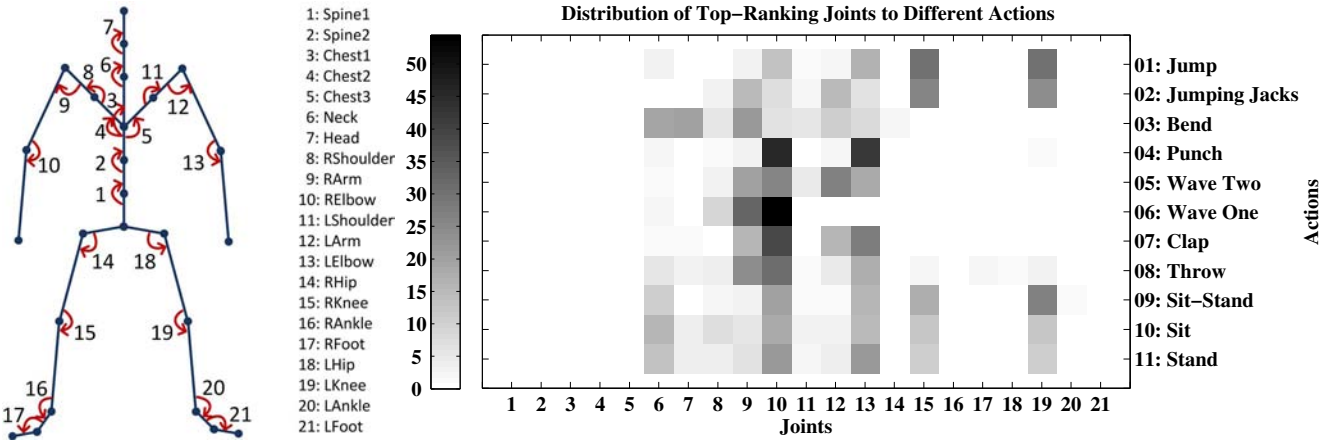
Common modeling methods such as LDS or HMM model the evolution of the time series of joint angles. However, instead of directly using the original joint angle time-series data A, one can also extract various types of features from A such as the *mean* or *variance* of joint angle time series, or the *maximum angular velocity* of each joint. For the sake of generality, we will denote this operation with $\mathcal{O}$ in the remainder of the 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 or across smaller segments of the time-series data. 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.

Our hypothesis presented in this paper is the following: Different actions require human 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, in other words, the invariant temporal structure of the action itself.

In order to visualize this phenomenon, let us consider the labeled joint angle configuration shown in Figure 1(a), and perform a simple analysis on Dataset #1, (see Section 3.2 for details about the datasets). The analysis is based on the following steps: i) partition an action sequence into a number of congruent segments, ii) compute the *variance* of the joint angle time series of each joint over each temporal segment (note that $\mathcal{O}$ 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.

Figure 1(b), shows the distribution of the *top-ranking* joints for different actions. We can see that some actions engage only a few joints (*e.g.*, actions 4 (*punch*) and 6 (*wave one*)) whereas other actions engage more joints. Nevertheless, the set of the most engaged joints are different for different actions. Joint 10 (*RElbow*) is the top-ranking joint 54% of the time, followed by joint 9 (*RArm*) 33% of the time in action 6 (*wave one*). Both joints 10 (*RElbow*) and 13 (*LElbow*) are the top-ranking joints more than 40% of the time in action 4 (*punch*). On the other hand, almost half of the joints appear in the top-ranking position at some point in actions 9 (*sit-stand*), 10 (*sit*) and 11 (*stand*); 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 action 9 (*sit-stand*) than in actions 10 (*sit*) and 11 (*stand*).

Figure 2 shows the histogram of the top-ranking 5 joints for four different actions. While the differences in the distribution of 1^{st}-, 2^{nd}-, or 3^{rd}-ranking joints, and so on, for actions 4 (*punch*) and 6 (*wave one*) are evident, actions 1 (*jump*) and 2 (*jumping jacks*) require closer look at the histograms. Specifically, even though joints 15 (*RKnee*) and 19 (*LKnee*) appear more than 25% of the time as either the 1^{st}- or 2^{nd}-ranking joint for both actions 1 (*jump*) and 2 (*jumping jacks*), joints 10 (*RElbow*) and 13 (*LElbow*) tend to rank in the top three at least 10% of the time for action 1 (*jump*) whereas joints 9 (*RArm*) and 12 (*LArm*) tend to rank in the top three for action 2 (*jumping jacks*). In short, different sets of joints reveal discriminative information about the underlying structure of the action. This is precisely the main observation that motivates us to consider sequences of the top N most informative joints as a new feature repre-



(a) (b)

Figure 1. (a) The structure of the skeleton used in Dataset#1 and corresponding set of 21 joint angles computed from it. (b) Distribution of the top-ranking joints, *i.e*., the most engaged joints, to different actions obtained for Dataset #1. Nonzero entries in a row show the set of joints that are engaged the most for the corresponding action. Some actions (such as 4 (*punch*) and 6 (*wave one*)) can be described easily by only a few number of joints whereas some actions (such as 9 (*sit-stand*), 10 (*sit*) and 11 (*stand*)) require many more joints.

sentation for human skeletal action recognition.

The new feature representation, which we call 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 evolution of the set of the most informative joints over all of the time segments. To extract this representation from the time-series data, we first partition the action sequence into N_s temporal segments and compute $\mathcal{O}$ over each segment. Let $\mathbf{a}_k^i = \{a_k^i\}_{k=t_s^k,\ldots,t_e^k}$ be a segment of $\mathbf{a}^i$ where t_s^k and t_e^k denote the start and the end frames for the segment k, respectively. Then, an action sequence is written as a collection of features, $\mathbf{F} = \{\mathbf{f}_k\}_{k=1,\ldots,N_s}$ where

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

The feature function, $\mathcal{O}(\mathbf{a}_k^i)$, provides a measure of information (*e.g*. the mean or variance of joint angles, or the maximum angular velocity) of the joint i in the temporal segment $\mathbf{a}_k$. We then rank-order all the joints in $\mathbf{f}_k$ based on the value of $\mathcal{O}$ and define SMIJ features as

$$\text{SMIJ} = \{\{\text{idof}\left(\text{sort}\left(\mathbf{f}_k\right), n\right)\}_{k=1,\ldots,N_s}\}_{n=1,\ldots,N}, \tag{2}$$

where the sort operator *sorts* the joints based on their local $\mathcal{O}$ score in descending order, the $\text{idof}\ (\cdot, n)$ operator returns the *id of* a joint that ranks n^{th} in the joint ordering, and N specifies the number of top-ranking joints included in the representation. 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). The resulting feature descriptor is $N_s \times N$-dimensional.

Metrics for SMIJ Since the proposed representation is a set of sequences over a fixed alphabet - the joints, we use the Levenshtein distance, $D_L(S_i, S_j)$, [10] for comparing the SMIJ features from two different sequences, S_i and S_j. The Levenshtein distance measures the amount of difference between two sequences of symbols such as strings. It is defined as the minimum number of operations required to transform one sequence into the other where the allowable operations are *insertion*, *deletion*, or *substitution* of a single symbol. We use a normalized version of the Levenshtein distance,

$$\bar{D}_L(S_i, S_j) = \frac{D_L(S_i, S_j)}{N_s \times N}, \tag{3}$$

where N_s is the number of time segments as defined in Equation 2. This allows us to compare pairs of distances that have been computed between two pairs of sequences that have different lengths. For example, normalization allows us to say that $\bar{D}_L(\text{abab}, \text{abad}) = \frac{1}{4} < \bar{D}_L(\text{ab}, \text{ad}) = \frac{1}{2}$, whereas the un-normalized version would give the same distance of 1 for both pairs of sequences.

The size of the SMIJ feature depends on the number of segments N_s, which depends on how the action sequence is partitioned. The Levenshtein distance between two sequences is at least equal to the difference in lengths of the two sequences. Since we require a distance of zero when two actions have the same rank-ordering, irrespective of their actual temporal length, one natural choice is to fix N_s to a constant value for all the action sequences.

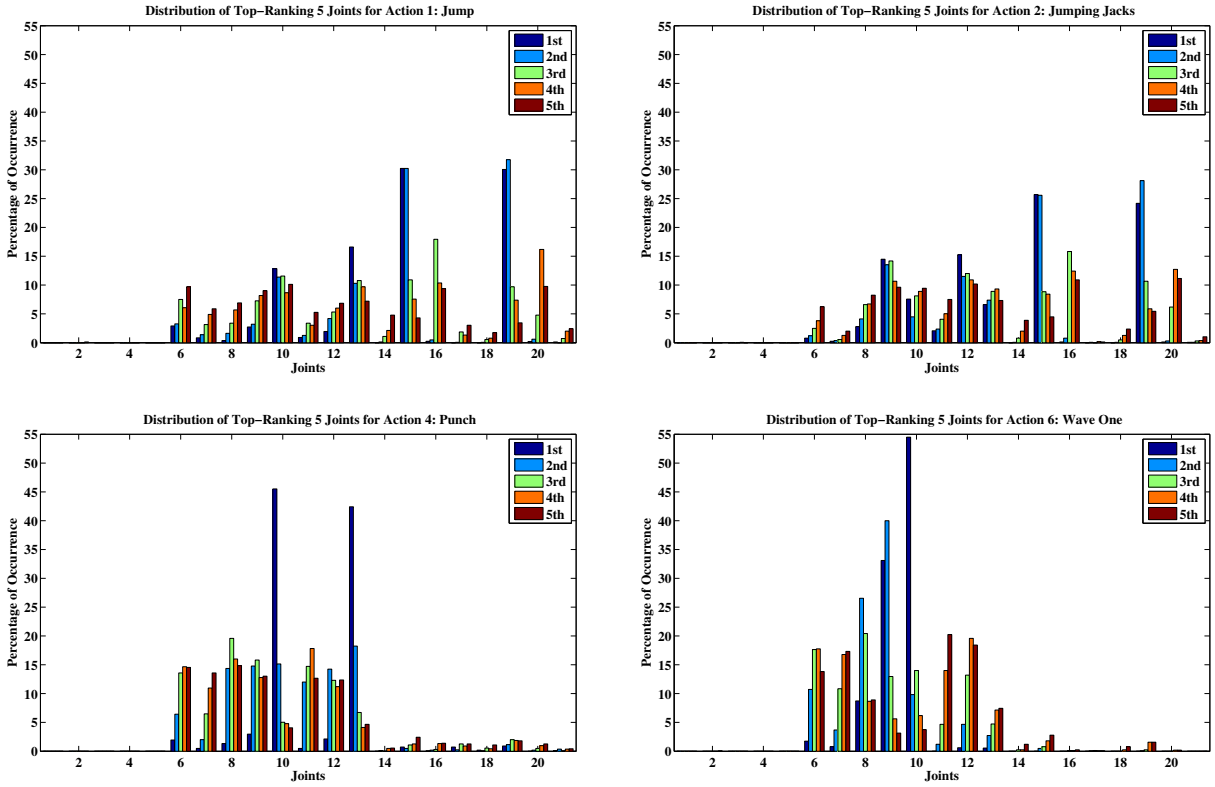


Figure 2. Histogram distribution of the top-ranking 5 joints for four actions selected from Dataset #1.

3. Evaluation of Feature Representations

In this section, we will compare the proposed Sequence of the Most Informative Joints (SMIJ) features against several other standard feature representations using three different action classification datasets. Baseline feature representations are briefly mentioned in Section 3.1. Details of the datasets are given in Section 3.2. The classification methods used are described in Section 3.3. Finally the comparative results are presented in Section 3.4.

3.1. Baseline Feature Representations

We first propose an alternative feature representation based on a similar idea to SMIJ. Instead of stacking the top-ranking N joints from all temporal segments into a single sequence of symbols, while keeping the temporal order of the joints intact, we create histograms separately for the 1^{st}-ranking joints, 2^{nd}-ranking joints, and so on, from all temporal segments; and then concatenate them as a feature descriptor, called Histograms of Most Informative Joints (HMIJ), to represent the action sequence, *i.e.*,

$$\text{HMIJ} = \left\{\text{hist}\left(\left\{\text{idof}\left(\text{sort}\left(\mathbf{f}_k\right), n\right)\right\}_{k=1,\ldots,N_s}\right)\right\}_{n=1,\ldots,N} \tag{4}$$

where the hist operator creates a J-bin l_1-normalized histogram from the input joint sequence, resulting in $J \times N$-dimensional feature descriptor. It is important to note that HMIJ features ignore the temporal order of the top-ranking N joints, and hence, will be used as a reference to assess the importance of preserving the temporal order information in the feature representation (as SMIJ features do).

Another popular method for representing an action sequence is based on the bag-of-motion words model. We first cluster the set of $\mathbf{f}_k$s into K clusters (*i.e.*, motion words) using k-means or k-medoids and then count the number of motion words that appear in a particular action sequence, yielding the Histogram-of-Motion Words (HMW) representation.

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.* in [4], and use LDS parameters (LDSP) as an alternative feature representation.

Even though we do not provide an exhaustive list of all possible feature representations, we believe that these three feature representations, *i.e.*, HMIJ, HMW and LDSP, are comprehensive enough to demonstrate the power of the proposed SMIJ features in terms of discriminability and interpretability for human action recognition.

3.2. Datasets

We evaluate the performance of each feature representation described above on three different human action

datasets of 3D skeleton data. Each dataset has almost completely distinct set of actions with different frame rates, different skeleton extraction method, and hence, skeleton data of various quality.

Dataset #1: 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). 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 the faulty one). We then extracted the skeleton data by post-processing the 3D optical motion capture data. The actions lengths vary from 773 to 14565 frames. In our experiments, we used 7 subjects (384 action sequences) for training and 5 subjects (275 action sequences) for testing. The set of actions consisted of *jump, jumping jacks, bend, punch, wave one hand, wave two hands, clap, throw, sit down, stand up, sit down/stand up*.

Dataset #2: From the HDM05 database [14] we used 11 actions performed by 5 subjects. In this dataset, subjects performed each action with various number of repetitions, resulting in 251 action sequences in total. In addition to marker location captured with the frequency of 120 Hz, the HDM05 database also provides the corresponding skeleton data. The duration of the action sequences ranges from 121 to 901 frames. In our experiments, we used 3 subjects (142 action sequences) for training and 2 subjects (109 action sequences) for testing. 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*.

Dataset #3: We also evaluated the action recognition on the MSR Action3D dataset [11] 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 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 15 to 76 frames. We used 5 subjects (226 action sequences) for training and 3 subjects (153 action sequences) for testing. 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, tennis serve*.

3.3. Classification Methods

In this section we examine the quality of different feature representations by evaluating their classification performance using well-known methods such as 1-nearest neighbor (1-NN) and support vector machine (SVM). Since we are investigating feature descriptors with different characteristics, we need to select the distance metrics according to the feature representations. We use the Levenshtein distance as explained in Section 2 for classification based on SMIJ. We use the χ^2 distance for classification based on histogram feature representations HMIJ and HMW. Finally, we use the Martin distance [6] as a metric between dynamical systems for classification based on LDSP. For SVM based classification, we follow one-vs-one classification scheme and use Gaussian kernel $K(S_i, S_j) = e^{-\gamma D^2(S_i,S_j)}$ with an appropriate distance function $D(S_i, S_j)$ depending on the feature type listed above. 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 in [18].

As proposed in Section 2, we use $\mathcal{O}$ to be the *variance* operator for Datasets #1 and #2 since it provides a good measure of activation (energy) of the skeletal joints. However, since the frame rate of Dataset #3 is very low, computing the *variance* over only a few number of samples did not seem reasonable. Therefore, we defined $\mathcal{O}$ to be the *maximum angular velocity* of the joints for Dataset #3 (which is more informative than the *variance* over segments with just one or two frames). We choose $N_s = 60$ for Dataset #1, $N_s = 55$ for Dataset #2, and $N_s = 11$ for Dataset #3 when computing SMIJ and HMIJ features. For HMW, we choose $K = 20$. For LDSP, a system order of 5 was used.

3.4. Experimental Results

Table 1 shows the performance of the proposed SMIJ representation. Note that using $N > 1$ most informative joints is better than using only the single most informative joint. The best classification performance is obtained for different values of N for different datasets; specifically, 94.91% when $N = 6$ for Dataset #1, 84.40% when $N = 2$ for Dataset #2, and 33.33% when $N = 5, 6$ for Dataset #3. The result of the best performance when using an intermediate value of N is not unexpected. In general, as N increases, the proposed representation captures more and more information about the action being performed. At the same time, the number of classification parameters increases with N, while the amount of training data remains the same. Therefore, there is a risk of over-fitting when N is large.

Table 2 shows the classification results when using HMIJ for several values of N. Notice that the performance of HMIJ is in general worse than that of SMIJ. This is to be expected, because HMIJ does not capture the temporal order of the sequence of ordered joints and therefore loses discriminability.

Table 3 shows classification results for all four feature representations. We choose to compare against SMIJ using $N = 2$ and HMIJ using $N = 4$. In general the best classification results are obtained by SMIJ using $N = 2$ for the first two datasets. In Dataset #3, however, the classi-

SMIJ (N)	Dataset #1		Dataset #2		Dataset #3	
	1-NN	SVM	1-NN	SVM	1-NN	SVM
1	72.73	88.73	72.48	77.98	16.34	24.18
2	78.91	94.18	80.73	**84.40**	24.18	29.41
3	**82.18**	94.18	**81.65**	77.06	22.22	28.76
4	82.91	93.09	77.98	77.06	22.22	28.10
5	82.18	93.82	81.65	67.89	23.53	33.33
6	82.55	**94.91**	81.65	64.22	**28.10**	**33.33**

Table 1. Classification results for SMIJ for several values of N.

HMIJ (N)	Dataset #1		Dataset #2		Dataset #3	
	1-NN	SVM	1-NN	SVM	1-NN	SVM
1	61.82	67.64	58.72	70.64	16.99	22.88
2	69.09	76.00	75.23	77.98	25.49	32.68
3	71.64	81.82	80.73	82.57	26.14	29.41
4	**72.73**	**82.91**	**80.73**	**82.57**	26.14	29.41
5	69.82	82.55	80.73	78.90	25.49	**33.33**
6	72.00	80.73	78.90	77.06	**28.76**	32.68

Table 2. Classification results for HMIJ for several values of N.

	Dataset #1		Dataset #2		Dataset #3	
	1-NN	SVM	1-NN	SVM	1-NN	SVM
SMIJ	**78.91**	**94.18**	**80.73**	**84.40**	24.18	29.41
HMIJ	72.73	82.91	80.73	82.57	26.14	29.41
HMW	70.91	81.09	78.90	78.90	21.57	32.68
LDSP	69.45	82.18	72.48	76.15	**43.14**	**47.06**

Table 3. Classification results for several baseline representations.

fication with $N = 2$ performs worse due to a low frame rate and more noisy skeleton data. Segmenting short action sequences into even smaller ones results in temporal segments consisting of only a few frames. Consequently, the statistical information extracted over such short segments is heavily biased by the noise in the data, resulting in poor performance of the classification algorithm. On the other hand, LDSP performs better for Dataset #3 since training LDS using the *entire* sequence provides better mechanism for handling the noise in the data and capturing the global temporal dynamics of the entire action.

4. Conclusion

In this paper, we have proposed a very intuitive and qualitatively interpretable skeletal motion feature, called Sequence of the Most Informative Joints (SMIJ). Instead of just being a set of matrix-valued parameters with no physical meaning, SMIJ has a very specific practical interpretation: the ordered set of the most informative joints in each temporal window, ordered in time for a given action. In our experiments, we have shown that our simple and computationally efficient feature performs better than standard parametric models such as LDS in action recognition tasks on three different datasets. With the increasing popularity of real-time 3D acquisition systems with human skeleton extraction, we believe SMIJ is ideal for human activity recognition tasks in practical applications.

Acknowledgements. This work was supported in part by the grants NSF 0941362, NSF 0941463 and NSF 0941382.

References

[1] J. Aggarwal and M. Ryoo. Human activity analysis: A review. *ACM Comput. Surv.*, 43:16:1–16:43, April 2011.

[2] J. K. Aggarwal and Q. Cai. Human motion analysis: A review. *Computer Vision and Image Understanding*, 73:90–102, 1999.

[3] S. Ali, A. Basharat, and M. Shah. Chaotic invariants for human action recognition. In *IEEE Int. Conf. on Computer Vision*, 2007.

[4] A. Bissacco, A. Chiuso, Y. Ma, and S. Soatto. Recognition of human gaits. In *IEEE Conf. on Computer Vision and Pattern Recognition*, volume 2, pages 52–58, 2001.

[5] A. Bissacco, A. Chiuso, and 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*, 29(11):1958–1972, 2007.

[6] K. D. Cock and B. D. Moor. Subspace angles and distances between ARMA models. *System and Control Letters*, 46(4):265–270, 2002.

[7] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *VS-PETS*, October 2005.

[8] I. Laptev. On space-time interest points. *Int. Journal of Computer Vision*, 64(2-3):107–123, 2005.

[9] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2008.

[10] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 1966.

[11] W. Li, Z. Zhang, and Z. Liu. Action recognition based on a bag of 3d points. In *Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 9 –14, june 2010.

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

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

[14] M. Müller, T. Röder, M. Clausen, B. Eberhardt, B. Krüger, and A. Weber. Documentation mocap database hdm05. Technical Report CG-2007-2, Universität Bonn, June 2007.

[15] G. W. Taylor and G. E. Hinton. Factored conditional restricted boltzmann machines for modeling motion style. In *Int. Conf. on Machine learning*, 2009.

[16] G. W. Taylor, G. E. Hinton, and S. Roweis. Modeling human motion using binary latent variables. In *Neural Information Processing Systems*, 2007.

[17] J. Yamato, J. Ohya, and K. Ishii. Recognizing human action in time-sequential images using hidden markov model. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 1992.

[18] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. *Int. J. Comput. Vision*, 73(2):213–238, June 2007.