# Flexible Dictionaries for Action Classification

Michalis Raptis, Kamil Wnuk, Stefano Soatto

# Flexible Dictionaries for Action Classification

Michalis Raptis, Kamil Wnuk, and Stefano Soatto

University of California, Los Angeles
Los Angeles, CA 90095
{mraptis,kwnuk,soatto}@cs.ucla.edu

**Abstract.** We present a simple approach to action classification which constructs a vector quantization of primitive motions from time series data corresponding to relative limb position estimates. The temporal scale, mean, and shape of primitive motion trajectories are independently modeled, thus creating a *flexible dictionary of action primitives*. We then explore two inference techniques that leverage our action dictionary representation, and evaluate their performance on both motion capture and video benchmark data. Our results indicate that even simplistic algorithms can outperform significantly more sophisticated ones in existing benchmark datasets.

## 1 Introduction

Classification of human activity has become a core interest domain in recent years, with a wide range of applications including surveillance, human-machine interfaces and video indexing. Human actions can be particularly difficult to classify due to potentially large variability in the time dimension in addition to the usual difficulties associated with processing of images. However, the temporal evolution contains a significant amount of information, as illustrated by [1]. In this work, we focus on classification of events with distinct temporal signatures.

Our core hypothesis is that the most complex human actions can be decomposed into short primitive actions. A simple, yet representative, example is the motion of the legs during walking. This motion can be decomposed into two components: one when the foot is in contact with the ground and the other when it is swinging above the ground [2]. The decomposition of actions into elementary motions has been explored previously in the literature. Marr and Vaina [3] proposed a method to segment movements into pieces, [4] defined primitive actions in terms of linear dynamical systems, and [5] proposed a probabilistic action decomposition framework.

In contrast to some recent work that reasons in the spatio-temporal domain, we do not deal with pixel or feature-based appearance during motions deemed to be interesting [6,7,8]. Instead we focus on the source of such appearance variations: the deformation of the human body. However, we do not seek to explicitly identify dynamical systems that drive a particular human skeleton [4]. Instead we seek to learn primitives of the motion of human limbs through examples, and perform classification decisions using a dictionary-based representation.

In an attempt to capture the available temporal information, we represent actions as multi-dimensional time series. We capitalize on this representation to easily extract trajectories of simple body movements. By clustering such primitive trajectories, we seek to construct a nonparametric quantization of the physically possible primitive motions, as they appear in the actions that we are interested in classifying. Because primitive limb motions contain a large amount of variability in temporal scale (speed of the action), as well as their relative location with respect to the center of the body, we explicitly and independently quantize these informative components to more effectively represent and match primitive trajectories.

We pursue the above goals by constructing dictionaries of action primitives. The dictionaries are obtained by clustering windows extracted at interesting points in our time series representation of actions. We demonstrate the performance of our dictionaries by applying two types of inference techniques. Our results show that:

1. Both of our inference techniques discriminate actions with well above 90% accuracy on commonly used benchmark datasets, using only a simple low-dimensional parameterization of human motion that we can track over time in a fully automatic fashion.
2. When exact pose is unavailable, our approach is robust to using a very rough approximation (bounding box), allowing us to succeed where other pose-based classifiers require supervision in the pose tracking procedure.
3. When exact pose information is available, such as in motion capture data, we show that we outperform recent dynamics-based approaches by up to 8.33% mean classification accuracy.

### 1.1 Related Work

Human action recognition algorithms can be separated into two main classes. The first one is the holistic approach, where a collection of statistics is extracted from a video sequence and used for classification [9,10,11,8,12,13]. The majority of such works represent a video sequence as spatio-temporal words [6,14] without taking into account causal constraints. Consequently, they lack the ability to discriminate between actions that share common basic movements in different order. In an attempt to incorporate the order of movements, Nowozin et al. [7] encoded the relative temporal sequence of such spatio-temporal features by finding a discriminative subsequence, again without regard to how the features were generated.

The second general approach is model-based, where the motion is characterized by the parameters of a model that is fit to the data. Hidden Markov Models [15,16] and finite-state models [17,18] have been used to model the temporal variability of human motion. However, these methods lose the details necessary to infer the dynamics as a result of the coarseness of their representation. Others model the dynamics of human gaits using hybrid linear models [19,20,21,22,23],

which exhibit great generative power; however, their discriminative power decreases with increased model complexity. This tradeoff makes it difficult to classify actions such as dancing.

Others in the second class of approaches propose to design invariants of the motion sequence [24]. However, these methods are difficult to generalize and require many samples of each action.

Our work attempts to capitalize on the best of both techniques. With our flexible dictionaries of action primitives we create a nonparametric model of possible human motions. The unique element of our approach is that such a model allows us to use statistical inference techniques typically reserved for holistic action approaches.

### 1.2 Actions as Time Series

As noted in our introductory statements, in this work we take a less common view of actions. Instead of representing an action as a collection of spatio-temporal patches [6,7,8], we interpret human motion as a multi-dimensional time series that captures the deformations of the body during activity. In particular, we track the positions of limb endpoints relative to the center of the body. As in [24], these include the arms, legs, and head (only 5 points).

In the case of motion capture data, where positions are expressed in three dimensions, our chosen representation produces a 15 dimensional time series ($x, y, z$ for each limb endpoint). For video we do not attempt to estimate 3D pose. Instead we roughly capture body deformations in the image plane. Because extraction of time series from video can be dataset dependent, details are provided below in the experimental section.

It is important to note that our methods are not bound to the above representation and can be used with time series generated from video or motion capture data using any number of techniques.

## 2 Dictionary of Action Primitives

Throughout this paper we define an *action primitive* as a time series subsequence that encodes a single dimension of a commonly occuring deformation. Due to our chosen method for obtaining time series, in our case such a primitive actually corresponds to a projection of a simple limb trajectory onto a single axis. For example, for a 3D trajectory of a single limb the time series resulting from the $x$, $y$ and $z$ coordinates each have their own action primitives. Independent action primitive dictionaries are constructed for all dimensions. This, for example, means 15 dictionaries in the case of motion capture data.

In the following sections we explore three incrementally more complex dictionary building procedures and their associated strengths and weaknesses for particular tasks. Each successive technique incorporates an additional degree of invariance by explicitly modeling a particular informative characteristic of action primitives.

### 2.1 Construction of a Simple Dictionary

The most basic approach to dictionary construction is to simply extract and cluster time series sub-sequences of a fixed size. Such clustering must be done with care, as recent literature on time series classification has shown that failing to tailor the sub-sequence sampling procedure to the signal can introduce artifacts that overpower the input, producing sinusoids independent of the underlying data [25,26]. For this reason we sample the signal by extracting sub-windows only around extrema points, which achieves local translation invariance as illustrated in [27]. To counter the effects of noise when detecting extrema, a small initial smoothing is applied. Sampling at the extrema of all the time series captures all significant changes of motion direction within each dimension of an action.

Once sub-windows are extracted for all dimensions in all action samples, we quantize them independently for each dimension using $K$-means. The $L_1$ distance was used to compare sub-windows. As shown in Fig. 1(a), the representative power of this simple approach suffers because several dictionary elements are wasted on encoding the same primitive shape occurring at different amplitudes. Thus, in our next approach we seek to increase representative power by independently quantizing the mean of the sub-windows.

### 2.2 Increasing Representative Power by Mean Quantization

To achieve clustering of trajectories invariant to their average amplitude, we subtract the mean from all extracted sub-windows before quantizing them as above. The average amplitude of a primitive trajectory is likely to be informative for discriminating actions, as it captures the relative position of the limb during motion. Thus we quantize the subtracted means. Hence, in addition to one dictionary of primitive trajectories per dimension of the time-series, we also add quantization of the mean. Each window extracted from the time series composing an action can now be represented with a best matching canonized trajectory and an assignment to a particular bin of the quantized mean.

Fig. 1(b) shows that this approach generates a dictionary spanning a greater variability of trajectories, thus improving its representational and discriminative power. This dictionary construction approach was found to perform well in classification tasks. However, modeling the temporal scale of dictionary elements still proved essential for achieving best results.

### 2.3 Modeling Time Scale for Temporal Flexibility

Human actions can have large temporal variability. Even within a single action instance the same primitives may be performed at greatly varying speeds. Picking up an object, for example, may be performed very rapidly when the object is solid or very slowly when the object is fragile and caution is required. Therefore, to increase the accurracy with which the dictionary can capture our data it is

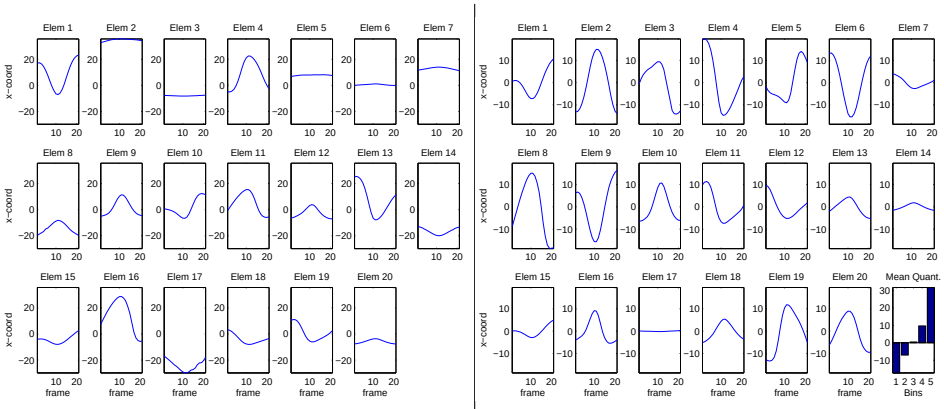



**Fig. 1.** The action primitive dictionaries constructed without, 1(a), and with, 1(b), mean quantization for the $x$-coordinate of the left hand in the FutureLight motion capture dataset. The dictionaries were constructed with $K$-means, using $K = 20$, and a sub-window size of 21. Notice that in 1(a) several clusters (elements 2, 5, 6, and 7 for example) capture the same trajectory, but at different mean locations. In addition to being inefficient, this means that when we represent a signal with this dictionary, much of the effort will go to matching the mean, diminishing the significance of the trajectory (shape). In 1(b), the mean was quantized (lower right) into 5 bins independent of the trajectory quantization. One can see, by comparing the two figures, that once the mean was modeled out from the trajectory quantization much more variability was captured in the 20 trajectory elements.

important to be able to represent action primitives at different time scales. For this reason, we extend our sub-window extraction procedure.

As before, the centers of time series sub-windows are selected at extrema points. However, instead of using a preset size we automatically select the size of each sub-window. This is done by symmetrically growing window borders until the nearest extremum is encountered.

Once the length of all sub-windows is known, we quantize this temporal scale using $K$-means. Each extracted sub-window is then matched to the closest quantized window size and resampled to this canonical size. Interpolation with B-splines was used to minimize loss during sub-window resampling. By this quantization of the time scale, we avoid comparing large scale phenomena with small scale trajectories, while maintaining invariance within local scale ranges.

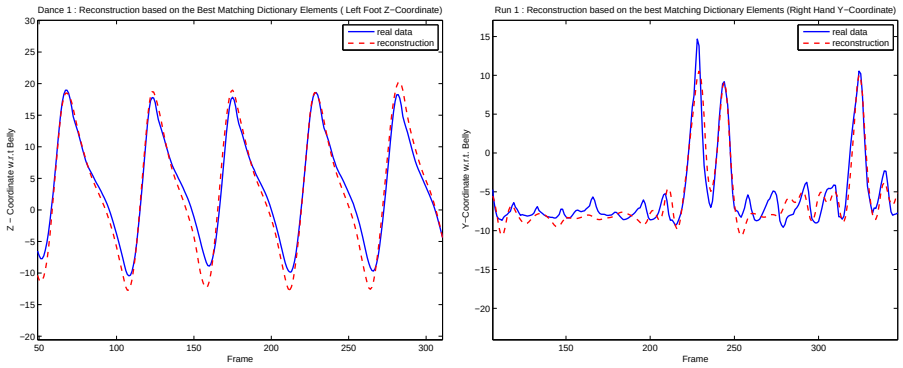
This procedure of rescaling should be understood in the context of achieving invariance with respect to reparameterization of the time axis, as an alternative to performing dynamic time warping when comparing two time series. However, both the extraction of invariants and comparing modulo reparameterization of

the temporal axis should be performed in a way that respects the dynamics of the underlying signal [27].

Following the rescaling, the mean quantization is performed as before, across all the extracted sub-windows. Finally, the clustering of the shape of the trajectories is performed independently within each canonical window size, thus creating a dictionary of action primitives spanning multiple time scales.

Adaptive window size selection does not only offer us a way of modeling the temporal variability, but also narrows down the importance of the choice of the window size, which needed to be manually specified in the previous approaches.

Moreover, as hypothesized, the power of the multi-scale dictionary becomes evident when applied to the reconstruction task, e.g. Fig. 2. The dictionary is able to capture the major components of variation of the time series. In order to generate the new canonical signal, a B-spline is fitted to the concatenated best matched dictionary elements.




(a) (b)

**Fig. 2.** Reconstruction of a time series representing a coordinate of the relative position of one of the limbs using the best matching dictionary elements. The original signal is shown in blue, and the approximation by the dictionary is shown in red. Fig. 2(a) shows a case where the quantization of the time scale, shape, and mean contains the correct elements needed for reconstruction. The example in Fig. 2(b) shows a case where several very rough approximations are made due to the coarseness of our quantization (mean $K = 5$, time scale $K = 5$ and shape at each scale $K = 4$).

## 3 Classification

We show that our dictionary of action primitives representation naturally lends itself to the action classification task, enabling a range of inference approaches. In

particular, we demonstrate two inference techniques: A bag-of-words approach, which ignores temporal constraints, and a string classification technique that explicitly takes into account the sequence in which action primitives occur.

### 3.1 Characterizing an Action

Once dictionaries are constructed for all dimensions, any action can be expressed by projecting each component time series onto the appropriate dictionary. The projection procedure follows a similar pipeline to dictionary construction. First, sub-windows are extracted in the same manner as during dictionary construction. Each sub-window is then matched to the closest quantized scale, mean, and trajectory in the given dictionary and assigned unique labels indicating the correspondences. For the simplest dictionaries only the trajectories are matched, and for the mean-quantized dictionaries only mean and trajectory are considered. In the case of multi-scale dictionaries the trajectory is resampled to the closest canonical size before comparing mean and shape. The result of projecting a time series onto a dictionary is thus a sequence of labels annotating all interest points.

### 3.2 Bagging Action Primitives

Our first classification approach discards all causal constraints. It is thus meant to serve as a baseline for more complex techniques. In this approach an action is characterized by comparing the distribution of dictionary elements within the whole action. This is known as the bag-of-words approach. It has shown to be very effective in the domains of document classification, category recognition in images, as well in behavior analysis [8,17]. To classify actions with this technique, we build a histogram of labels for each dimension of the time series. The dimensionality of the histogram for each time series is equal to the number of elements in the dictionary used. Each histogram bin records the number of times a particular dictionary element occurred during the action. For mean quantized dictionaries, we consider distributions of mean and shape in independent histograms. We found empirically that this independence assumption shows almost no difference from results given when performing inference with multi-dimensional histograms to couple the mean and trajectory labels. Each histogram is turned into a distribution independent of action duration via normalization. Finally, all histograms from each dimension are stacked to create a single action descriptor.

Once the stack of histograms characterizing each action is computed, we use a standard supervised SVM classification approach [28]. Given a partition of the action samples into labeled testing and training data, we train a multi-class $\chi^2$ kernel SVM on the histograms. Any action or action segment can then be classified with the SVM, once it is converted to the histogram of action primitives representation.

### 3.3 Exploiting Temporal Constraints

To incorporate temporal constraints into our inference, we map the problem of classifying time series to a problem of classifying strings [29,30]. By projecting the interest points of a time series to our dictionary, as described, we obtain a sequence of labels which can be viewed as a "string" representation of the time series. The alphabet elements are the elements of our dictionary. Our approach is motivated by algorithms from the domain of protein sequence similarity detection [31]. Our classification procedure can be described in three steps:

First, we compute the pairwise similarity between all the string sequences that describe each sample of an action. There are several techniques for string matching in the literature [32,33], that are commonly used to align proteins or nucleotide sequences. In our implementation, we used the Smith-Waterman (SW) algorithm; a dynamic programming algorithm which finds the local alignment of two string sequences while allowing gaps. In order for the similarity score to be independent of the length of the two sequences that we compare, we normalize the score by the length of the aligned sequences. Similarities computed for each dimension of time series of an action are equally weighted so as to obtain an overall similarity score for each action sample. We also experimented with jointly aligning all channels with respect to a single temporal warping, as opposed to independently aligning each channel. While this approach is physically more plausible (there is, after all, one temporal dimension), the resulting model achieves classification results that are slightly worse than those using independent alignment on each channel. This empirical finding remains to be fully understood.

The second step is to represent each action via its similarity to the actions in the training set. More specifically, the action $F$ is represented by a vector of scores:

$$X_F = [x_{f_1}, x_{f_2}, \ldots, x_{f_n}] \tag{1}$$

where $f_i$ is the $i$th training set sequence, and $x_{f_i}$ is the value of the SW score between sequence $F$ and $f_i$. This procedure can be viewed as the vectorization step of a kernel method approach, such as a SVM.

The third step is the definition of the positive definite kernel function:

$$K(X,Y) = \frac{X \cdot Y}{\sqrt{(X \cdot X)(Y \cdot Y)}} \quad . \tag{2}$$

This kernel $K(\cdot,\cdot)$ is then transformed into a radial basis kernel using the induced distance:

$$D(X,Y) = \sqrt{K(X,X) + K(Y,Y) - 2K(X,Y)} \tag{3}$$

Once we have the radial basis kernel, we train a multi-class SVM, and then classify actions.

## 4 Experiments

We applied both of our inference approaches to human motion classification in both motion capture and video data. Both fixed and adaptive window size

dictionaries were used. We show that when using motion capture data, the ideal case for our approach, we achieve state of the art performance, demonstrating 5 to 8% improvement in classification over recent dynamics-based approaches. We also show that our methods perform competitively in video, despite the availability of only a rough estimate of relative limb locations.

### 4.1 Datasets

As presented in [24], the FutureLight motion capture dataset [34] contains a total of 158 action samples which represent 5 actions, with a variable number of samples of each. The actions include dance, jump, run, sit and walk. Each action contains considerable intra-class variability, including several fairly ambiguous samples, such as a ballet dance that appears visually similar to a walk. As mentioned earlier, we choose to represent each recorded action with a 15 dimensional time series.

To demonstrate the applicability of our approach to video-based action classification, we evaluate performance on the Weizmann Human Action dataset provided by [10]. This dataset contains video of 9 different people, performing 9 unique actions, including: run, walk, sideways run, jump, jump in place, jumping jack, one-handed wave, two-handed wave, and bend. The dataset is annotated with silhouettes for each sample action, obtained using a background subtraction algorithm. Based on the silhouette properties we represent each action as a 5 dimensional time series. Specifically we split the silhouette into 4 quadrants at its center of mass and track the max width of the silhouette in each quadrant. We also track the absolute height of the silhouette over time. These features provide a very rough estimate of the positions of the body extrema over time and it is computationally more efficient than the spatio-temporal "cubes" that [10] used or the orientations of rectangular patches of the silhouette used by [17].

### 4.2 Results

Performance was evaluated on the datasets above using a leave-one-out cross validation procedure. Below we report the best results achieved for each combination of dictionary type and inference technique after exploring a range of various parameters:

**Classification Results**

| | **FutureLight** | **Weizmann** |
|---|---|---|
| our BAG+SVM | 95.30 | 91.36 |
| our SW+SVM | 96.04 | 97.52 |
| our BAG+Temporal Scale+SVM | 97.40 | 95.06 |
| our SW+Temporal Scale+SVM | **98.03** | **100** |
| [24] | 89.7 | 92.6 |
| [10] | – | **100** |

Surprisingly, when ignoring global temporal information and using the bag-of-words classification approach, we were able to achieve excellent classification

performance in both motion capture and video data. Using dictionaries with fixed window size (BAG+SVM) we obtained 95.3% mean classification accuracy on the FutureLight dataset (Table 1, window size: 21 samples) and 91.36% mean accuracy on the Weizmann action dataset (Table 5, window size: 17 samples). Using multi-scale dictionaries (BAG+Temporal Scale+SVM) we obtained 97.40% mean classification accuracy on the FutureLight dataset (Table 3), and 95.06% mean classification on the Weizmann dataset (Table 7).

Incorporating temporal constraints (SW) with our second inference method, we obtained improved performance in both datasets for fixed window and adaptive window size dictionaries. This is shown in Tables 2 and 4 for the FutureLight dataset and Tables 6 and 8 for the Weizmann dataset. For the FutureLight dataset performance increased by a little over a half percent, but for classification in the Weizmann dataset, we obtained a significant improvement of roughly 5 to 6%. The small increase in improvement in FutureLight is due to the already excellent performance of the bag-of-words approach. It is also worthy to note that these datasets do not contain actions where the global temporal order of the action is crucial.

| | Dance | Jump | Run | Sit | Walk |
|---|---|---|---|---|---|
| Dance | **28** | 1 | | | 2 |
| Jump | | **13** | | | 1 |
| Run | 1 | | **28** | | 1 |
| Sit | | | | **35** | |
| Walk | | | | | **48** |

Table 1. BAG+SVM

| | Dance | Jump | Run | Sit | Walk |
|---|---|---|---|---|---|
| Dance | **29** | 2 | | | |
| Jump | | **14** | | | |
| Run | | 4 | **26** | | |
| Sit | | | | **35** | |
| Walk | | | | | **48** |

Table 2. SW+SVM

| | Dance | Jump | Run | Sit | Walk |
|---|---|---|---|---|---|
| Dance | **28** | 1 | | | 2 |
| Jump | | **14** | | | |
| Run | | | **29** | | 1 |
| Sit | | | | **35** | |
| Walk | | | | | **48** |

Table 3. BAG+Temporal Scale+SVM

| | Dance | Jump | Run | Sit | Walk |
|---|---|---|---|---|---|
| Dance | **28** | 2 | | | |
| Jump | | **14** | | | |
| Run | | | **29** | | 1 |
| Sit | | | | **35** | |
| Walk | | | | | **48** |

Table 4. SW+Temporal Scale+SVM

| | A1 | A2 | A3 | A4 | A5 | A6 | A7 | A8 | A9 |
|---|---|---|---|---|---|---|---|---|---|
| A1 | **9** | | | | | | | | |
| A2 | | **9** | | | | | | | |
| A3 | | | **9** | | | | | | |
| A4 | | | | **9** | | | | | |
| A5 | | | | | **7** | | 2 | | |
| A6 | | | | | | **8** | | 1 | |
| A7 | | | | | | | **6** | 3 | |
| A8 | | | | | | | 1 | **8** | |
| A9 | | | | | | | | | **9** |

Table 5. BAG+SVM

| | A1 | A2 | A3 | A4 | A5 | A6 | A7 | A8 | A9 |
|---|---|---|---|---|---|---|---|---|---|
| A1 | **8** | 1 | | | | | | | |
| A2 | | **9** | | | | | | | |
| A3 | | | **9** | | | | | | |
| A4 | | | | **9** | | | | | |
| A5 | | | | | **9** | | | | |
| A6 | | | | | | **9** | | | |
| A7 | | | | | | | **9** | | |
| A8 | | | | | | | | **9** | |
| A9 | | | | | | | 1 | | **8** |

Table 6. SW+SVM

| | A1 | A2 | A3 | A4 | A5 | A6 | A7 | A8 | A9 |
|---|---|---|---|---|---|---|---|---|---|
| A1 | **9** | | | | | | | | |
| A2 | | **9** | | | | | | | |
| A3 | | | **9** | | | | | | |
| A4 | | | | **8** | 1 | | | | |
| A5 | | | | 1 | **8** | | | | |
| A6 | | | | | | **9** | | | |
| A7 | | | | | 1 | | **7** | 1 | |
| A8 | | | | | | | | **9** | |
| A9 | | | | | | | | | **9** |

Table 7. BAG+Temporal Scale+SVM

| | A1 | A2 | A3 | A4 | A5 | A6 | A7 | A8 | A9 |
|---|---|---|---|---|---|---|---|---|---|
| A1 | **9** | | | | | | | | |
| A2 | | **9** | | | | | | | |
| A3 | | | **9** | | | | | | |
| A4 | | | | **9** | | | | | |
| A5 | | | | | **9** | | | | |
| A6 | | | | | | **9** | | | |
| A7 | | | | | | | **9** | | |
| A8 | | | | | | | | **9** | |
| A9 | | | | | | | | | **9** |

Table 8. SW+Temporal Scale+SVM

A1: Run, A2: Walk, A3: Side, A4: Jump, A5: PJump, A6: Jack, A7: Wave 1, A8: Wave 2, A9: Bend

Confusion Matrices for FutureLight (Tables 1-4) and Weizmann (Tables 5-8) results.

## 5 Conclusion

This work proposed an approach for constructing dictionaries of primitive actions suitable for action classification from time series data. We explored the representative and discriminative power of these dictionaries using them in reconstruction and classification tasks, respectively.

We achieved state of the art classification results in the FutureLight motion capture dataset using both bag-of-words and string matching inference tech-

niques. Also, we demonstrated competitive results in video data, while using significantly simpler features with only 5 dimensions. Together, these results make a compelling argument for viewing actions from the perspective of multi-dimensional time series.

**Acknowledgments.** We would like to acknowledge the support of AFOSR FA9550-06-1-0138, ONR 67F-1080868 and NSF ECS-0622245.

## References

1. Johansson, G.: Visual perception of biological motion and a model for its analysis. percPsycho **14** (1973) 201–211
2. McGeer, T.: Passive Dynamic Walking. The International Journal of Robotics Research **9**(2) (1990) 62
3. Marr, D., Vaina, L.: Representation and Recognition of the Movements of Shapes. Proceedings of the Royal Society of London. Series B, Biological Sciences **214**(1197) (1982) 501–524
4. Bissacco, A., Chiuso, A., Ma, Y., Soatto, S.: Recognition of human gaits. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Volume 2. (December 2001) 52–57
5. Bregler, C.: Learning and recognizing human dynamics in video sequences. IEEE Conference on Computer Vision and Pattern Recognition (1997) 568–574
6. Niebles, J., Wang, H., Fei-Fei, L.: Unsupervised learning of human action categories using spatial-temporal words. Proc. BMVC (2006)
7. Nowozin, S., Bakir, G., Tsuda, K.: Discriminative subsequence mining for action classification. In: 11th IEEE International Conference on Computer Vision, Los Alamitos, CA, USA, IEEE Computer Society (10 2007) 1919–1923
8. Dollar, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on (15-16 Oct. 2005) 65–72
9. Zelnik-Manor, L., Irani, M.: Statistical analysis of dynamic actions. IEEE Trans. Pattern Anal. Mach. Intell. **28**(9) (2006) 1530–1535
10. Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R.: Actions as Space-Time Shapes. Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on **2** (2005)
11. Niebles, J., Fei-Fei, L.: A hierarchical model of shape and appearance for human action classification. Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on (17-22 June 2007) 1–8
12. Efros, A.A., Berg, A.C., Mori, G., Malik, J.: Recognizing action at a distance. In: IEEE International Conference on Computer Vision, Nice, France (2003) 726–733
13. Jhuang, H., Serre, T., Wolf, L., Poggio, T.: A biologically inspired system for action recognition. In: Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on. (2007) 1–8
14. Ikizler, N., Duygulu, P.: Human action recognition using distribution of oriented rectangular patches. In: Workshop on Human Motion. (2007) 271–284
15. Wilson, A.D., Bobick, A.F.: Parametric hidden markov models for gesture recognition. In: IEEE Trans. on Pattern Analysis and Machine Intelligence. Volume 21(9). (Sept. 1999) 884–900

16. Song, Y., Feng, X., Perona, P.: Towards detection of human motion. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2000). (2000) 810–817
17. Ikizler, N., Forsyth, D.: Searching video for complex activities with finite state models. Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on (17-22 June 2007) 1–8
18. Hongeng, S., Nevatia, R., Bremond, F.: Video-based event recognition: activity representation and probabilistic recognition methods. Comput. Vis. Image Underst. **96**(2) (2004) 129–162
19. Pavlovic, V., Rehg, J.M., MacCormick, J.: Learning switching linear models of human motion. In: NIPS. (2000) 981–987
20. Vidal, R., Soatto, S., Sastry, S.: An algebraic geometric approach to the identification of a class of linear hybrid systems. In: Proceedings of the IEEE Conference on Decision and Control. Volume 1. (December 2003) 167–172
21. Agarwal, A., Triggs, B.: Tracking articulated motion using a mixture of autoregressive models. In Pajdla, T., Matas, J., eds.: ECCV (3). Volume 3023 of Lecture Notes in Computer Science., Springer (2004) 54–65
22. Bissacco, A., Soatto, S.: Classifying human dynamics without contact forces. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Volume 2. (June 2006) 1678–1685
23. Li, R., Tian, T.P., Sclaroff, S.: Simultaneous learning of nonlinear manifold and dynamical models for high-dimensional time series. Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on (14-21 Oct. 2007) 1–8
24. Ali, S., Basharat, A., Shah, M.: Chaotic invariants for human action recognition. Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on (14-21 Oct. 2007) 1–8
25. Lin, J., Keogh, E., Truppel, W.: Clustering of streaming time series is meaningless. Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery (2003) 56–65
26. Fujimaki, R., Hirose, S., Nakata, T.: Theoretical analysis of subsequence time-series clustering from a frequency-analysis viewpoint. In: SDM, SIAM (2008) 506–517
27. Soatto, S.: On the distance between non-stationary time series. In: Modeling, Estimation and Control. Springer Verlag (2007)
28. Vapnik, V.N.: The nature of statistical learning theory. Springer-Verlag New York, Inc., New York, NY, USA (1995)
29. Teo, C.H., Vishwanathan, S.V.N.: Fast and space efficient string kernels using suffix arrays. In: ICML '06: Proceedings of the 23rd international conference on Machine learning, New York, NY, USA, ACM (2006) 929–936
30. Yang, C., Guo, Y., Sawhney, H.S., Kumar, R.: Learning actions using robust string kernels. In: Workshop on Human Motion. (2007) 313–327
31. Liao, L., Noble, W.S.: Combining pairwise sequence similarity and support vector machines for remote protein homology detection. In: RECOMB '02: Proceedings of the sixth annual international conference on Computational biology, New York, NY, USA, ACM (2002) 225–232
32. Smith, T.F., Waterman, M.S.: Identification of common molecular subsequences. Journal of Molecular Biology **147** (1981) 195–197
33. Needleman, S., Wunsch, C.: An efficient method applicable to the search for similarities in the amino acid sequences of two proteins. Journal of Molecular Biology **48**(3) (1970) 444–453
34. R&D division of Santa Monica Studios: FutureLight