Elsevier Editorial System(tm) for Image and Vision Computing
Manuscript Draft

Manuscript Number:

Title: Model-Based Recognition of Human Actions by Trajectory Matching in Phase Spaces

Article Type: Full Length Article

Keywords: action recognition; nonlinear dynamical systems; time-delay embeddings; trajectory matching; pose estimation; motion capture; pose estimation.

Corresponding Author: Mr. Adolfo López, Ph.D.

Corresponding Author's Institution: Technical University of Catalonia (UPC)

First Author: Adolfo López, Ph.D.

Order of Authors: Adolfo López, Ph.D.; Josep R Casas, Doctor

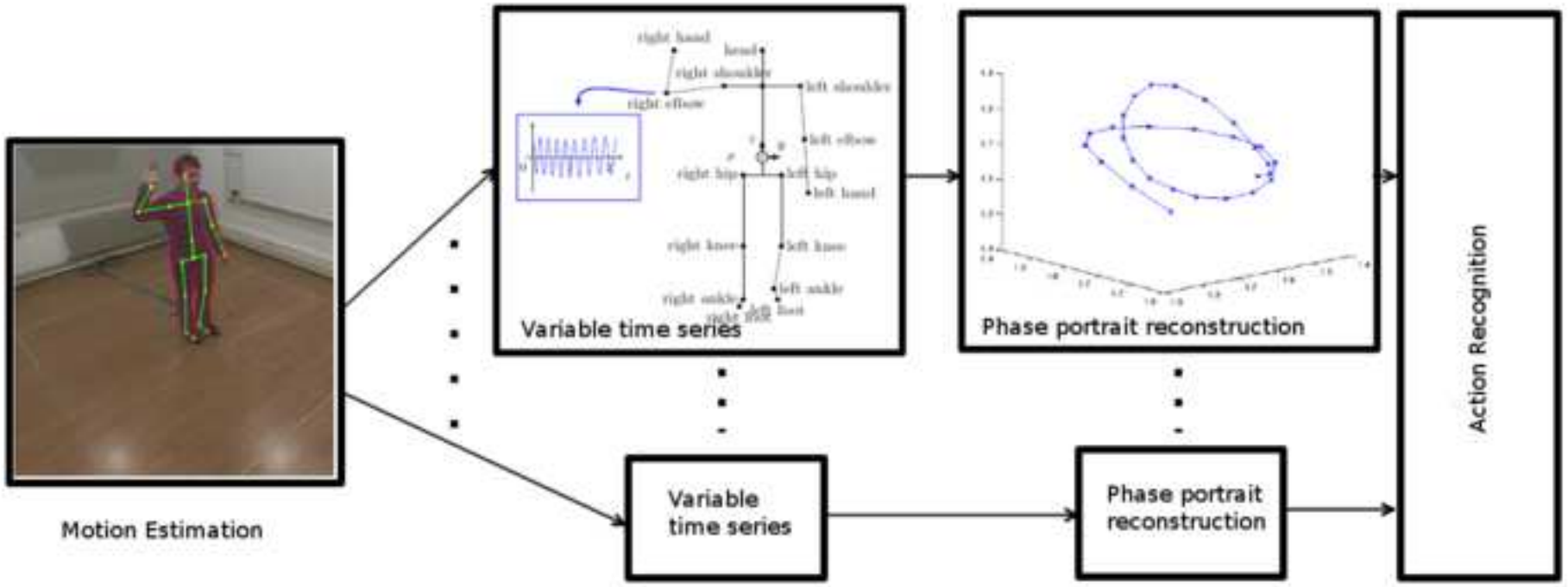
Abstract: This paper presents a human action recognition framework based on the theory of nonlinear dynamical systems. The ultimate aim of our method is to recognize actions from multi-view video. We estimate and represent human motion by means of a virtual skeleton model providing the basis for a view-invariant representation of human actions. Actions are modeled as a set of weighted dynamical systems associated to different model variables. We use time-delay embeddings on the time series resulting of the evolution of model variables along time to reconstruct phase portraits of appropriate dimensions. These phase portraits characterize the underlying dynamical systems. We propose a distance to compare trajectories within the reconstructed phase portraits and a method to learn a set of weights reflecting the discriminative power of a given model variable in a given action class. Our approach presents a good behavior on noisy data, even in cases where action sequences last just for a few frames. Experiments with marker-based and markerless motion capture data show the effectiveness of the proposed method. To the best of our knowledge, our contribution is the first to apply time-delay embeddings on data obtained from multi-view video.

Suggested Reviewers:

Abstract

This paper presents a human action recognition framework based on the theory of nonlinear dynamical systems. The ultimate aim of our method is to recognize actions from multi-view video. We estimate and represent human motion by means of a virtual skeleton model providing the basis for a view-invariant representation of human actions. Actions are modeled as a set of weighted dynamical systems associated to different model variables. We use time-delay embeddings on the time series resulting of the evolution of model variables along time to reconstruct phase portraits of appropriate dimensions. These phase portraits characterize the underlying dynamical systems. We propose a distance to compare trajectories within the reconstructed phase portraits and a method to learn a set of weights reflecting the discriminative power of a given model variable in a given action class. Our approach presents a good behavior on noisy data, even in cases where action sequences last just for a few frames. Experiments with marker-based and markerless motion capture data show the effectiveness of the proposed method. To the best of our knowledge, this contribution is the first to apply time-delay embeddings on data obtained from multi-view video.

Graphical Abstract



***Highlights**

>Human action recognition framework based on nonlinear dynamical systems>Use of a view-invariant representation of motion by means of a body model>Action model: set of weighted dynamical systems associated to each model variable >Dynamical systems modeled as phase portraits obtained by time-delay embeddings>Novel application in multi-view video data with promising results

Model-Based Recognition of Human Actions by Trajectory Matching in Phase Spaces

Adolfo López-Méndez[a], Josep R. Casas[b]

[a] *Technical University of Catalonia (UPC)*
email: adolf.lopez@upc.edu
[b] *Technical University of Catalonia (UPC)*
email: josep.ramon.casas@upc.edu

Abstract

This paper presents a human action recognition framework based on the theory of nonlinear dynamical systems. The ultimate aim of our method is to recognize actions from multi-view video. We estimate and represent human motion by means of a virtual skeleton model providing the basis for a view-invariant representation of human actions. Actions are modeled as a set of weighted dynamical systems associated to different model variables. We use time-delay embeddings on the time series resulting of the evolution of model variables along time to reconstruct phase portraits of appropriate dimensions. These phase portraits characterize the underlying dynamical systems. We propose a distance to compare trajectories within the reconstructed phase portraits and a method to learn a set of weights reflecting the discriminative power of a given model variable in a given action class. Our approach presents a good behavior on noisy data, even in cases where action sequences last just for a few frames. Experiments with marker-based and markerless motion capture data show the effectiveness of the proposed method. To the best of our knowledge, this contribution is the first to apply time-delay embeddings on data obtained from multi-view video.

Keywords: action recognition, multi-view, markerless motion capture, nonlinear dynamical systems, time-delay embeddings

1. Introduction

Automatic recognition of human actions from video is a challenging problem that has attracted the attention of researchers in the recent decades.

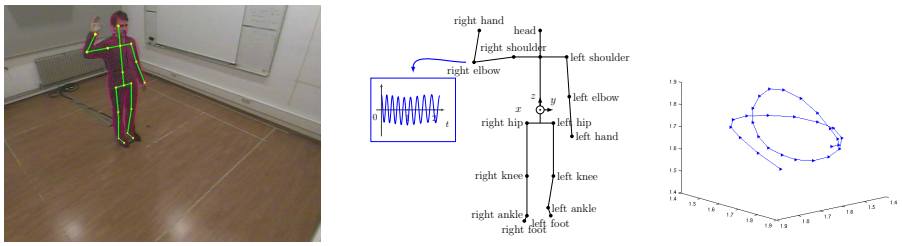
Human computer interfaces, surveillance or entertainment, to name a few, are among the wide range of applications of this technology.

From the computer vision perspective, the evolution of certain features in the temporal dimension is crucial to recognize human actions. The most common approach consists in extracting these features directly from pixels [1, 2]. While one can obtain accurate recognition results even with low-level features, approaches relying on pixels and appearance-based descriptors are normally limited by view-points. In our work, however, we focus on using body models to overcome the view-point dependency. Human body models introduce motion constraints, offer a compact representation of human motion and are able to represent the underlying hierarchical relation between joints. Due to these properties, human body models are of interest for both pose inference and action recognition. We therefore understand the recognition of human actions as a two-step problem: pose inference and modeling the temporal evolution of a set of human model variables. This paper focuses on the second step.

The aim of this paper is to exploit the theory of nonlinear dynamical systems and, more specifically, time-delay embeddings, to model the dynamics of human actions using a set of parameters resulting from a motion capture procedure. In contrast to a few existing approaches in the literature [3], we propose a method to compare the set of trajectories lying in phase spaces reconstructed with experimental data. This comparison relies on the topology of trajectories in a phase space, and overcomes problems related to the availability of long-term observations. Our ultimate goal is to recognize a set of human actions in a multi-view environment. We experimentally illustrate our achievements by inferring the pose of humans and recognizing their actions in a multi-camera scenario. To the best of our knowledge, this is the first application of time-delay embeddings on data obtained from markerless motion capture in multi-view video. We also provide comparative performance with existing approaches by using a 3D MoCap dataset obtained with marker-based sensors.

2. Related Work

The term human action can be defined according to [4], thus meaning a simple motion pattern usually lasting a short period of time and typically executed by a single person. Provided this definition, approaches addressing the problem of human action recognition from video can be divided into



(a) (b) (c)

Figure 1: System overview. (a) Motion Capture estimate. (b) View-invariant representation of motion through the human model (c) Phase space reconstruction of elbow motion using a time-delay embedding.

two large categories, depending on whether they employ or not an explicit representation of the human body. In this paper, we use the term model-based for approaches using a human body model, usually in the form of a skeletal structure. In contrast, we term model-free the approaches that work directly with image or video features. These latter approaches are copious in number, and can be applied in a wide variety of situations. Nonetheless, they usually rely on view-dependent features. For a deeper analysis of model-free methods, we refer the reader to [4, 5].

Model-based approaches are able to tackle the problem of view-dependency whenever the model can be represented in the 3-dimensional world in a view-invariant manner [6]. This is the case of virtual skeleton representations, where motion sequences can be easily represented independently of the overall rotation and translation parameters. The main drawback is the pose estimation problem associated to these approaches: it requires capturing the motion from video. This problem is normally addressed by tracking approaches either in multi-view settings [7, 8], using range sensors [9, 10, 11], such as Kinect, or by example-based approaches [12, 13, 14].

Within model-based approaches, our work is related to the analysis of time series and the underlying dynamical systems that describe the motion of joints in a human model. Pavlovic and Rehg [15] infer the kinematics of walking and running to model dynamical systems that can be used to recognize both actions. With similar kinematic features, Bisacco et al. [16] classify human gaits by defining a distance in the space of dynamical systems. Campbell and Bobick [17] represent human motion using joint angles to construct curves in a phase space. They are able to segment fundamental

steps in ballet dance. Lv et al. [18], infer human actions by representing them as a set of weighted channels consisting in the temporal evolution of 3D joint coordinates. Raptis et al. [19] compare time series by considering a time warping model that accounts for the intrinsic characteristics of an underlying dynamical system. More recently, Raptis et al. [20], have proposed a method to model human actions as a linear time-invariant dynamical model of relatively small order that generates multivariate time series in the form of joint angle dynamics along time.

Some authors focus on dimensionality reduction methods to deal with the action recognition problem by embedding input feature spaces into low-dimensional spaces. Dimensionality reduction methods have been widely applied also in the field of model-free action recognition [21, 2, 22], due to the high dimensionality of many image features. However, when one deals with actions and time series, keeping the geometric structure of input spaces is often not sufficient, and there is a need for constraining embedding techniques to cope with the underlying dynamics. Lewandowski et al. [23], propose Temporal Laplacian Eigenmaps to embed time series of motion capture data and video data, obtaining good results on both human motion reconstruction and action recognition.

With the aim of tackling any modeling limitation produced by assumptions made on the dynamical model, Ali et al. [3] propose to exploit the theory of chaotic systems to recognize human actions. They embed one-dimensional joint signals in an m-dimensional phase space to compute metric and dynamical invariants. They are able to recognize actions using MoCap data and video data. In a recent work, Basharat and Shah [24] extend the univariate embedding to multivariate analysis in order to synthesize human motion and dynamic textures. Our proposal is based on this latter sort of nonlinear dynamical models [25] to infer human actions from video.

3. Proposed Approach

In our work, human actions are modeled as a collection of dynamical system models, each one characterizing the temporal evolution of a set of body model variables. These models are obtained by phase space reconstructions according to the theory of nonlinear time series analysis [25]. In contrast to other embedding techniques, in our approach we embed 1-D signals onto a higher dimensional space that is topologically equivalent to the dynamical system generating those signals.

In the following, we present the choice of variables used in order to learn motion patterns, we outline the phase space reconstruction method employed, and we propose a metric to compare reconstructed phase portraits. We finally describe the training and the recognition strategy proposed to recognize human actions.

3.1. Pose Representation

The main motivation of using a model-based approach to human action recognition is the view-invariant nature of the representation of human poses and actions provided by a skeletal model [6]. We employ a kinematic tree structure that encodes an articulated skeleton. In this modeling framework, and provided that bone elongations are fixed, human pose is represented by a translation $\mathbf{r} \in \mathbb{R}^3$, a global rotation $\phi \in \mathbb{R}^3$ and a set of joint angle rotations $\theta_i \in \mathbb{R}$. Each joint can have up to 3 associated rotations, depending on its degrees of freedom, and these rotations are independent of the global translation and rotation. This collection of variables matches the data provided by marker-based motion capture systems and it is also appropriate to perform markerless motion capture from video.

Due to the independence of joint rotations with respect to global variables, we achieve invariance by discarding those model variables whose evolution along time is independent of the type of human action. Specifically, we discard the model translation on the transverse plane (XY in Fig.1b) and the rotation around the vertical axis of the model, i.e., the orientation (z in Fig. 1b). The remaining vertical translation is normalized using the model height. Let us clarify this choice with walking as an example. A human can walk in several directions on a scenario. If the coordinate axes are aligned such that the XY plane is parallel to the floor, then all these directions are reflected on the orientation (rotation around the vertical axis) and the translation on the XY plane. If the human model is rotated, say 90°, around any other global axes of rotation, then walking is not possible in this scenario (the human is likely to be lying on the floor, for instance). Similarly, a pronounced translation in the vertical direction is not possible while walking (unless the terrain is not an horizontal plane, but we do not consider this case). Therefore, the translation on the XY plane and the orientation are not characteristic of any action.

As a result, we have a set of time series consisting in the angular evolution of joint angles, a normalized translation and two global rotation variables.

From now on, we use ξ to denote this set of body model variables describing human motion.

3.2. Phase Space Reconstruction

Let us define a dynamical system as the possibly nonlinear map $f : \mathbb{R}^m \rightarrow \mathbb{R}^m$ that describes the temporal evolution of state variables $\mathbf{x}(t) = [x_1(t), x_2(t), ..., x_m(t)] \in \mathbb{R}^m$. Similarly, let $h : \mathbb{R}^m \rightarrow \mathbb{R}$ be the mapping of the state-space variables to one-dimensional observables z that conform scalar time series. In our case, the temporal evolution of ξ variables, $z^{\xi}(t)$, are these scalar time series, and we want to characterize the underlying system f by reconstructing an m-dimensional phase space where time series are unfolded.

The reconstructed phase space is a metric space and, as shown by Takens [26], for a sufficiently large m, this space is an homeomorphism of the true dynamical system that generated the time series. Takens' theorem provides the theoretical justification for reconstructing state spaces using time-delay embeddings:

$$\hat{\mathbf{x}}^{\xi}(t) = [z^{\xi}(t), z^{\xi}(t+\tau), ..., z^{\xi}(t+(m-1)\tau)] \tag{1}$$

where $\hat{\mathbf{x}}^{\xi}(t)$ is a point in the reconstructed phase space, m is the embedding dimension and τ is the embedding delay. Hence, for a sufficiently large m, the phase space is reconstructed by stacking sets of m temporally equispaced samples of the input scalar time series. But not only dimension is essential: embedding delay τ also determines the properties of the reconstructed phase space. We first determine the embedding delay using the mutual information method [27] and then we employ the estimated delay to find the appropriate embedding dimension using the method by Cao [28].

Embedding delay. The basic idea behind the method for inferring the embedding delay is that an appropriate embedding delay must provide independent state variable coordinates, so that the time series data gets effectively unfolded in a higher dimensional space. In essence, the method in [27] computes the mutual information between the time series and delayed versions of the same series. The optimal delay is then obtained by taking the first minimum of the mutual information function (see Algorithm 1 and Fig. 2a).

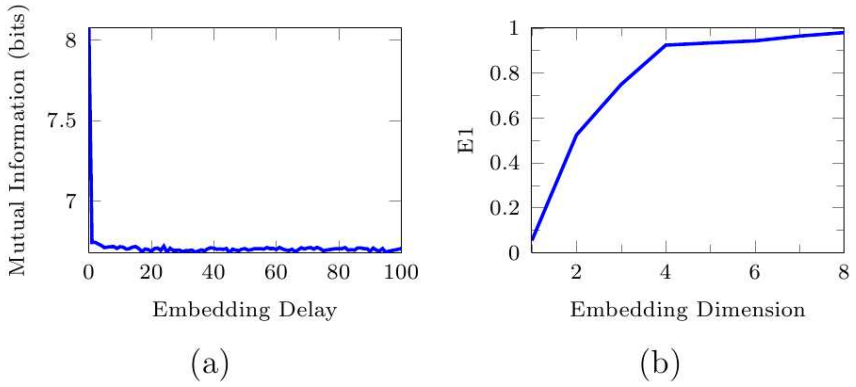


Figure 2: Determining the embedding delay and a suitable embedding dimension for a joint angle scalar time series. (a) Mutual information of the series and its delayed version. The first minimum is used as the optimal embedding delay. (b) E1 measure, obtained by Cao's method [28]. Suitable embedding dimensions are found when the E1 measure starts converging to a stable, high value. In the example, m=4 is a sufficient embedding dimension.

Algorithm 1: Embedding Delay

$\mathcal{Z}^{\xi} = \{z^{\xi}(0), ..., z^{\xi}(T-1)\}$: Time series obtained from the human body model variable ξ
Normalize $\mathcal{Z}^{\xi}$ between 0 and 1
Set $minMI >> 1$
for $\tau = 1 \dots max\ \tau$ **do**
    Delayed time series : $\mathcal{Z}^{\xi}_{\tau} = \{z^{\xi}(\tau), ..., z^{\xi}((T-1+\tau) \mod (T))\}$
    $I(\tau) = -\sum_{Z^{\xi}} \sum_{Z^{\xi}_{\tau}} p(Z^{\xi}, Z^{\xi}_{\tau}) \log \frac{p(Z^{\xi}, Z^{\xi}_{\tau})}{p(Z^{\xi}) p(Z^{\xi}_{\tau})}$
    if $I(\tau) < minMI$
    then
        $minMI = I(\tau)$
    end
    else
        break //τ is the optimal embedding delay//
    end
end

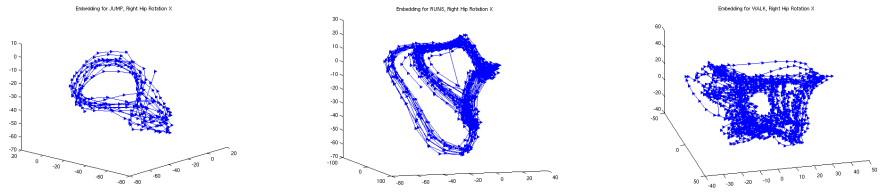


Figure 3: Examples of reconstructed 3-D phase portraits. Hip rotations for jump, run and walk actions in the Future Light dataset [29].

Embedding dimension. Kennel et al. [30] noted that, since the phase space of a dynamical system folds and unfolds smoothly, two points that are sufficiently close in a given m-dimensional reconstructed phase space, should keep close in $m+1$ dimensions. They reasoned that, as the optimal embedding dimension is reached, the number of false nearest neighbors (points that are nearest neighbors in m dimensions but not in $m+1$ dimensions) should be negligible. Noting the limitations of the false nearest neighbors method for short-term experimental data, Cao [28] proposed a novel method for finding the optimal embedding dimension (see Algorithm 2 and Fig. 2b).

Once both the embedding delay and the embedding dimension have been estimated, the phase space reconstruction is performed as follows:

$$\hat{\mathbf{X}}^{\xi} = \begin{bmatrix} z^{\xi}(0) & z^{\xi}(\tau) & \cdots & z^{\xi}((m-1)\tau) \\ & \vdots & & \\ z^{\xi}(t) & z^{\xi}(t+\tau) & \cdots & z^{\xi}(t+(m-1)\tau) \\ & \vdots & & \\ z^{\xi}(N-1-(m-1)\tau) & z^{\xi}(N-1-(m-2)\tau) & \cdots & z^{\xi}(N-1) \end{bmatrix} \tag{2}$$

Our objective is to use the phase space portraits $\hat{\mathbf{X}}^{\xi}$ as signatures. To match these signatures, we exploit some topological and geometric features of the obtained trajectories in high dimensional spaces. As mentioned throughout the paper, each one of the model variables constitutes a time series from which we reconstruct a phase space. In order to easily compare the reconstructed phase portraits without using invariants [3], we must enforce a common embedding dimension for all the actions and model variables.

Algorithm 2: Embedding Dimension

$\hat{\mathbf{x}}_m^{\xi}(i) = \left[z^{\xi}(i) \quad z^{\xi}(i+\tau) \quad \cdots \quad z^{\xi}(i+(m-1)\tau)\right]$ // Reconstructed phase space point in m dimensions //

N //Number of points in time series $\mathcal{Z}^{\xi}$ //

for $m = 1 \ldots M$ **do**

  for $i = 0 \cdots N-1-m\tau$ **do**

    Find the Nearest Neighbor $\hat{\mathbf{x}}^{\xi}(j)$

    $a(i,m) = \frac{\|\hat{\mathbf{x}}_{m+1}^{\xi}(i) - \hat{\mathbf{x}}_{m+1}^{\xi}(j)\|}{\|\hat{\mathbf{x}}_{m}^{\xi}(i) - \hat{\mathbf{x}}_{m}^{\xi}(j)\|}$

  end

  $E(m) = \frac{1}{N-m\tau} \sum_{i=0}^{N-1-m\tau} a(i,m)$

  if $m \geq 2$ **then**

    $E1(m-1) = E(m)/E(m-1)$

  end

  if $m \geq 3$ **then**

    if $\|E1(m-1) - E1(m-2)\| < Th$ **then**

      break //$m-1$ is a suitable embedding dimension//

    end

  end

end

3.3. Distance between phase portraits

The proposed recognition method is based on comparing the set of reconstructed phase portraits of a target action (one per rotation) with a set of phase portraits from representative templates. In order to perform such a comparison, we propose a distance between two reconstructed phase portraits $\hat{\mathbf{X}}_a^{\xi}$ and $\hat{\mathbf{X}}_b^{\xi}$.

We use the map $f(\mathbf{x}(t)) \rightarrow \mathbf{x}(t+1)$ to refer to the underlying dynamical system associated to some model variable and $\hat{f}(\mathbf{x}_a(n)) \rightarrow \mathbf{x}_a(n+1) \;\; \mathbf{x}_a \in \hat{\mathbf{X}}_a^{\xi}$ to the system representing the evolution in the phase portrait $\hat{\mathbf{X}}_a^{\xi}$ reconstructed with the model time series (we use n instead of t to denote that we have a sampled signal). From this notation we introduce the nonlinear time evolution operator $\hat{f}^n$ such that:

$$\hat{f}^0 = \text{identity} \quad (3)$$
$$\hat{f}^1 = \hat{f} \quad (4)$$
$$\hat{f}^n \hat{f}^s = \hat{f}^{n+s} \quad (5)$$

Due to the nature of human motion, we assume that f is smooth. We also assume that important topological characteristics of the underlying dynamics are reflected on the trajectories of the reconstructed phase portraits (Fig. 3), in spite of the fact that the data employed is a sampled version of a continuous signal.

Our proposal for the distance between reconstructed phase portraits relies on these properties and it is inspired on DTW [31]. The underlying idea is to traverse one phase portrait $\hat{\mathbf{X}}_a^\xi$ and look for the nearest neighbor at each time instant in another phase portrait $\hat{\mathbf{X}}_b^\xi$. If the portraits have been generated by similar motion (and possibly by the same action) nearest neighbors should be in closer distances along the respective reconstructed phase portraits. Nearest neighbor retrieval is, therefore, restricted in such a way that the temporal ordering is preserved. This is equivalent to measuring the prediction error of points in $\hat{\mathbf{X}}_a^\xi$ given $\hat{\mathbf{X}}_b^\xi$. Algorithm 3 describes the computation of the distance between two reconstructed phase portraits of some body model rotation.

The search for nearest neighbors is restricted by δ to the near future of the phase portrait. This variable aims at compensating for the different lengths of the available action samples. Note that while the distance performs a one-pass traversal through $\hat{\mathbf{X}}_a^\xi$, we allow $\hat{f}_b$ to go from the last reconstructed state space point to the first one. This is clearly necessary for periodic or quasi-periodic human motion, such as walking. Note also that the distance between $\hat{\mathbf{X}}_a^\xi$ and $\hat{\mathbf{X}}_b^\xi$ is, in general, different from the distance between $\hat{\mathbf{X}}_b^\xi$ and $\hat{\mathbf{X}}_a^\xi$. Hence, in order to have a true metric, we define the final distance between two phase portraits as:

$$e(\hat{\mathbf{X}}_a^\xi, \hat{\mathbf{X}}_b^\xi) = \frac{d(\hat{\mathbf{X}}_a^\xi, \hat{\mathbf{X}}_b^\xi) + d(\hat{\mathbf{X}}_b^\xi, \hat{\mathbf{X}}_a^\xi)}{2} \quad (6)$$

3.4. Training and Recognition

Human actions are not necessarily characterized by full body motion. Clearly, understanding the action *raise hand* does not require knowledge of

Algorithm 3: d($\hat{\mathbf{X}}_a^\xi$, $\hat{\mathbf{X}}_b^\xi$)

Start at $\mathbf{x}_a(0) \in \hat{\mathbf{X}}_a^\xi$
Look for its nearest neighbor $\mathbf{x}_b(n_0) \in \hat{\mathbf{X}}_b^\xi$
$\mathbf{x}_{bl} = \mathbf{x}_b(n_0)$
$d = \|\mathbf{x}_a(0) - \mathbf{x}_b(n_0)\|^2$
$n = 1$
while $\mathbf{x}_a(n) \neq \mathbf{x}_a(0)$ **do**
    $l = \underset{k}{argmin} \quad \|\mathbf{x}_a(n) - \hat{f}_b^k(\mathbf{x}_{bl}))\|^2$ restricted to $0 \leq k < \delta$.
    $d = d + \|\mathbf{x}_a(n) - \hat{f}_b^l(\mathbf{x}_{bl}))\|^2$
    $\mathbf{x}_{bl} = \hat{f}_b^l(\mathbf{x}_{bl})$
    $n = n + 1$
end
$d = d/n$

the motion of legs. Even in the case of having full body motion, it might happen that some parts are more important for understanding a human action. Motivated by the potential saliency of different body parts in different actions, we use training data to learn a set of weights reflecting the importance of the motion of different body model variables in each action class.

Due to the independence of the variables in the model-based representation of human actions, we propose a strategy that performs fusion of the similarity scores obtained in each model variable ξ. Let $\{\Gamma_\xi, Y_\xi\}$ be the training data consisting in the set of reconstructed phase portraits $\hat{\mathbf{X}}_i^\xi$ and action class labels for the variable ξ, respectively. We aim at maximizing the classification accuracy by weighting each one of the variables ξ based on their discriminative power on the training data. For this reason, we first classify each training sample using the nearest neighbor, based on the distance $e(\hat{\mathbf{X}}_i^\xi, \hat{\mathbf{X}}_j^\xi), \quad i \neq j$. Using the training data false positive, false negative (misses) and true positive classifications for the action class label y in each model variable ξ ($\mathrm{fp}_{Y_\xi=y}$, $\mathrm{fn}_{Y_\xi=y}$ and $\mathrm{tp}_{Y_\xi=y}$ respectively), we define the weights for each variable ξ and action y as:

$$w_{Y_\xi=y} \propto \frac{\mathrm{tp}_{Y_\xi=y}}{\mathrm{fp}_{Y_\xi=y} + \mathrm{fn}_{Y_\xi=y} + \mathrm{tp}_{Y_\xi=y}} \tag{7}$$

To classify a test action based on the set of reconstructed phase portraits

$\hat{\mathbf{X}}_{\rho}^{\xi}$, we normalize the weights $w_{Y_\xi=y}$ such that $\sum w_{Y_\xi=y} = 1$, and we fuse the similarity scores with respect to the training samples :

$$\hat{y} = \underset{y \in Y}{argmax} \sum_{\xi} w_{Y_\xi=y} \exp\left(-e(\hat{\mathbf{X}}_{\rho}^{\xi}, \hat{\mathbf{X}}_{i,Y_\xi=y}^{\xi})\right) \quad (8)$$

The proposed strategy takes advantage of training data to learn the variable model weights in a discriminative manner, i.e., the importance of each model variable depends on the extent to which its motion pattern in a given action differs from motion patterns in other actions.

4. Experimental Results

In order to evaluate our method, we have conducted two different types of experiments. On the one hand, we have tested our method with a 3D MoCap dataset [29] for comparative performance with previously reported results in the literature [3, 20, 32]. On the other hand, we have tested the proposed technique with markerless motion capture data obtained in a multi-view environment with our own implementation of a hierarchical particle filter algorithm for human pose estimation. The latter experiments aim at evaluating the trajectory matching in phase spaces with less accurate motion data.

4.1. MoCap Data

The motion capture data dataset provided by FutureLight [29] contains 158 sequences of 5 action classes: *dance*, *jump*, *run*, *sit* and *walk*. These sequences have important intra-class variations that make this dataset a challenging one. For instance, dance sequences contain several ballet moves but also 4 walk-style dance moves that, actually, are similar to walking. Run and walk class sequences present important variations in speed, arm swing or bouncing, and include stops and turnings. Similarly, jump class sequences are performed in place or while walking. A good description of this dataset, illustrated with some samples, can be found in [3].

We employ a total of 27 variables (arm and leg joint rotations, 2 global rotations and normalized translation in the vertical axis), thus obtaining 27 reconstructed phase portraits per action sequence. Classification performance is evaluated using leave-one-out cross-validation, as has been done

by other authors working in action recognition with this dataset [3, 20, 32]. Classification results are shown in Table 1. Classification accuracy (91.77%) validates the performance of reconstructed phase portraits and the proposed distance for the action recognition task, as the results obtained are comparable to existing approaches [20, 3], but are ranked just below the approach by Raptis et al. [32] (see Table 2). They use an efficient time series dictionary in these dataset, thus achieving robustness against intra-class variations. On the contrary, our method makes a comparison that is rather exhaustive, in the sense that all points of the available trajectories are matched. Therefore, our method presents limitations in coping with significant performance differences between the same action class.

The comparative performance of the method by Ali et al. [3] and our method is relevant, since they use chaotic invariants and time-delay embeddings. The accuracy of our algorithm on this data is slightly better but, in overall, both results are consistent, thus proving that methods using time-delay embeddings perform well in recognizing human actions from MoCap data. Apart from the improved recognition performance, our method has the advantage of being suitable for short term actions, as we show in experiments with video data (see Section 4.2).

Using the same model variables, we evaluate the recognition rate as a function of the embedding dimension (Fig. 4). In the case of $m = 1$ we perform DTW [31] using a window to constrain the warping to near time instants of the evaluated sample. In this way, the computation time is dramatically reduced. As expected, the recognition rate improves as the embedding dimension m grows, until a limit is reached (in our case, this limit is reached at $m = 4$). If we look at the E1 measure computed using the false nearest neighbor method, we see that at $m = 4$ the slope changes notably and the E1 measures starts settling down to a convergence value. This is an interesting result, since it shows the correlation between the recognition performance and the theoretical correctness, both topologically and geometrically, of the reconstructed phase space (according to the Takens' theorem and the embedding dimension m).

4.2. Video Data

To test our method, we record several sequences involving four actors performing several actions. This test scenario is intended to simulate a smart environment where each user interacts with the room through 4 calibrated

	Dance	Jump	Run	Sit	Walk
Dance	**28**	3	0	0	0
Jump	0	**11**	1	0	2
Run	1	3	**25**	0	1
Sit	0	0	3	**33**	0
Walk	0	0	0	0	**48**

Table 1: Confusion Matrix of the Future Light dataset for an embedding dimension $m = 5$ and δ=10. Overall classification accuracy is 91.77%.

Method	Performance on the FutureLight Dataset
[3]	89.7
[32]	98.03
[20]	83.63
Our method	91.77

Table 2: Comparison of classification results

cameras located at the ceiling corners (see Fig. 5). In this context, intra-class variations of the actions performed are not as relevant as in the case of the Future Light dataset, because in this dataset many actions serve as commands, and are executed in a somewhat precise manner. In spite of that, some stylistic variations exist, especially between actors. We employ 7 minutes of video data containing the following actions: *walk*, *raise hand*, *crouch*, *wave hand*, *bounce* (an arm movement similar to bouncing a ball), *jump*, *clap*, *kick*, *sit* and *stand up*. Actions involving a single arm or leg motion are performed with the right limb. Except for walk, all the actions last a few seconds (typically between 1 and 4 seconds). This fact produces reconstructed phase portraits with a few points in several cases, making unfeasible to apply chaotic invariants reliably unless some re-sampling is performed [3]. Concatenation of sequences has been proposed also as a possible solution, but in some actions concatenation may produce jumps that would violate the continuity of the dynamics. The metric distance that we propose in this paper does not require a very good quality of the reconstructed phase portraits, since we are analyzing the trajectories in the phase space. Since the estimated time delays and the embedding dimensions allow having at least 15 reconstructed points in the worst case, we do not perform any up-sampling nor concatenation.

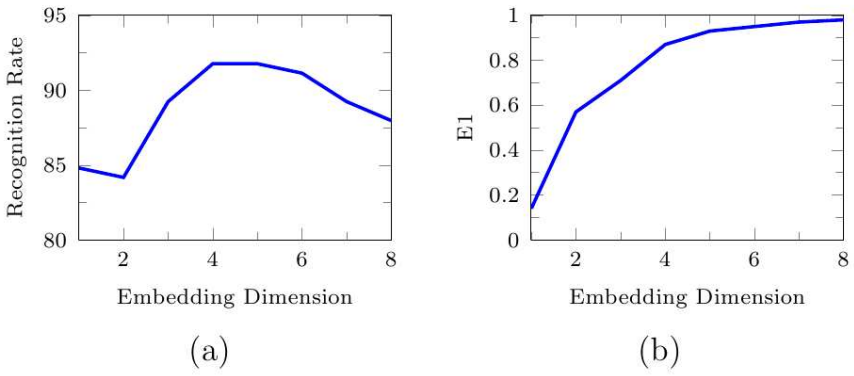


Figure 4: Analysis of the recognition rate and the embedding dimensions in the Future Light dataset (a) Recognition rate as a function of the embedding dimension. (b) Mean E1 measure [28] on the whole Future Light dataset.

Figure 5: Multi-view data-set samples. Top row: The four available viewpoints during walk action. Bottom row: Kick, jump, sit and bounce samples.

Tracking. To track the actors, we employ a markerless motion capture method based on our previous work on Hierarchical Particle Filter-based pose and anthropometrics estimation [33]. Although the state vector strictly comprises pose and anthropometric variables, we simplify the problem by computing anthropometric variables only once at the beginning of each subject's sequence. Our tracking method uses a body model comprising an articulated skeleton (a set of joints connected by bones) and a triangular mesh surface model to generate pose and anthropometric candidates. In this modeling framework, pose is formulated as a global rotation and translation and a

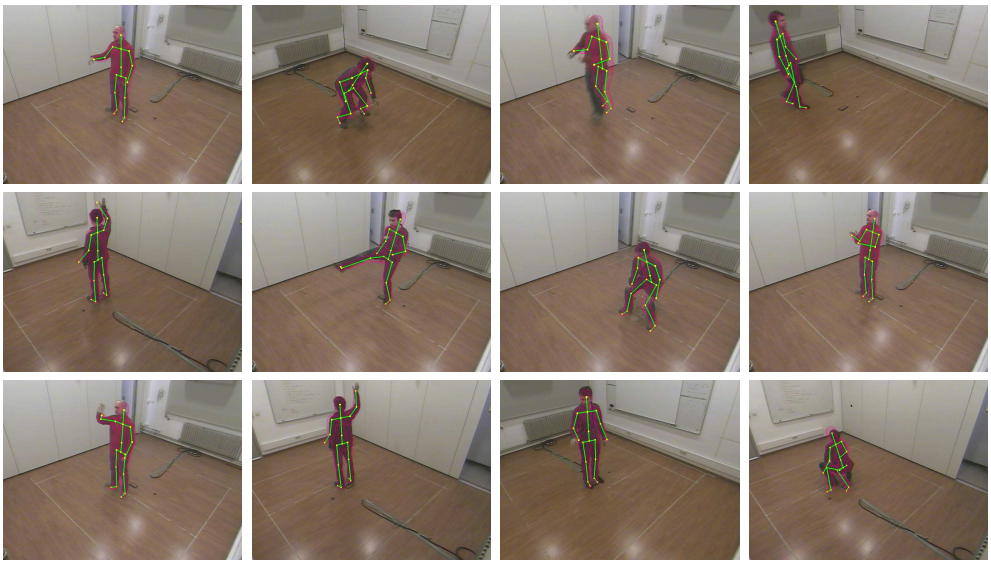


Figure 6: Selected tracking samples.

set of joint angles, while the anthropometric profile is expressed in terms of bone elongations and mesh deformations. In this way, all the joint positions and mesh vertices can be controlled by varying the state space vector. Pose and anthropometric candidates are projected onto the multiple views and the mesh projection is compared with the foreground silhouettes extracted in each corresponding view in order to define the weight of each particle. In the HPF formalism, the state space vector is sampled and filtered by using hierarchical partitions; specifically, we divide the state space variables into global and torso variables, arms and legs in a total of 7 layers. From the tracking results we obtain the necessary rotations. We use 19 variables, since our model is simpler than the one resulting from marker-based motion capture in [29]. We estimate an average tracking error of 15 cm for all the considered data, although errors occur especially in some arm motions (clapping), in fast motion (kick, jump) and in crouching actions, where the tracker makes important errors in estimating the legs' pose. In spite of that, the tracking provides an approximate representation of the motion involved in each action (see Fig. 6).

Evaluation. To evaluate the classification performance, we divide the data into several training and testing sets. We pick 2/3 of the action sequences for training and 1/3 for testing. We repeat this procedure 10 times with

	walk	raise hand	crouch	wave hand	bounce	jump	clap	kick	sit	stand up
walk	**1.00**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
raise hand	0.00	**0.66**	0.00	0.04	0.24	0.0	0.06	0.00	0.00	0.00
crouch	0.00	0.00	**0.96**	0.00	0.00	0.01	0.00	0.00	0.03	0.00
wave hand	0.00	0.03	0.00	**0.83**	0.10	0.00	0.04	0.00	0.00	0.00
bounce	0.00	0.09	0.00	0.36	**0.56**	0.00	0.00	0.00	0.00	0.00
jump	0.00	0.00	0.00	0.00	0.00	**0.93**	0.00	0.06	0.01	0.00
clap	0.00	0.02	0.00	0.00	0.00	0.00	**0.97**	0.01	0.00	0.00
kick	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**1.00**	0.00	0.00
sit	0.00	0.00	0.00	0.00	0.00	0.17	0.00	0.00	**0.83**	0.00
stand up	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**1.00**

Table 3: Confusion Matrix for Tracking Data. Training sets of 2/3 of the available data. Overall classification performance is 87.5% ($m = 4$, δ=5).

random partitions of the data. The classification results, for embedding dimension $m = 4$ and parameter $\delta = 5$, are shown in Table 3. The proposed method shows an excellent performance with actions involving large motions (the majority of these actions are classified with rates higher than 90%). The training method, based on model variable fusion, provides effective results with these actions. For instance, in *walk* action, although almost all the model variables have non-negligible weights (except for the vertical translation), the importance is set on the leg variables. Similarly, *crouch* has 75% of the weight distributed in the knees' rotation and the vertical translation. In contrast, actions involving a single arm motion are classified with less accuracy. We have observed that in these actions, right shoulder rotations get heavily weighted during training phase. Specifically, the actions *raise hand* and *wave hand* have high weights on x and y axes of rotation in the shoulder joint (in our model, the z rotation axis is defined along the upper arm), as it is a salient feature in the dataset. In the case of *wave hand* and *bounce*, right elbow joint is weighted over the mean, because the pattern in this joint is particular of these two actions. However, in action *bounce*, the shoulder's y-axis, a differential trait from *wave hand*, has no importance. This is actually because it is differential with respect to *wave hand*, but it is very similar to many other actions in the dataset. Similarly, elbow rotation in action *raise hand*, as a differential trait versus *wave hand*, is not significantly weighted during training for the same reason. To summarize, our method is able to

focus on distinct human model variables in an automated manner, with effective results when large motion takes place in the performance of an action. However, weighting model variables in a discriminative manner considering them isolatedly has limitations with a single arm motion.

To further analyze the behavior of our action recognition method, we repeat the experiment reducing the amount of training data. Specifically, we use 1/2 and 1/3 of the data as training in two additional experiments. As before, we repeat the procedure 10 times for each one of the experiments. The classification accuracies for 1/2 and 1/3 of training data are 84.3% and 76.7% respectively, showing that the proposed method can perform accurately with a few action exemplars.

In a final experiment, we use the action sequences of 3 actors for training and the remaining for testing. In this way, we evaluate the performance of the algorithm under stylistic variations. In this last experiment, the recognition rate is 79.9%, confirming that the dense matching of trajectories in the phase space is sensitive to some stylistic variations. In overall, the proposed method has a promising performance with multi-view video data in spite of some important tracking errors.

5. Conclusions

In this paper we have presented a novel human action recognition method based on the analysis of time series by phase space reconstructions with time-delay embeddings. We employ a set of variables of a skeletal model as the set of time series, each of which is used to reconstruct a phase portrait. We propose a distance to compare each phase portrait without requiring the computation of chaotic invariants, which might need a high number of samples to be significant. We also propose to perform fusion of the different classifications performed at the variable level, a strategy that allows focusing on a subset of variables that produces sufficient information to classify an action. We have experimentally shown the feasibility of such an approach to recognize human actions from multi-view video for the first time, in a novel application of theory of nonlinear dynamical systems.

We have observed that a dense matching of embedded time series has limited accuracy against intra-class variations. Therefore, finding of new ways of comparing phase space points is a future research line. On the other hand, although experimental results have shown that the discriminative learning of weights per model variable is an effective method of removing

nuisance from the pose data, the proposed method depends solely on training data, thus ignoring the topological information of the human body model. We believe that by introducing topological information of the skeletal model we can avoid problems derived from weighting isolated variables, thus avoiding the algorithm to ignore some information that may be of interest to recognize actions with more subtle motions.

References

[1] H. Jhuang, T. Serre, L. Wolf, T. Poggio, A biologically inspired system for action recognition, in: ICCV 2007, pp. 1 –8.

[2] L. Wang, D. Suter, Recognizing human activities from silhouettes: Motion subspace and factorial discriminative graphical model, in: CVPR 2007, pp. 1 –8.

[3] S. Ali, A. Basharat, M. Shah, Chaotic invariants for human action recognition, in: ICCV 2007, pp. 1–8.

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

[5] R. Poppe, A survey on vision-based human action recognition, Image and Vision Computing 28 (2010) 976–990.

[6] V. Parameswaran, R. Chellappa, View invariance for human action recognition, Int. J. Comput. Vision 66 (2006) 83–101.

[7] J. Gall, B. Rosenhahn, T. Brox, H. Seidel, Optimization and filtering for human motion capture - a multi-layer framework, International Journal of Computer Vision 0 (2010) 1–18.

[8] J. Bandouch, M. Beetz, Tracking humans interacting with the environment using efficient hierarchical sampling and layered observation models, in: ICCV-HCI 2009.

[9] M. Siddiqui, G. Medioni, Human pose estimation from a single view point, real-time range sensor, in: CVPR 2010 Workshops, pp. 1 –8.

[10] V. Ganapathi, C. Plagemann, D. Koller, S. Thrun, Real time motion capture using a single time-of-flight camera, in: CVPR 2010, pp. 755 –762.

[11] J. Shotton, A. Andrew Fitzgibbon, M. Cook, A. Blake, Real-time human pose recognition in parts from single depth images, in: CVPR 2011.

[12] S. Hou, A. Galata, F. Caillette, N. Thacker, P. Bromiley, Real-time body tracking using a gaussian process latent variable model, Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on (2007) 1–8.

[13] G. Taylor, L. Sigal, D. Fleet, G. Hinton, Dynamical binary latent variable models for 3d human pose tracking, in: CVPR 2010, pp. 631 –638.

[14] R. Urtasun, T. Darrell, Sparse probabilistic regression for activity-independent human pose inference, in: CVPR 2008, pp. 1 –8.

[15] V. Pavlovic, J. Rehg, Impact of dynamic model learning on classification of human motion, in: CVPR 2000, volume 1, pp. 788 –795 vol.1.

[16] A. Bissacco, A. Chiuso, Y. Ma, S. Soatto, Recognition of human gaits, in: CVPR 2001, volume 2, pp. II–52 – II–57 vol.2.

[17] L. Campbell, A. Bobick, Recognition of human body motion using phase space constraints, in: ICCV 1995, pp. 624–630.

[18] F. Lv, R. Nevatia, M. W. Lee, 3d human action recognition using spatio-temporal motion templates, in: ICCV-HCI 2005, pp. 120–130.

[19] M. Raptis, M. Bustreo, S. Soatto, Time warping under dynamic constraints (2007).

[20] M. Raptis, K. Wnuk, S. Soatto, Spike train driven dynamical models for human actions, in: CVPR 2010, pp. 2077–2084.

[21] T.-J. Chin, L. Wang, K. Schindler, D. Suter, Extrapolating learned manifolds for human activity recognition, in: ICIP 2007, volume 1, pp. I –381 –I –384.

[22] J. Blackburn, E. Ribeiro, Human motion recognition using Isomap and Dynamic Time Warping, in: Proceedings of the 2nd conference on Human motion: understanding, modeling, capture and animation, 2007, Springer-Verlag, Berlin, Heidelberg, 2007, pp. 285–298.

[23] M. Lewandowski, J. Martinez-del Rincon, D. Makris, J.-C. Nebel, Temporal extension of laplacian eigenmaps for unsupervised dimensionality reduction of time series, in: Pattern Recognition (ICPR), 2010 20th International Conference on, pp. 161 –164.

[24] A. Basharat, M. Shah, Time series prediction by chaotic modeling of nonlinear dynamical systems., in: ICCV 2009, pp. 1941–1948.

[25] H. Kantz, T. Schreiber, Nonlinear time series analysis, Cambridge U. Press, 2004.

[26] F. Takens, Detecting strange attractors in turbulence, Dynamical Systems and Turbulence (1981) 366–381.

[27] A. M. Fraser, H. L. Swinney, Independent coordinates for strange attractors from mutual information, Physical Review A 33 (1986) 1134–1140.

[28] L. Cao, Practical method for determining the minimum embedding dimension of a scalar time series, Physica D: Nonlinear Phenomena 110 (1997) 43 – 50.

[29] Future Light R&D division of Santa Monica, ????

[30] M. B. Kennel, R. Brown, H. D. I. Abarbanel, Determining embedding dimension for phase-space reconstruction using a geometrical construction, Phys. Rev. A 45 (1992) 3403–3411.

[31] D. J. Berndt, J. Clifford, Advances in knowledge discovery and data mining, American Association for Artificial Intelligence, Menlo Park, CA, USA, 1996, pp. 229–248.

[32] M. Raptis, K. Wnuk, S. Soatto, Flexible dictionaries for action classification, in: Proceedings of the International Workshop on Machine Learning for Vision-based Motion Analysis, in conjunction with ECCV 2008.

[33] M. Alcoverro, J. Casas, M. Pardàs, Skeleton and Shape Adjustment and Tracking in Multicamera Environments, in: F. J. P. López, R. B. Fisher (Eds.), AMDO, volume 6169 of *Lecture Notes in Computer Science*, Springer, 2010, pp. 88–97.