Probabilistic Discriminative Dimensionality Reduction for Pose-Based Action Recognition

Valsamis Ntouskos, Panagiotis Papadakis and Fiora Pirri

Abstract We examine the problem of classifying action sequences given a small set of examples for each type of action. Based on the presumption that human motion resides in a low dimensional space, we introduce a probabilistic dimensionality reduction model able to recover the structure of a low-dimensional manifold where all the involved actions reside. Requiring that sequences of the same action are placed apart from other sequences, we are able to achieve higher classification rates, with respect to other commonly used techniques, by performing the classification on this manifold. The main contribution is the introduction of a new model, based on Back-constrained GP-LVM which can be used for the efficient classification of sequences. We compare our method with the classification based on the Dynamic Time Warping distance and with the V-GPDS model, adapted for classification. Results are provided for sequences taken from two publicly available datasets which highlight different aspects of the method.

Keywords Action recognition · Dimensionality reduction · Manifold learning · Time series models · Motion capture

1 Introduction

Human action recognition is one of the most challenging applications in the field of computer vision. It requires to infer an action model from the observation of a motion sequence, hence it requires the solution of an inverse problem [18]. Furthermore, the modelling process is based on several steps tackling, in turn, different sub-problems: data acquisition, motion analysis and segmentation in individual actions, alignment

V. Ntouskos (✉) · P. Papadakis · F. Pirri
ALCOR Laboratory, Sapienza University of Rome, Rome, Italy
e-mail: ntouskos@dis.uniroma1.it

P. Papadakis
e-mail: papadakis@dis.uniroma1.it

F. Pirri
e-mail: pirri@dis.uniroma1.it

A. Fred and M. De Marsico (eds.), *Pattern Recognition Applications and Methods*,
Advances in Intelligent Systems and Computing 318,
DOI 10.1007/978-3-319-12610-4_9

between sequences and classification with respect to a given taxonomy. All these steps are computationally expensive, while ideally recognition should be performed online.

In this paper we address the alignment and classification part of the complete pipeline. Namely, we assume that a sequence that captures an individual action is already available and the task is to recognize the performed action. To this end we introduce a model based on the the Back-Constrained GP-LVM introduced in [9, 10], and extend it for the application of action recognition, exploiting the strength of a lower dimensional manifold. In detail, we derive a discriminative, probabilistic dimensionality reduction model for mapping motion capture sequences in a low dimensional latent space which assists the action classification process. The proposed model introduces a latent space featuring a fixed set of actions and constrains feature distances in data space to be suitably projected in the latent space, in order to preserve the clustering of common patterns. Actions are represented as a sequence of poses, which can be taken from motion capture (MoCap) data. This projection ensures a discriminative power to the GP-LVM model and it also exploits the peculiar property of action sequences of being reducible to a lower dimensional manifold [17].

In Sect. 2 we briefly review recent works on pose-based action recognition and dimensionality reduction, showing the major trends of research in this field. In Sect. 3 we overview the theoretical foundation of GP-LVM on which our model is based. In Sect. 4 we present our discriminative model. Section 5 demonstrates the latent space structure recovered by the proposed model and examines its performance on human action classification. We compare our method with a sequence classification method based on Dynamic Time Warping as well as the Variational Gaussian Process Dynamical Systems [6] recently proposed for modelling high dimensional dynamical systems. We conclude the work addressing possible extensions.

2 Related Work

In this section we review some of the main approaches to action recognition and mainly those which refer to manifold learning or treat the problem of action recognition in MoCap sequences.

So far many techniques have been proposed in the literature regarding action recognition where stochastic, volumetric or non-parametric models are most commonly employed. Detailed reviews of the techniques which have been considered in the research on human motion analysis and on action recognition can be found in [1, 12, 26]. Several works address the problem of modelling and recognizing human motion by learning the structure of the low dimensional manifold where it resides, and by recovering a mapping between the high dimensional observations and this manifold.

In [7] the authors consider MoCap sequences and they learn the structure of a unidimensional smooth manifold by applying the tensor voting technique [13]. A motion distance score is used to compute the similarity between the actions recorded

in two different sequences. The setting provides the possibility to compare also actions extracted from videos with actions taken from MoCap sequences.

In [34] the authors consider a two dimensional manifold with a toroidal topology in order to estimate human motion. They build on the idea of Gaussian Process Latent Variable Models (GP-LVM) [9] to identify a manifold which jointly captures gait and pose, via three different models. They introduce a new model (JGPM) which they compare to two constrained latent variable models based on GP-LVM and Local Linear GP-LVM [29] respectively.

In [23] the authors propose a non-linear generative model for human motion data that considers binary latent variables. The introduced architecture makes on-line inference efficient and allows for a simple approximate learning procedure. The method performance is evaluated by synthesizing various motion sequences and by performing on-line filling in of data, lost during motion capture.

Following a different perspective, in [21] the authors explore the space of actions, spanned by a set of action-bases, to identify some action invariants with respect to viewpoint, execution rate and subject's body shape. Action recognition is performed for four different kind of actions (sitting, standing, running and walking) and the results show that it is possible to correctly classify most of these actions using the proposed method.

The redundancy of the original representation of MoCap sequences is also exploited in [11] where a compressive sensing method is introduced. Here the authors argue that human actions are sparse in the action space domain as well as the time domain, and they seek therefore a sparse representation. The sparse representation introduced can assist in different applications regarding MoCap data like motion approximation, compression, action retrieval and action classification.

Finally, in [32] (see also [30, 33]) the authors examine whether and to what extent the use of information about the subject's pose assists recognition. In this case, several pose-based features are used, based on the relative pose features introduced in [14, 15]. Their results suggest that knowing the pose of the subject leads to better results, in terms of classification rate. It is also shown that pose based features alone are usually sufficient, as their combination with appearance based features is usually not leading to higher classification rate.

3 Gaussian Process Latent Variable Models

In this section we review Gaussian Process Latent Variable Models [9]. A Gaussian process is a collection of random variables such that any finite collection of them has a Gaussian distribution [19]. Namely, a random variable of a Gaussian process is $f(\mathbf{x}_i) = \mathcal{GP}(\mu(\mathbf{x}_i), k(\mathbf{x}_i, \mathbf{x}_j))$, with μ and $k(\mathbf{x}, \mathbf{x}')$ the mean and covariance function of the process respectively, indexed over the set $\mathcal{X}$ of all the possible inputs. The Gaussian process is a non parametric prior for the random variable $f(\mathbf{x}_i)$ where $\mathbf{x}_i$ is the deterministic input. Gaussian processes have been successfully used for both regression and classification tasks.

In [9] the author shows that Principal Component Analysis (PCA) can be interpreted as a product of Gaussian processes mapping latent-space points to points in data-space, when the covariance function is linear; when instead a non-linear covariance function is used, such as an RBF kernel then the mapping is non-linear. Lawrence shows the advantages in using Gaussian Processes Latent Variable Models (GP-LVM); for example, for optimization purposes, the data can be divided in active and inactive, according to some rule. Then, because points in the inactive set project into the data-space as Gaussian distributions, due to the properties of the variance the likelihood of each data point can be optimized independently.

In addition to the advantage in terms of visualization and computational efficiency highlighted in [9], GP-LVM turns out to be a powerful unsupervised learning algorithm. Indeed, GP-LVM can manage, via the non-linear mapping of the latent variables to the data-space, noisy or incomplete input data, when Gaussian processes are used as non parametric priors for them.

At this point, we introduce some preliminary definitions that we will refer throughout the following sections

Let $\mathbf{Y}$ be the normalized data in $\mathbb{R}^{N\times d}$, for example specifying the pose of a subject in space, with respect to a coordinate frame; let $\mathbf{X}$ be the mapped positions in latent-space, with $\mathbf{X} \in \mathbb{R}^{N\times q}$, with $q \leq d$. Let f be a mapping, such that:

$$y_{nj} = f(\mathbf{x}_n, \mathbf{w}_j) + \epsilon_{nj}, \tag{1}$$

Here, y_{nj} is the observed element of the nth row and jth column of $\mathbf{Y}$, ϵ_{nj} denotes the noise affecting the mapping and $\mathbf{x}_n$, the nth row of $\mathbf{X}$, and $\mathbf{w}_j$ are the parameters of the mapping f. Given a Gaussian process as a prior on f, when the prior is the same on each of the f functions one obtains [9]:

$$p(\mathbf{Y}|\mathbf{X}, \theta) = \prod_{j=1}^{d} \mathcal{N}(\mathbf{y}_j|\mathbf{0}, \mathbf{K}) \tag{2}$$

Here, $\mathbf{y}_j$ is the jth column of $\mathbf{Y}$ and $\mathbf{K}$ is the $N \times N$ kernel of the Gaussian process. We see that (2) suggests a conditional independence in the data space, given the latent space representation.

Learning amounts to maximizing the likelihood of the position of the latent variables $\mathbf{X}$ and θ, which are the parameters of the kernel:

$$L(\mathbf{X}, \theta) = -\frac{d}{2}\log|\mathbf{K}| - \frac{1}{2}\mathrm{Tr}\left(\mathbf{K}^{-1}\mathbf{Y}\mathbf{Y}^{\top}\right) \tag{3}$$

In order to optimize the non-linear model, it is necessary to initialize the model using appropriate initial values for the positions of the latent-space points. It is also necessary to initialize the hyperparameters of the model. Optimization is obtained by an iterative minimization of the objective function, by using a gradient based algorithm. As the model is non-linear, the hypersurface is subject to local-minima,

so the initialization of the positions of the latent-space points is crucial. When non-linear dimensionality reduction methods are used for the initialization, like local linear embedding (LLE) [20] or ISOMAP [24], the structure of the manifold is expected to be more accurately recovered. GP-LVM have been exploited in many applications as for example in [27–29, 31].

4 Discriminative Sequence Back-Constrained GP-LVM

As mentioned in the previous sections, models from the family of GP-LVM methods are well suited for predicting missing values or missing samples of time sequences. However, they do not seem to perform equally well when they are used for clustering and classification problems, particularly for time-series data. This drawback of the classical GP-LVM methods can be also witnessed by observing that it is hard to recover the structure of a common latent-space for a set of sequences, as their latent space representations are scattered across the latent-space and no relation can be drawn between sequences corresponding to the same action. This is due to the fact that standard GP-LVM models do not provide a mechanism to encourage points to be placed closer to each other in the latent-space when they belong to the same class and the same also holds at the level of individual sequences.

Local distances can be directly used in GP-LVM to provide a common latent-space representation as they are well suited for classification purposes. In fact local distances in data-space provide some information regarding the intra-class variation. Lawrence and Quiñonero-Candela in [10] have introduced Back-Constrained GP-LVM which considers local distances in the data-space. The GP-LVM model uses a product of Gaussian processes to map from the latent-space to the data-space. Each of these processes refers to a different dimension of the data-space and it is governed by the coordinates of the latent-points. In order to obtain a smooth mapping in the opposite direction, the authors in [10] propose to construct this mapping by means of a kernel based regression. Adopting this technique, the latent points are constrained to be the product of a smooth mapping from the data-space. This forces small distances in data-space to lead to small distances between the corresponding points in the latent-space. The smoothness of the mapping from the data-space to the latent-space is determined by the kernel function. Using this mapping, it is not needed to perform a new optimization to approximate the latent-space representation of new data.

The previous method cannot be directly applied on data originating from sequences, as it is expected that individual elements of a sequence do not provide sufficient information regarding the characteristics of the entire sequence. Building on the same principle, namely the use of local distances in the data-space as back-constraints, we formulate a GP-LVM variant which considers entire sequences rather than individual data points.

Before introducing our model, we briefly review the Dynamic Time Warping (DTW) algorithm, as well as a set of sequence alignment kernels based on DTW and its variants, which will be used for the derivation of our model.

4.1 Dynamic Time Warping and Sequence Alignment Kernels

Dynamic Time Warping is used to match two time dependent sequences by nonlinearly warping one sequence onto the other. Let us consider two vector sequences $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_N)$ with $N \in \mathbb{N}$ and $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_M)$ with $M \in \mathbb{N}$. Each vector in the sequence belongs to a n-dimensional feature space $\mathcal{F}$ so $\mathbf{y}_n, \mathbf{z}_m \in \mathcal{F}$. A local distance measure is defined to compare a pair of features, provided by an appropriate kernel function:

$$\kappa : \mathcal{F} \times \mathcal{F} \to \mathbb{R}^+ \tag{4}$$

A warping path is a sequence $p = (p_1, \ldots, p_L)$ where each element is a pair $p_l = (n_l, m_l)$. The total cost of a warping path p, according to the predefined distance measure, is:

$$c_p(\mathbf{y}_n, \mathbf{z}_m) = \sum_{l=1}^{L} \kappa(\mathbf{y}_{n_l}, \mathbf{z}_{m_l}) \tag{5}$$

The Dynamic Time Warping distance between two sequences is defined as the minimal total cost among all possible warping paths. To obtain this value we have to solve the following optimization problem:

$$DTW(\mathbf{Y}, \mathbf{Z}) = \min_p \left\{ c_p(\mathbf{Y}, \mathbf{Z}) \right\} \tag{6}$$

We can also identify an optimal warping path (not necessarily unique):

$$p^* = \arg\min_p \left\{ c_p(\mathbf{Y}, \mathbf{Z}) \right\} \tag{7}$$

The DTW distance is well-defined, even though there may exist many warping paths of minimal total cost. Moreover, it is symmetric if the distance measure is also symmetric, but it is not a proper metric, as it does not satisfy the triangle inequality. In order to apply DTW on MoCap sequences, we must first define the local cost measure κ. Two popular choices are to use the sum of the geodesic distances between the unit-quaternions representing the joint angles, as well as the optimal alignment distance between the three dimensional positions of the joints [14].

Based on the notions of the DTW distance and the optimal warping path, alignment kernels have been proposed which consider entire sequences as a whole. As an example we cite here [2, 5, 22].

4.2 Sequence Back-Constrained GP-LVM

In this section we show how to enforce a clustering of the sequences in the latent-space, governed by their respective similarity, which will enable a more accurate classification of a new sequence. To ensure that data instances which are close to each other in the data-space, are mapped to positions which are close also in the latent-space, we apply a similarity measure for comparing different sequences and identify a characteristic feature, summarizing the entire sequence.

Here we consider that each frame of a motion sequence is represented as a d-dimensional array. An entire sequence, with index s, is represented thus as a set of d dimensional arrays of cardinality L_s, forming a matrix $\mathbf{Y}_s \in \mathbb{R}^{L_s \times d}$. A collection of S motion sequences is represented as the concatenation of the respective sub-matrices forming the data-matrix $\mathbf{Y} \in \mathbb{R}^{N \times d}$, with $N = \sum_{s=1}^{S} L_s$. Let $\mathcal{J}_s$ be the set of indices of the sth sequence in the data matrix, the corresponding representation of the data-points in the q dimensional latent-space form a matrix $\mathbf{X} \in \mathbb{R}^{N \times q}$. The coordinates of the centroid of the latent-space representation of the sth sequence, is defined as:

$$\mu_{sq} = \frac{1}{L_s} \sum_{n \in \mathcal{J}_s} x_{nq} \tag{8}$$

The likelihood of the GP-LVM model is given by (3). The centroid of the latent positions of the data points is taken to be the characteristic feature of the sequence. Therefore, we require that the local distances between the sequences in data-space, computed via the DTW technique, are preserved in latent-space; thus they are specified as the distances between the centroids $\boldsymbol{\mu}_s$. Hence, we consider a mapping to the latent-space governed by an alignment kernel k:

$$g_q(\mathbf{Y}_s) = \sum_{m=1}^{S} a_{mq} k(\mathbf{Y}_s, \mathbf{Y}_m) \tag{9}$$

The degree to which the local distances in the data-space are preserved depends on the particular characteristics of the kernel employed for the mapping.

We, thus, have to maximize a constrained likelihood, instead of maximizing the likelihood of the original GP-LVM model.

Each of the $S \cdot q$ constraints can be written as:

$$g_q(\mathbf{Y}_s) - \mu_{sq} = 0 \tag{10}$$

Maximizing the constrained likelihood of the model, we expect to obtain a latent-space representation, where similar sequences are better grouped together, with respect to the representation obtained by the original model. Another important advantage of this approach is that we can use the inverse mapping recovered in the

learning phase, for the purposes of fast inference. In this way, we avoid the costly operation of reoptimisation, which is otherwise necessary to obtain the latent-space representation of new sequences.

Up to this point, we did not consider the labels of each type of sequence. In the following section, we modify our model by replacing the Gaussian prior with a prior which will make the model more discriminative.

4.3 Discriminative Sequence Back-Constrained GP-LVM

Discriminative GP-LVM (D-GPLVM) has been originally introduced in [27]. In order to make the Sequence Back-Constrained GP-LVM (SB-GPLVM) model more discriminative, we can consider a measure of the between-group variation and the within-group separation. Referring to Fisher's Discriminant Analysis, in case we need to estimate a linear projection of the data, such that an optimal separation is achieved, we need to maximize the ratio of the *between-group-sum of squares* to the *within-group-sum of squares*.

We thus seek the direction of projection given by the vector $\mathbf{a}$ which provides a good separation of the data. Denoting as $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^{\mathrm{T}}$ the low dimensional representation of the data points $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_N]^{\mathrm{T}}$, the *between-group-sum of squares* is given as:

$$\mathbf{a}^{\mathrm{T}}\mathbf{B}\mathbf{a} = \sum_{c=1}^{C} \frac{N_c}{N} \mathbf{a}^{\mathrm{T}}(\mu_c - \mu_0)(\mu_c - \mu_0)^{\mathrm{T}}\mathbf{a} \tag{11}$$

The *within-group-sum of square* is given as:

$$\mathbf{a}^{\mathrm{T}}\mathbf{W}\mathbf{a} = \frac{1}{N} \sum_{c=1}^{C} \sum_{n=1}^{N_c} \mathbf{a}^{\mathrm{T}}(\mathbf{x}_n^{(c)} - \mu_c)(\mathbf{x}_n^{(c)} - \mu_c)^{\mathrm{T}}\mathbf{a} \tag{12}$$

Here $\mathbf{X}^{(c)} = [\mathbf{x}_1^{(c)}, \ldots, \mathbf{x}_{N_c}^{(c)}]^{\mathrm{T}}$ are the N_c points which belong to the class c, $\boldsymbol{\mu}_c$ is the mean of the elements of class c and $\boldsymbol{\mu}_0$ is the mean computed across all the points.

The criterion used for maximizing between-group separability and minimizing within-group variability is the following [8]:

$$J(\mathbf{X}) = \mathrm{Tr}(\mathbf{W}^{-1}\mathbf{B}) \tag{13}$$

Based on the previous discussion, in order to transform the SB-GPLVM model making it discriminative, it is necessary to replace the Gaussian prior with a prior which depends on (13). This prior takes the following form:

$$p(\mathbf{X}) = \frac{1}{\alpha} \exp\left\{-\frac{\gamma}{2} J^{-1}\right\} \tag{14}$$

where α is a normalization constant, possibly depending on p, and γ represents the scaling factor of the prior.

The log likelihood associated with the discriminative model becomes:

$$L = -\frac{d}{2}\log|\mathbf{K}| - \frac{1}{2}\mathrm{Tr}\left(\mathbf{K}^{-1}\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\right) - \frac{\gamma}{2}\mathrm{Tr}\left(\mathbf{B}^{-1}\mathbf{W}\right) \tag{15}$$

The parameter γ controls the relative importance of the discriminative prior and it reflects the ability of the model to be more discriminative or more generalizing, according to the value it takes.

4.4 Classification Based on D-SBGPLVM

In this section we illustrate how to compute the latent representation of the data points belonging to a new sequence. This will allow to classify any new sequenced according to the introduced D-SBGPLVM model. Let $\mathbf{Y}_*$ be the data-space representation of a new sequence and $\mathbf{X}_*$ the corresponding latent-space representation. The new sequence's centroid in latent-space can be estimated orders of magnitude faster than $\mathbf{X}_*$ by making use of Eq. (9) introduced in Sect. 4.2. Thus, the coordinates of the test sequence centroid, in each dimension of the latent space are given by:

$$\forall\, q:\ \mu_{*q} = g_q(\mathbf{Y}_*) = \sum_{s=1}^{S} a_{qs}\, k(\mathbf{Y}_*, \mathbf{Y}_s) \tag{16}$$

where μ_{*q} is the qth dimension coordinate of the centroid $\boldsymbol{\mu}_*$ of the test sequence. In this case, no minimization is required and the time, necessary for computing the coordinates of the centroid of the test sequence, is proportional to the time needed to compute the kernel values.

At this point, any multi-class classification method can be employed, in order to perform classification. As the latent-space has a dimensionality much smaller than the original data-space, it is expected that classification is more robustly performed in the latent representation of the sequences. Moreover, the proposed method provides a concise way to classify sequences as a whole, as the model treats them explicitly as individual entities.

5 Results

The ability of the Discriminative Sequence Back-Constrained GP-LVM model to provide a latent-space representation suitable for efficient and robust classification of sequences, is examined in this section.

Evaluation on the HDM05 "Cuts" *Dataset* [16]. Part of the "Cuts" sequences, contained in the HDM05 dataset, has been used for evaluating the model we propose, in comparison to other methods which can be used for sequence classification. This dataset includes the following actions: *Clapping hands*-5 repetitions (17 sequences); *Hopping on right leg*-3 reps. (12 seqs.); *Kick with right foot in front*-2 reps. (15 seqs.); *Running on place*-4 steps (15 seqs.); *Throwing high with right hand while standing* (14 seqs.); *Walking starting with right foot*-4 steps (16 seqs.).

The sequences are sampled at a frequency of 120 frames per second. For this dataset, sequences are already accurately segmented, in order to contain a single action with the same number of repetitions.

The results of the proposed method are compared with the classification results, obtained by directly using the DTW distances of the sequences in the data-space, as well as using the highest class-conditional densities obtained by the Variational Gaussian Process Dynamical Systems (V-GPDS) method [6]. All results are taken by Cross-Validation. Each experiment is performed by keeping all action sequences of one of the five subjects as test sequences and by using the sequences of the other four subjects as training instances. Finally, the results are averaged over the five individual experiments.

Table 1 gives the accuracy rate achieved with each of these three methods for each action as well as in average. Regarding the results obtained by the proposed method, relative features are used and the dimensionality of the latent-space space is fixed to four. Moreover, for the back-constraints the kernel proposed in [2] is used and the initial positions of the latent points are obtained by using the Local Linear Embedding algorithm [20]. Finally, classification in latent-space is performed by SVMs using the RBF kernel function. Figure 1 shows the corresponding confusion matrix obtained by using the D-SBGPLVM model.

One can see from the results provided in Table 1 that our method gives the best results, both for each individual type of action, except for Hop, as well as in average. We observe that the classification accuracy is relatively high for the DTW distance alone. This depends also on the fact that this dataset is specifically constructed in such a way, that actions of the same kind can be aligned with a very small cost. This is possible as they are defined at a high detail level regarding their execution and they

Table 1 Comparison of the classification results for the HDM05 "Cuts" dataset

	DTW (%)	V-GPDS (%)	D-SBGPLVM (%)
Clap	70.6	16.7	88.2
Hop	100	66.7	83.3
Kick	40.0	33.3	53.3
Run	66.7	33.3	80.0
Throw	64.3	50.0	78.6
Walk	100	83.3	100
Average	73.0	47.2	80.9

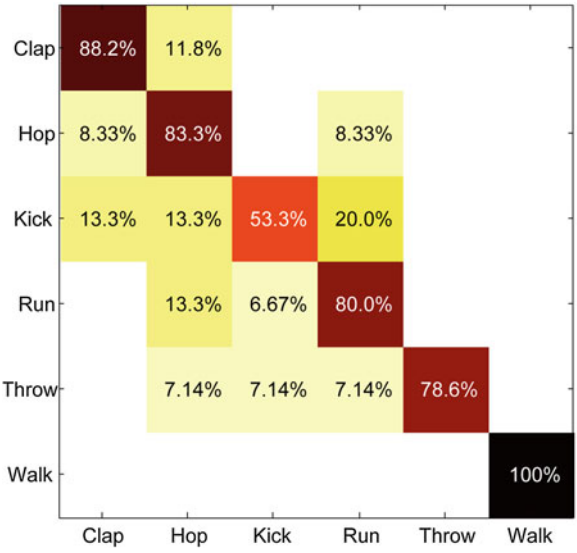

Fig. 1 Confusion matrix by using D-SBGPLVM model in combination with SVM on the HDM05 "Cuts" dataset. Average accuracy: 80.9 %

have been also accurately segmented manually. Regarding classification of human actions using the V-GPDS model, it is necessary to train a different model for each individual type of action. After a model has been trained for each type of action, it is possible to compute the class conditional densities for the new sequence.

Considering that the analogous model of V-GPDS, which does not consider time dynamics introduced in [25], provides good classification results (e.g. on the USPS Handwritten Digits Dataset) we expected higher classification rates for the adapted V-GPDS model. Searching the cause of this issue, we have noticed that models for certain actions tend to provide quite high conditional densities most of the time. Further investigation is needed in this direction, as the experiments performed using V-GPDS were not sufficient to derive safe conclusions and possibly a more suitable adaptation of the model for classification purposes is needed.

In the case of D-SBGPLVM, the model is trained by optimising the latent coordinates of the sequences and the hyper-parameters of the model by using all training sequences. By the optimisation process, we recover also the parameters of the kernel based regression, which forms the inverse mapping from the data-space to the latent-space. We provide some examples of bi-dimensional latent-spaces recovered by training the model using sequences of the HDM05 "Cuts" dataset in Fig. 2. In these figures, each color corresponds to a different class of action, crosses are the latent representations for each individual data point, triangles correspond to the centroids of the training sequences and finally the squares correspond to the estimated position of the testing sequences centroids computed using the back-constraints. In Fig. 2 the recovered latent-spaces are shown for three different types of representa-

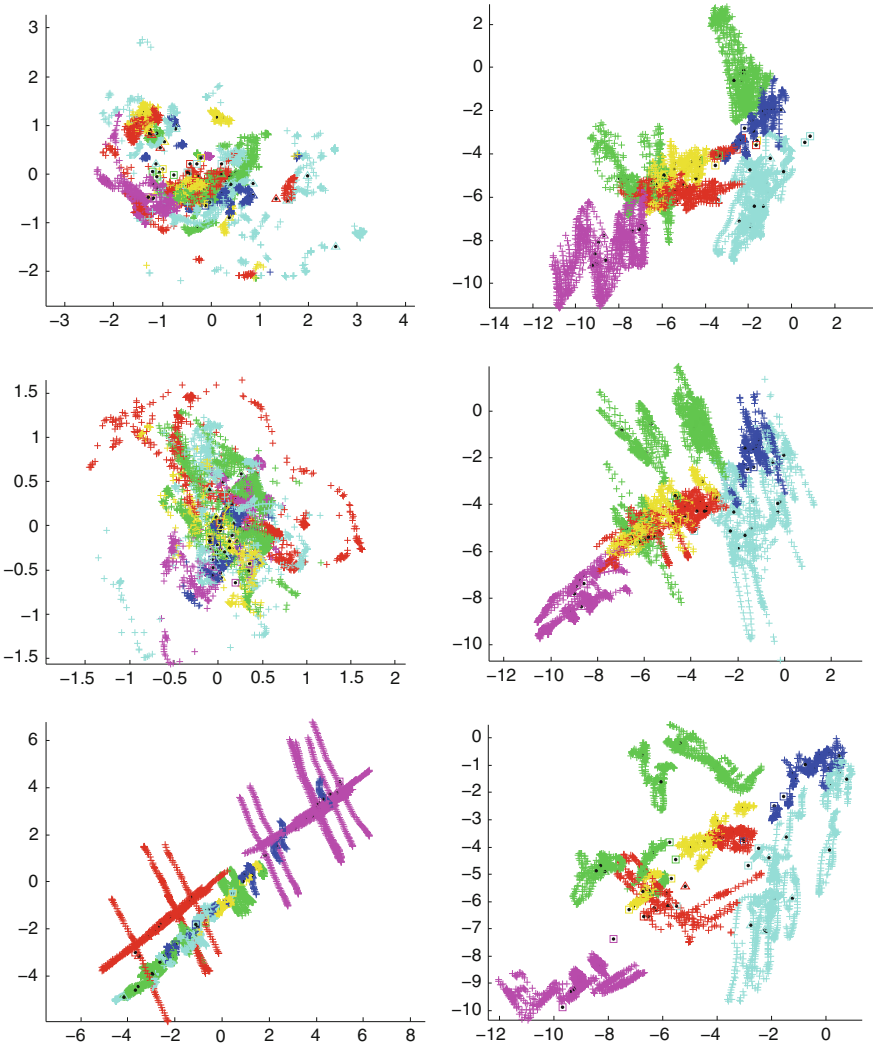

Fig. 2 *Left Column* Latent-space representation by PPCA initalization and considering Euler Angles (*Top*), Unit-Quaternion (*Middle*) and 3D Point Cloud (*Bottom*) representations *Right Column* latent-space representation considering relative features representation and PPCA (*Top*), LLE (*Middle*) and ISOMAP (*Bottom*) initalization

tions considered for the sequences and by using Probabilistic PCA in order to retrieve initial values for the latent points. In the case of Euler Angles and Unit-Quaternions, one can notice that different sequences are placed on top of each other and thus we expect classification rates to be low.

Our interpretation is that this mainly depends on the high non-linearity of the data-space and the fact the PPCA, being a linear dimensionality reduction technique, is not

able to provide suitable initial values for the latent points. As our model is non-linear and it is optimized by using a gradient based algorithm, it is susceptible to local minima. However, in the case of 3D point cloud representation, the data-space does not show excessive non-linearity and even PPCA initialization seems to be sufficient to recover a better structure for the latent-space.

The case of Relative Features (as in [14], but without discretization based on some threshold) is examined also in Fig. 2. Relative features include for example the distance between two specified joints, the distance of a joint with respect to the plane defined by three other, the angle between two successive joints etc. Here we can better observe the impact of the initialization technique on the resulting structure of the latent-space. It is evident that the use of more sophisticated non-linear dimensionality reduction techniques to obtain the initial values, helps recovering a better structure of the common latent-space.

Evaluation on actions of the CMU Dataset [4]. Seven actions from the CMU dataset have been also considered for evaluating the model we propose. This dataset includes the following actions: *Walking* (15 sequences); *Running* (15 seqs.); *Jumping* (15 seqs.); *Sitting-Standing* (7 seqs.); *Throwing-Tossing* (15 seqs.); *Boxing* (9 seqs.); *Dancing* (9 seqs.).

Each of these actions is performed from a different actor. Moreover, the actions have not been hand-picked and their label only relies on the default labelling provided by the publishers of the dataset. Finally, motion sequences have not been manually segmented. We perform classification instead by just considering the first two seconds of each sequence. For these reasons, we can see that this dataset represents a more challenging and realistic instance of the action recognition problem. Five-fold cross-validation has been used here for obtaining the final classification results.

The classification accuracy achieved by the proposed method, compared with the results of DTW distances and V-GPDS method, are provided in Table 2. Here, Euler angles are considered as features provided to the D-SBGPLVM, while the rest of the setting is the same with the one described for the "Cuts" experiments. In Fig. 3 we provide the corresponding confusion matrix and the overall classification rate, when the D-SBGPLVM model is used.

Table 2 Comparison of the classification results for the actions taken from CMU dataset

	DTW (%)	V-GPDS (%)	D-SBGPLVM (%)
Walk	80.0	40.0	66.7
Run	60.0	40.0	66.7
Jump	86.7	40.0	73.3
Throw-Toss	80.0	40.0	80.0
Sit-Stand	46.7	40.0	80.0
Box	100	20.0	80.0
Dance	26.7	80.0	73.3
Average	63.5	42.9	72.9

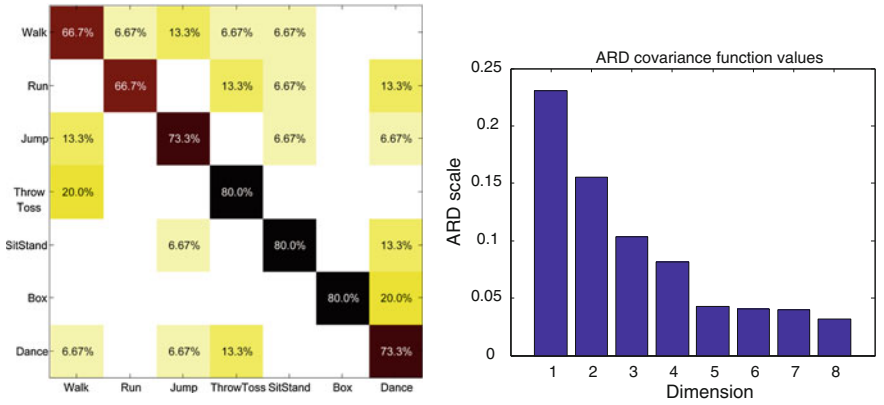

Fig. 3 *Left* Confusion matrix by using D-SBGPLVM model in combination with SVM on the CMU dataset. Average accuracy: 72.9 % *Right* sorted ARD covariance function values obtained after training the model for the same dataset using the RBF-ARD kernel. Average accuracy: 74.1 %

We can observe here, that the results for the "CMU" dataset are analogous to the ones corresponding to the "Cuts" dataset. We expect that the lower rate achieved in general by all algorithms mainly depend on the particular difficulties which characterise this dataset, as mentioned above. Considering these difficulties, one can see that the proposed model gives satisfying classification results. This also demonstrates the generalization capabilities of the proposed probabilistic model, which based on this characteristic leads to an overall accuracy that exceeds the accuracy achieved by the other two methods considered here.

The same experiments were also performed by considering the recently proposed 'path kernel' [3] providing equivalent results. The classification rate was slightly lower but this may be related to the particular selection of the parameters of the kernel. Moreover, we performed trials using the automatic relevance determination (ARD) squared exponential kernel as in [6, 25]. In this case, considering eight dimensions for the latent space, we obtained a classification rate of 74.1 % for the CMU dataset. What is important to note here are the values of the relative importance of each dimension after training the model, shown in Fig. 3. One can see here that most of the information for the actions is embedded in a four dimensional sub-manifold. This result is in accordance with the ones reported in [17].

6 Conclusions

In this paper, we have introduced a novel GP-LVM variant in order to recover the structure of a lower dimensional manifold for a set of sequences of different action types. We have shown that the model, according to our approach, attains increased classification accuracy by working in the low dimensional latent-space instead of

the original data-space. By exploiting the inverse mapping, from the data-space to the latent-space, our approach is able to infer the class of a new sequence within a few seconds (Matlab implementation tested on the following system: 2.2 GHz quad-core AMD Phenom, 4 GB RAM). This provides a crucial advantage with respect to other GP-LVM models which require several minutes to complete this task, having to deal with a new optimization to obtain the latent-space representation of the new data instances. We have further shown that the proposed D-SBGPLVM model attains classification rates equivalent to the current state-of-the-art when combined with a standard classifier, as for example SVM, for classification in the latent-space.

Within the directions of our future work, we further consider the combination of the proposed method with some pose recovery algorithm. In this way, it would be possible to train the model by using action sequences taken from a MoCap dataset and classify sequences recovered from videos by means of the pose recovery algorithm. This would make action recognition from 2D video sequences also possible. Finally, we are currently considering automated ways for the segmentation of motion sequences to sub-sequences of individual actions without prior knowledge of the actions preformed. This step is important for allowing the processing of sequences containing multiple actions with the method described in this work.

Acknowledgments This paper describes research done under the EU-FP7 ICT 247870 NIFTi project.

References

1. Aggarwal, J.K., Cai, Q.: Human motion analysis: a review. Comput. Vis. Image Underst. **73**(3), 428–440 (1999)
2. Bahlmann, C., Haasdonk, B., Burkhardt, H.: On-line handwriting recognition with support vector machines-a kernel approach. In: International Workshop on Frontiers in Handwriting Recognition, pp. 49–54 (2002)
3. Baisero, A., Pokorny, F.T., Kragic, D., Ek, C.H.: The path kernel. In: Proceedings of the International Conference on Pattern Recognition Applications and Methods (2013)
4. CMU: Carnegie-mellon mocap database, http://mocap.cs.cmu.edu/ (2003)
5. Cuturi, M., Vert, J.P., Birkenes, O., Matsui, T.: A kernel for time series based on global alignments. Comput. Res. Repos. (2006)
6. Damianou, A.C., Titsias, M.K., Lawrence, N.D.: Variational gaussian process dynamical systems. In: Neural Information Processing Systems Conference, pp. 2510–2518 (2011)
7. Gong, D., Medioni, G.: Dynamic manifold warping for view invariant action recognition. In: International Conference on Computer Vision (2011)
8. Härdle, W., Simar, W.: Applied Multivariate Statistical Analysis. Springer, New York (2003)
9. Lawrence, N.D.: Gaussian process latent variable models for visualisation of high dimensional data. In: Neural Information Processing Systems Conference (2003)
10. Lawrence, N.D., Candela, J.Q.: Local distance preservation in the gp-lvm through back constraints. In: International Conference on Machine learning, pp. 513–520 (2006)
11. Li, Y., Fermüller, C., Aloimonos, Y., Ji, H.: Learning shift-invariant sparse representation of actions. In: International Conference on Computer Vision and Pattern Recognition, pp. 2630–2637 (2010)
12. Moeslund, T.B., Hilton, A., Krüger, V.: A survey of advances in vision-based human motion capture and analysis. Comput. Vis. Image Underst. **104**(2–3), 90–126 (2006)

13. Mordohai, P., Medioni, G.G.: Dimensionality estimation, manifold learning and function approximation using tensor voting. J. Mach. Learn. Res. **11**, 411–450 (2010)
14. Müller, M.: Information Retrieval for Music and Motion. Springer, Heidelberg (2007)
15. Müller, M., Röder, T., Clausen, M.: Efficient content-based retrieval of motion capture data. In: SIGGRAPH, pp. 677–685 (2005)
16. Muller, M., Roder, T., Clausen, M., Eberhardt, B., Krüger, B., Weber, A.: Documentation mocap database hdm05. Technical report CG-2007-2, Universität Bonn (2007)
17. Ntouskos, V., Papadakis, P., Pirri, F.: A comprehensive analysis of human motion capture data for action recognition. In: Proceedings of the International Conference on Computer Vision Theory and Applications, pp. 647–652 (2012)
18. Poggio, T.: Early vision: from computational structure to algorithms and parallel hardware. Comput. Vis. Graph. Image Process. **31**(2), 139–155 (1985)
19. Rasmussen, C., Williams, C.: Gaussian processes for machine learning. Adaptive Computation and Machine Learning. MIT, Cambridge (2006)
20. Roweis, S., Saul, L.: Nonlinear dimensionality reduction by locally linear embedding. Science **290**(5500), 2323–2326 (2000)
21. Sheikh, Y., Sheikh, M., Shah, M.: Exploring the space of a human action. Int. Conf. Comput. Vis. **1**, 144–149 (2005)
22. Shimodaira, H., Noma, K., Nakai, M., Sagayama, S.: Dynamic time-alignment kernel in support vector machine. Neural Inf. Process. Syst. Conf. **2**, 921–928 (2001)
23. Taylor, G.W., Hinton, G.E., Roweis, S.T.: Modeling human motion using binary latent variables. In: Neural Information Processing Systems Conference, pp. 1345–1352 (2006)
24. Tenenbaum, J.B., Silva, V.D., Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction. Science (2000)
25. Titsias, M.K., Lawrence, N.D.: Bayesian gaussian process latent variable model. J. Mach. Learn. Res. Proc. Track **9**, 844–851 (2010)
26. Turaga, P.K., Chellappa, R., Subrahmanian, V.S., Udrea, O.: Machine recognition of human activities: a survey. IEEE Trans. Circuits Syst. Video Technol. **18**(11), 1473–1488 (2008)
27. Urtasun, R., Darrell, T.: Discriminative gaussian process latent variable model for classification. In: International Conference on Machine Learning, pp. 927–934 (2007)
28. Urtasun, R., Fleet, D.J., Fua, P.: 3d people tracking with gaussian process dynamical models. In: International Conference on Computer Vision and Pattern Recognition, pp. 238–245 (2006)
29. Urtasun, R., Fleet, D.J., Geiger, A., Popovic, J., Darrell, T., Lawrence, N.D.: Topologically-constrained latent variable models. In: International Conference on Machine Learning, pp. 1080–1087 (2008)
30. Waltisberg, D., Yao, A., Gall, J., Van Gool, L.: Variations of a hough-voting action recognition system. In: International conference on Pattern Recognition, pp. 306–312 (2010)
31. Wang, J.M., Fleet, D.J., Hertzmann, A.: Gaussian process dynamical models. Neural Inf. Proc. Syst. Conf. **18**, 1441–1448 (2006)
32. Yao, A., Gall, J., Fanelli, G., Gool, L.V.: Does human action recognition benefit from pose estimation? In: British Machine Vision Conference, pp. 67.1–67.11 (2011)
33. Yao, A., Gall, J., Gool, L.J.V.: A hough transform-based voting framework for action recognition. In: International Conference on Computer Vision and Pattern Recognition, pp. 2061–2068 (2010)
34. Zhang, X., Fan, G.: Joint gait-pose manifold for video-based human motion estimation. In: European Conference on Computer Vision, pp. 47–54 (2011)