HAL
archives-ouvertes.fr

Discriminative Sequence Back-constrained GP-LVM for MOCAP based Action Recognition

Valsamis Ntouskos, Panagiotis Papadakis, Fiora Pirri

▶ **To cite this version:**

Valsamis Ntouskos, Panagiotis Papadakis, Fiora Pirri. Discriminative Sequence Back-constrained GP-LVM for MOCAP based Action Recognition. International Conference on Pattern Recognition Applications and Methods, Feb 2013, Barcelona, Spain. 10.5220/0004268600870096 . hal-00832895

HAL Id: hal-00832895
https://hal.inria.fr/hal-00832895

Submitted on 11 Jun 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Discriminative Sequence Back-Constrained GP-LVM for MOCAP Based Action Recognition

Valsamis Ntouskos, Panagiotis Papadakis, Fiora Pirri
ALCOR Lab, Department CCME, Sapienza, University of Rome
{*ntouskos,papadakis,pirri*}*@dis.uniroma1.it*

Keywords: Action Recognition; Motion Capture Sequences; Manifold Learning; GPLVM.

Abstract: In this paper we address the problem of human action recognition within Motion Capture sequences. We introduce a method based on Gaussian Process Latent Variable Models and Alignment Kernels. We build a new discriminative latent variable model with back-constraints induced by the similarity of the original sequences. We compare the proposed method with a standard sequence classification method based on Dynamic Time Warping and with the recently proposed V-GPDS model (Damianou et al., 2011). The proposed methodology exhibits high performance even for datasets that have not been manually preprocessed while it further allows fast inference by exploiting the back constraints.

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 an observed motion sequence, hence it requires the solution of an inverse problem (Poggio, 1985). Furthermore, the complete modelling process is composed of several intermediate stages, namely: data acquisition, that in general requires a sophisticated technology, motion analysis and segmentation into individual actions, alignment between sequences and classification with respect to a given taxonomy. While all these stages are computationally expensive the main goal remains to obtain real-time recognition.

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 Gaussian Process Latent Variable Model (GP-LVM) and the Back-Constrained GP-LVM introduced in (Lawrence, 2003) and (Lawrence and Quiñonero Candela, 2006) respectively, 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, from motion capture (MoCap) data, and constrains feature distances in data space to be suitably projected in the latent space, in order to preserve the clustering of common patterns. This ensures a discriminative power to the GP-LVM model and it also exploits the characteristic property of action sequences of being reducible to a lower dimensional manifold (Ntouskos et al., 2012).

We organize the remainder of this paper as follows: In Section 2 we briefly review recent works on action recognition based on MoCap sequences and dimensionality reduction, showing the major trends of research in this field. In Section 3 we unfold the theoretical foundation of GP-LVM on which our model is based, in Section 4 we present our discriminative model and in Section 5 we demonstrate the latent space structure recovered by the proposed model and examine 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 (Damianou et al., 2011) recently proposed for modelling high dimensional dynamical systems. We conclude our work by discussing possible extensions.

2 RELATED WORK

The problem of human action recognition has been addressed from a plurality of perspectives that range

from stochastic and volumetric to non-parametric models, the latter being most commonly employed. Extended reviews on human motion analysis and action recognition can be found in (Aggarwal and Cai, 1999), (Moeslund et al., 2006) and (Turaga et al., 2008).

A distinctive branch of research concerns approaches that 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 the manifold. Our focus mainly resides onto this category of methods and in particular to those wherein actions have been captured as MoCap sequences, a data representation that has lately been gaining momentum and led to the proliferation of human action repositories. In the following, we briefly review a number of works that are representative of this area and within the same spirit of the proposed approach.

In (Gong and Medioni, 2011) the authors consider MoCap sequences and they learn the structure of a unidimensional smooth manifold by applying the tensor voting technique (Mordohai and Medioni, 2010). 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 (Zhang and Fan, 2011) 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) (Lawrence, 2003) 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 (Urtasun et al., 2008) respectively.

In (Taylor et al., 2006) 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's 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 (Sheikh et al., 2005) 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 a constrained set of actions 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 (Li et al., 2010) where a compressive sensing method is introduced. The authors argued that human actions are sparse in the action space domain as well as the time domain, and therefore they seeked a sparse representation. The introduced sparse representation could assist in different applications regarding MoCap data like motion approximation, compression, action retrieval and action classification.

Finally, in (Yao et al., 2011) (see also (Yao et al., 2010) and (Waltisberg et al., 2010)) 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 (Müller et al., 2005) and (Müller, 2007). 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 did not appear to increase classification rate.

3 GAUSSIAN PROCESS LATENT VARIABLE MODELS

A Gaussian process is a collection of random variables such that any finite collection of them has a Gaussian distribution (Rasmussen and Williams, 2006). 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, as it is assumed to be observed. Gaussian processes have been successfully used for both regression and classification tasks.

In (Lawrence, 2003) it is shown 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; instead, when 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, since 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 in-

Figure 1: Rendered poses of a subject performing a "fighting" action taken from a MoCap sequence (courtesy of (mocapdata.com, 2011)).

dependently.

In addition to the advantage in terms of visualization and computation, highlighted in (Lawrence, 2003) 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.

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 silhouette element in space (see Figure 1), 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:

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

Here, y_{nj} the element of the n-th row and j-th column of $\mathbf{Y}$, ε_{nj} denoting noise and $\mathbf{x}_n$ the n-th row of $\mathbf{X}$, and $\mathbf{w}_j$ 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 (Lawrence, 2003):

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

Here, $\mathbf{y}_j$ is the j-th column of $\mathbf{Y}$ and $\mathbf{K}$ is the $N \times N$ kernel of the Gaussian process. Equation 2 suggests a conditional independence of the data space dimensions given the latent space representation.

Learning amounts to maximizing the likelihood with respect to the position of the latent variables $\mathbf{X}$ and θ, 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 by setting the initial values for the positions of the latent-space points, and the hyperparameters of the model as well as perform an iterative minimization of the objective function, by using a gradient based optimization algorithm. As the model is non-linear, it 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) (Roweis and Saul, 2000) or ISOMAP (Tenenbaum et al., 2000), it is expected that the structure of the manifold is more accurately recovered. GP-LVM have been exploited in many applications as for example in (Urtasun et al., 2006), (Urtasun and Darrell, 2007), (Urtasun et al., 2008) and (Wang et al., 2006).

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 handicap of the classical GP-LVM methods can be also witnessed by observing the latent-space representations of the data.

One can notice that recovering the structure of a common latent-space of a set of sequences, their latent space representations are scattered across the latent-space and no relation is evident 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 neighboring points to be placed closer to each other in the latent-space, while the same also holds at the level of individual sequences.

In cases where local distances in data-space provide some information regarding the intra-class variation, these can be directly used in the GP-LVM model, in order to provide a common latent-space representation better suited for classification purposes. Lawrence and Quiñonero-Candela in (Lawrence and Quiñonero Candela, 2006) have introduced Back-Constrained GP-LVM which considers local distances in the data-space. The GP-LVM model provides a direct mapping from the latent-space to the data-space by means of a product of Gaussian processes. 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 (Lawrence and Quiñonero Candela, 2006) 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 enforces small distances in data-space to lead to small distances between the

neighboring points in the latent-space. The smoothness of the mapping from the data-space to the latent-space is determined by the kernel function. An interesting property in the construction of an inverse mapping from the data-space to the latent-space, is the possibility to estimate the latent-position of a new data point without the need of re-optimisation.

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 variations, which will be used for the derivation our model.

4.1 Dynamic Time Warping and Sequence Alignment Kernel

Dynamic Time Warping is used to match two time dependent sequences by nonlinearly warping the one against the other. Let us consider two vector sequences $\mathbf{Y} \doteq (\mathbf{y}_1, \ldots, \mathbf{y}_N)$ with $N \in \mathbb{N}$ and $\mathbf{Z} \doteq (\mathbf{z}_1, \ldots, \mathbf{z}_M)$ 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:

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

A warping path is a sequence $p = (p_1, \ldots, p_L)$ where each element is a tuple $p_l = (n_l, m_l)$. 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:

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

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

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

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 does not define 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 (Müller, 2007).

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 ((Shimodaira et al., 2001), (Bahlmann et al., 2002) and (Cuturi et al., 2006)).

4.2 Sequence Back-Constrained GP-LVM

In order to be able 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. Once these conditions are accommodated, we can 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.

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 by 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 $N = \sum_{s=1}^{S} L_s$. The set $\mathcal{J}_s$ contains the indices of the s^{th} sequence in the data matrix. The corresponding representation of the data-points in the q dimensional latent-space form a matrix $\mathbf{X} \in \mathbb{R}^{N \times q}$. We also consider the coordinates of the centroid of the latent-space representation of the s^{th} sequence, 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 Equation 3. The centroid of the latent positions of the data points are 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 μ_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.

Instead of maximizing the likelihood of the original GP-LVM model, now he have to maximize a constrained likelihood.

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 clustered 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 re-optimisation, which otherwise would be 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 (Urtasun and Darrell, 2007). 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 squares* 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, μ_c is the mean of the elements of class c and μ_0 is the mean computed over all the points.

The criterion used for maximizing between-group separability and minimizing within-group variability is the following (Härdle and Simar, 2003):

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

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 Equation (13). This prior takes the following form:

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

where C_p is a normalization constant 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 order to classify a new sequence according to the D-SBGPLVM model, it is necessary first to compute the latent representation of the data points belonging to the sequence. Let $\mathbf{Y}_*$ be the data-space representation of the new sequence. The corresponding latent-space representation $\mathbf{X}_*$ can be estimated by maximizing $p(\mathbf{Y}_*, \mathbf{X}_* | \mathbf{X}, \mathbf{Y}, \theta)$.

Alternatively we can perform inference by only using the position of the centroid of the latent representation of the test sequence. The new sequence's centroid in latent-space can be estimated orders of magnitude faster by making use of the sequence back-constraints introduced in Section 4.2 (Equation 9). Thus, the coordinates of the test sequence's centroid, in each dimension of the latent space are given by:

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

where μ_{*q} is the q^{th} dimension coordinate of the centroid μ_* 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 practically equal 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

In this Section, we evaluate the ability of the Discriminative Sequence Back-Constrained GP-LVM model to provide a latent-space representation, that allows robust and effective classification of human action sequences.

Evaluation on the HDM05 "Cuts" Dataset Part of the "Cuts" sequences, contained in the HDM05 (Muller et al., 2007) 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 repetitions - 12 sequences
- Kick with right foot in front, 2 repetitions - 15 sequences
- Running on place, 4 steps - 15 sequences
- Throwing high with right hand while standing - 14 sequences
- Walking starting with right foot, 4 steps - 16 sequences

The sequences are sampled with a frequency of 120 frames per second and 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 (Damianou et al., 2011).

All results have been extracted 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 (Bahlmann et al., 2002) is considered and the initial positions of the latent points are obtained by using the Local Linear Embedding algorithm (Roweis and Saul, 2000). Finally, classification in latent-space is performed by SVMs using the RBF kernel function. Figure 2 shows the corresponding confusion matrix obtained by using the D-SBGPLVM model.

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%

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

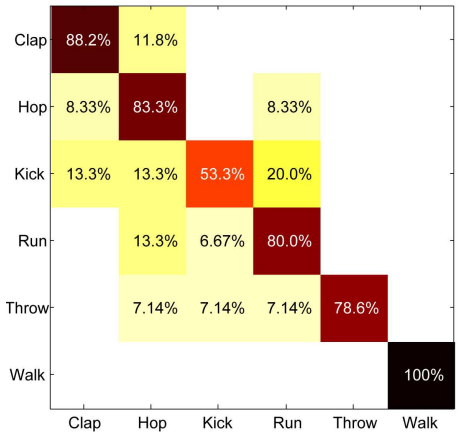

Figure 2: Confusion matrix by using D-SBGPLVM model in combination with SVM on the HDM05 "Cuts" dataset. Average accuracy: 80.9%

regarding their execution and they have been also accurately segmented manually. Finally, similar actions always start in the same way. 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.

Strangely, the classification rate of the V-GPDS model was not as high as expected considering that the analogous model which does not consider time dynamics (see (Titsias and Lawrence, 2010)), is reported to provide good classification results (e.g. on the USPS Handwritten Digits Dataset). Searching the cause of this issue, we have noticed that models for certain actions tend to provide higher conditional densities most of the time. Visually examining the latent-space representations of these models, by training them considering a three-dimensional latent-space, we have also observed that these particular sequences cover a much greater portion of the latent-space with respect to the other sequences. Further investigation is needed in this direction, as the experiments performed using V-GPDS were not sufficient to derive safe conclusions.

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, the one 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 Figures 3 and 4. In these figures,

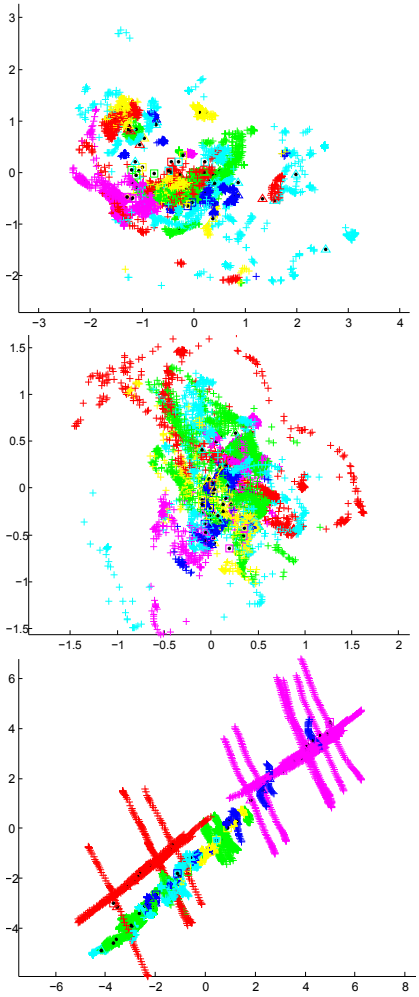

Figure 3: **Top** Latent-space representation considering Euler Angles representation and PPCA initalization, **Middle** Latent-space representation considering Unit-Quaternion representation and PPCA initalization, **Bottom** Latent-space representation considering 3D Point Cloud representation and PPCA initalization,

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, as they are computed using the back-constraints. In Figure 3 the recovered latent-spaces are shown for three different types of representations 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. We expect

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 (Müller, 2007) but without discretization based on some threshold) is examined in Figure 4. 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 Seven actions from the CMU dataset (CMU, 2003) have been also considered for evaluating the model we propose. This dataset includes the following actions:

- Walking - 15 sequences
- Running - 15 sequences
- Jumping - 15 sequences
- Sitting-Standing - 7 sequences
- Throwing-Tossing - 15 sequences
- Boxing - 9 sequences
- Dancing - 9 sequences

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

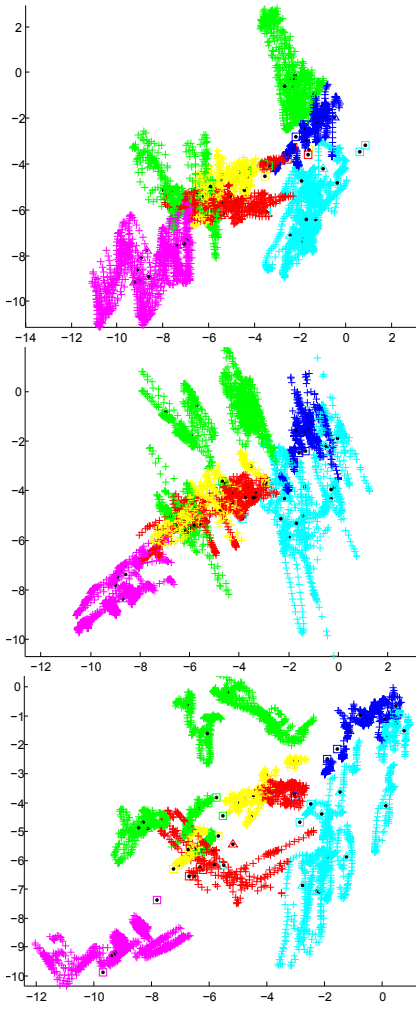

Figure 4: **Top** Latent-space representation considering Relative features representation and PPCA initalization, **Middle** Latent-space representation considering Relative features representation and LLE initalization, **Bottom** Latent-space representation considering Relative features representation and ISOMAP initalization,

provided to the D-SBGPLVM, while the rest of the setting is the same with the one described for the "Cuts" experiments. In Figure 5 we provide the corresponding confusion matrix and the overall classification rate, when the D-SBGPLVM model is used.

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 this difficulties, one can see that the proposed model gives satisfying classification results. This also demonstrates the generalization capabilities of the proposed

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%

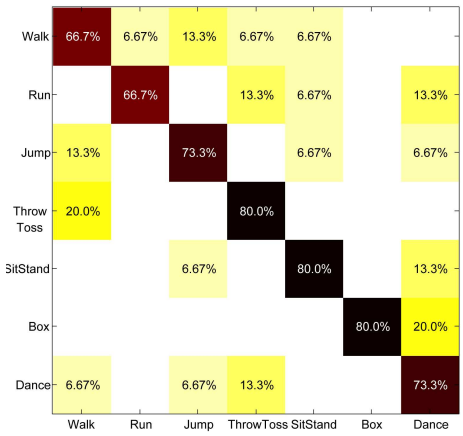

Figure 5: Confusion matrix by using D-SBGPLVM model in combination with SVM on the CMU dataset. Average accuracy: 72.9%

probabilistic model, which based on this characteristic leads to an overall accuracy that exceeds the accuracy achieved by the other two methods considered here.

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 types that 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 using a contemporary computer and a non-optimized implementation. This provides a crucial advantage with respect to other GP-LVM models which, by resorting to a new optimization to obtain the latent-space representation of the new data instances, require several minutes to complete this task. We have further shown that the proposed D-SBGPLVM model attains classification rate equivalent to the current state-of-the-art when combined with a standard classifier, as for example SVM, for classification in the latent-space.

We have focused our work on sequences originating from motion capture datasets. However, we expect to obtain satisfying results also by using sequences acquired by consumer depth cameras (e.g. Kinect (Microsoft, 2010)). There is a series of problems which should be addressed though in this case. The first regards the high level of noise of the data acquired by using this type of devices. The induced noise significantly degrades the quality of the data which in turn impairs classification in contrast to the case where highly accurate data obtained by professional 3D motion capture techniques, are used. We have also found that sampling rate plays an important role in the recognition accuracy and the fact that these devices are limited to an acquisition rate of 30 frames per second may render the classification process even more difficult. Finally, the situation is further complicated by the fact that using such devices, the acquired skeleton may not be complete as some part of the performer's body may lay outside the field of view of the device.

Within the directions of our future work, we further consider the combination of the proposed method with a 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 further allow us to perform action recognition from 2D video sequences as well.

ACKNOWLEDGEMENTS

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

REFERENCES

Aggarwal, J. K. and Cai, Q. (1999). Human motion analysis: A review. *Computer Vision and Image Understanding*, 73(3):428–440.

Bahlmann, C., Haasdonk, B., and Burkhardt, H. (2002). On-line handwriting recognition with support vector machines-a kernel approach. In *International Workshop on Frontiers in Handwriting Recognition*, pages 49–54.

CMU (2003). Carnegie-mellon mocap database. http://mocap.cs.cmu.edu/.

Cuturi, M., Vert, J.-P., Birkenes, O., and Matsui, T. (2006). A kernel for time series based on global alignments. *Compute Research Repository*.

Damianou, A. C., Titsias, M. K., and Lawrence, N. D. (2011). Variational gaussian process dynamical systems. In *Neural Information Processing Systems Conference*, pages 2510–2518.

Gong, D. and Medioni, G. (2011). Dynamic manifold warping for view invariant action recognition. In *International Conference on Computer Vision*.

Härdle, W. and Simar, W. (2003). *Applied Multivariate Statistical Analysis*. Springer Verlag.

Lawrence, N. D. (2003). Gaussian process latent variable models for visualisation of high dimensional data. In *Neural Information Processing Systems Conference*.

Lawrence, N. D. and Quiñonero Candela, J. (2006). Local distance preservation in the gp-lvm through back constraints. In *International Conference on Machine learning*, pages 513–520.

Li, Y., Fermüller, C., Aloimonos, Y., and Ji, H. (2010). Learning shift-invariant sparse representation of actions. In *International Conference on Computer Vision and Pattern Recognition*, pages 2630–2637.

Microsoft, C. (2010). Kinect. http://www.xbox.com/en-US/kinect.

mocapdata.com (2011). Eyes, japan co. ltd. http://www.mocapdata.com/.

Moeslund, T. B., Hilton, A., and Krüger, V. (2006). A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding*, 104(2-3):90–126.

Mordohai, P. and Medioni, G. G. (2010). Dimensionality estimation, manifold learning and function approximation using tensor voting. *Journal of Machine Learning Research*, 11:411–450.

Müller, M. (2007). *Information Retrieval for Music and Motion*. Springer Verlag.

Müller, M., Röder, T., and Clausen, M. (2005). Efficient content-based retrieval of motion capture data. In *SIGGRAPH*, pages 677–685.

Muller, M., Roder, T., Clausen, M., Eberhardt, B., Krüger, B., and Weber, A. (2007). Documentation mocap database hdm05. Technical Report CG-2007-2, Universität Bonn.

Ntouskos, V., Papadakis, P., and Pirri, F. (2012). A comprehensive analysis of human motion capture data for action recognition. In *Proceedings of the International Conference on Computer Vision Theory and Applications*, pages 647–652.

Poggio, T. (1985). Early vision: From computational structure to algorithms and parallel hardware. *Computer Vision, Graphics, and Image Processing*, 31(2):139–155.

Rasmussen, C. and Williams, C. (2006). *Gaussian processes for machine learning*. Adaptive computation and machine learning. MIT Press.

Roweis, S. and Saul, L. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326.

Sheikh, Y., Sheikh, M., and Shah, M. (2005). Exploring the space of a human action. *International Conference on Computer Vision*, 1:144–149.

Shimodaira, H., Noma, K., Nakai, M., and Sagayama, S. (2001). Dynamic Time-Alignment Kernel in Support Vector Machine. *Neural Information Processing Systems Conference*, 2:921–928.

Taylor, G. W., Hinton, G. E., and Roweis, S. T. (2006). Modeling human motion using binary latent variables. In *Neural Information Processing Systems Conference*, pages 1345–1352.

Tenenbaum, J. B., Silva, V. D., and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*.

Titsias, M. K. and Lawrence, N. D. (2010). Bayesian gaussian process latent variable model. *Journal of Machine Learning Research - Proceedings Track*, 9:844–851.

Turaga, P. K., Chellappa, R., Subrahmanian, V. S., and Udrea, O. (2008). Machine recognition of human activities: A survey. *IEEE Transasctions on Circuits and Systems for Video Technology*, 18(11):1473–1488.

Urtasun, R. and Darrell, T. (2007). Discriminative gaussian process latent variable model for classification. In *International Conference on Machine Learning*, pages 927–934.

Urtasun, R., Fleet, D. J., and Fua, P. (2006). 3d people tracking with gaussian process dynamical models. In *International Conference on Computer Vision and Pattern Recognition*, pages 238–245.

Urtasun, R., Fleet, D. J., Geiger, A., Popovic, J., Darrell, T., and Lawrence, N. D. (2008). Topologically-constrained latent variable models. In *International Conference on Machine Learning*, pages 1080–1087.

Waltisberg, D., Yao, A., Gall, J., and Van Gool, L. (2010). Variations of a hough-voting action recognition system. In *International conference on Pattern Recognition*, pages 306–312.

Wang, J. M., Fleet, D. J., and Hertzmann, A. (2006). Gaussian process dynamical models. In *Neural Information Processing Systems Conference*, volume 18, pages 1441–1448.

Yao, A., Gall, J., Fanelli, G., and Gool, L. V. (2011). Does human action recognition benefit from pose estimation? In *British Machine Vision Conference*, pages 67.1–67.11.

Yao, A., Gall, J., and Gool, L. J. V. (2010). A hough transform-based voting framework for action recognition. In *International Conference on Computer Vision and Pattern Recognition*, pages 2061–2068.

Zhang, X. and Fan, G. (2011). Joint gait-pose manifold for video-based human motion estimation. In *European Conference on Computer Vision*, pages 47–54.