Generalised Laplacian Eigenmaps for Modelling and Tracking Human Motions

Jesus Martinez-del-Rincon, *Member, IEEE,* Michal Lewandowski, *Member, IEEE,* Jean-Christophe Nebel, *Senior Member, IEEE,* and Dimitrios Makris, *Member, IEEE*

***Abstract*—This paper presents Generalised Laplacian Eigenmaps, a novel dimensionality reduction approach designed to address stylistic variations in time series. It generates compact and coherent continuous spaces whose geometry is data-driven. This work also introduces Graph-based Particle Filter, a novel methodology conceived for efficient tracking in low dimensional space derived from a spectral dimensionality reduction method. Its strengths are a propagation scheme which facilitates the prediction in time and style, and a noise model coherent with the manifold, which prevents divergence, and increases robustness. Experiments show that a combination of both techniques achieves state-of-the-art performance for human pose tracking in underconstrained scenarios.**

***Index Terms*—Human articulated tracking, Human motion modelling, Dimensionality reduction, Particle filtering.**

I. Introduction

A variety of applications in computer vision, such as visual surveillance, gesture analysis, human-computer interfaces and computer animation, requires the interpretation of human poses and their dynamics. Due to the complexity of human motion, computer vision systems usually rely on learning from human motion time sequences [1]–[10]. Despite the fact human pose recovery is a very active research field, it still remains a major challenge. Since there are so many ways of performing even the simplest activity, such variability affects significantly performance of applications, especially when the activity is performed by an individual who is not present in the training set [11]. In this work, we propose a novel method to represent ‘styles’ in time sequences and its integration in a novel tracking framework for human motion analysis.

Here, we use the term ‘style’ to express a variation of a given activity or movement that does not affect its intrinsic nature, which means that a styled instance of an activity is still recognisable as belonging to the same activity class. Variability is caused not only by morphological and biomechanical differences between people, but also by multiple factors affecting an individual’s behaviour such as mood, clothing, speed of movement and environment. The combination of all these factors is expressed in a continuous space or ‘style’ space.

This approach contrasts with simplified models where style is decomposed into a small set of ad-hoc discrete states or labels, such as identity or gait, which identify univocally a subject and/or an activity [3]–[8], [12], [13]. In such framework, only interpolation between those user-defined styles can be contemplated which limits its practical usage. Nevertheless, even under these limiting conditions, the generation of realistic motion synthesis by combining linearly discrete styles [13] illustrates the continuous nature of style.

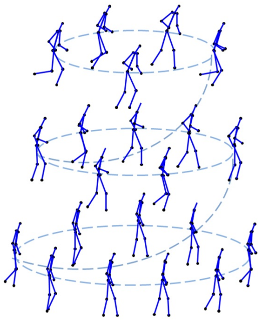


Figure 1. Stylistic variations for the gait cycle considering different subjects and different speeds

Those ’style’ spaces are embedded in the multi dimensional space of possible human poses which is high-dimensional in its traditional limb based parameterisation [9], [10], [14]. However, it has been shown that the pose space for a given activity has a significantly smaller intrinsic dimensionality. For instance walking can be embedded into a two dimensional space [15]–[18]. Since most methodologies suffer from the high variability of human poses, dimensionality reduction (DR) is an intuitive approach to generate a low dimensional representation which facilitates motion analysis. Unfortunately, stylistic variability, which is intrinsic to human locomotion, is usually lost during that process. As a consequence, this may result in a loss of specificity by compressing important information such as intra-activity variance and inter-subject variability.

Style-preserving modelling of human activities will lead to more robust and successful human tracking systems. In this work, we define the robustness of a tracking system as the capacity of not diverging and maintaining its performance when observation degrades, i.e. when ambiguity increases. In tracking applications, the high dimensionality of data impacts negatively on their performance [19]. This led to the development of tracking approaches that address the size of the solution space, either using efficient search strategies such as annealing [20] and space partition [21] or by reducing its dimensionality [1], [2], [6], [22], [23]. However, their modelling limitations in terms of preserving stylistic variations impact on tracking results in a loss of accuracy and robustness when dealing with different subjects and/or activities.

J. Martinez is with ECIT, Queen’s University of Belfast, e-mail: j.martinez-del-rincon@qub.ac.uk.

M. Lewandowski, J.C Nebel and D. Makris are with Kingston University.

Manuscript received -, 2013; revised -, 2014.

This paper focuses on the preservation of stylistic variations of activities while reducing the dimensionality of human motion feature space and on the usage of style modelling for improving articulated tracking frameworks. As a first contribution, we introduce a novel technique, Generalised Laplacian Eigenmaps (GLE), which is able to produce, in an unsupervised manner, continuous low-dimensional space activity manifolds whose geometries are data-driven. By introducing explicitly style as a discriminative constraint in the generation of embedded spaces, the new methodology is able to model variability due to stylistic variations caused both by a person's identity and the type of activity. In such continuous spaces, the boundaries between similar activities, such as walking and running, are fuzzy. Therefore, it is possible to consider that a specific locomotion is a stylistic extension of another [13] (see Figure 1), even though they are biomechanically different.

Our second contribution is a novel and integrated articulated tracking scheme, Graph-based Particle Filter (GbPF), which integrates the formation process of graph-based DR methods into the tracking paradigm. As a result, this new scheme is able to address natural limitations of model priors for articulated motion tracking: it improves search efficiency, reduces risks of divergence and increases the likelihood of recovering after failure.

A. State of the art

1) Human Pose Modelling: In the context of human pose modelling, a low dimensional representation not only has to provide a compact and functional space, but also must be sufficiently general to capture human pose variations. The inability of linear methodologies like PCA to deal simultaneously with both requirements, [24], has led to the development of many non linear DR techniques which can be classified in two main categories: embedding-based and mapping-based approaches.

Embedded-based approaches such as Laplacian Eigenmaps (LE) [25], Isomap [26] and Local Linear Embedding (LLE) [27], estimate the structure of the underlying manifold by preserving geometrical properties of the data structure. However, since they do not provide any mapping between low and high dimensional spaces, this usually needs to be estimated in a second optimisation step [28].

Mapping-based approaches, such as kernel PCA [29] or Gaussian process latent variable model (GPLVM) [30], use nonlinear functions to map the embedded space to the data space and vice versa. They are optimised in conjunction with the latent variables in a single optimisation process, aiming at better results. However, such an approach increases the computational complexity from of $O(N^2)$ for embedded methods [31] to $O(N^3)$ [32]. As a consequence, their usage when dealing with large datasets is problematic. In addition, high complexity leads to problems of convergence [15], [30].

The exploitation of non-linear DR techniques for tracking in a lower-dimensional space requires preservation of locality and temporality in the low dimensional space: nearby points in time and in high dimensional space must be mapped to nearby points in low dimensional space. If this property is not preserved, a smooth trajectory in the high dimensional space will map to a discontinuous trajectory in the low dimensional space. Exploiting such manifold would require artificially high values in a noise model and/or empirical dynamic models to deal with discontinuities, which leads to inconsistent pose tracking [1], [17]. Several techniques, such as ST-Isomap [33], back constraint GPLVM (BC-GPLVM) [34], Gaussian process dynamical model (GPDM) [17] and Temporal Extension of Laplacian Eigenmaps (TLE) [15], attempt to address this issue by introducing temporal constrains to ensure smooth transition in the latent space. Despite their success in modelling a given activity, all these methods fail to represent stylistic variations, such as different people performing the same activity or the same person performing different variations of an activity. This is due to two different factors. First, many non-linear dimensionality methods are not able to generate a consistent manifold when different stylistic variations are present in a training set [25]–[27], [30], [34], [35]. Second, some methods are style independent, i.e. style information is discarded on purpose to generate subject independent models [15], [16], [33].

A few approaches have been proposed to express style. Some previous works focus on learning multi style models over conventional low dimensional spaces [3], [7], [8]. Due to the limitations of those DR techniques, style is lost during the process and these methodologies need to mitigate the elimination of stylistic variations by reintroducing style in a second learning step. Other approaches attempt to quantise styles within the learning process of the low dimensional space. Thus, Elgammal et al. [16] suggested a generative model that explicitly decomposes the intrinsic body configuration as a function of content and style. However, the complexity of this method increases exponentially with the number of considered styles. Pan et al. [4] proposed a more elegant solution based on a hierarchical methodology that learns a latent distribution for each given style. Finally, Wang et al. [13] presented a multi factor Gaussian process model that parameterises the space of human motion styles by a small number of low-dimensional factors, i.e. gait and identity. All these supervised [3], [5], [7], [8], [12], [16], [36] and semi-supervised [4], [13] methodologies approximate the pose/style space by a set of known discrete states related to people and/or activity labels. However, they fail to model style variations due to other factors such as environment, speed, clothes and mood since factors that define a particular style are often poorly defined, certainly non-discrete and hardly quantifiable. Only unsupervised and data-driven methodologies offer a framework which may be able to include the whole range of style variability into a model.

To date, very few unsupervised approaches are able to deal with style. Urtasun et al. [19] proposed a locally-linear GPLVM (LLGPLVM) that requires some prior knowledge about the activity of interest to constrain the optimisation process. By imposing a tubular manifold geometry, an extra dimension is allocated to the representation of stylistic variations. An important issue regarding this ad-hoc constrain is that, the artificially imposed geometry is not always representative of the actual geometry of the data as explained later in section IV-B3.

2) Human Pose Tracking: Due to the high complexity of articulated human tracking, tracking paradigms that aim at finding efficiently a solution in the high dimensional space possess a significant advantage. Partitioned sampling [21] proposes division of the search space into several partitions and sequential application of dynamics for each of them followed by some weighted resampling. However, annealed particle filter (APF) [20] is better suited for pose tracking since its layered and hierarchical methodology takes into account the hierarchy of articulation. By refining gradually the resolution of the particle filter (PF) fitness function, the complexity of the search is reduced drastically. While these techniques perform satisfactorily in a calibrated multi-camera environment, they are not suitable in monocular or uncalibrated scenarios [37] due to the inherent ambiguity and the underconstrained nature of the problem.

In these circumstances, human pose models have the potential to constrain the solution space, which has led to the inclusion of priors specifically for tracking. First attempts were based on the simplistic assumption that human motion is smooth, which can be modelled as a low-order Markov model [38], [39]. In response to the non smoothness of the human motion, early activity specific models [18], [40], [41] proposed Gaussian mixture as a mean to capture and model human pose priors. However, despite some success, their models limit the ability to describe the complexity of the human motion space. More recently, in order to address the inherent problems associated with Gaussian approaches, more advanced prior learning methodologies were integrated within tracking paradigms [1], [3], [7], [42]. Unfortunately, almost none of the DR techniques which aims at preserving style has been quantitatively validated within a pose tracking framework [4], [13], [16], [19], [43]. To our knowledge, only dynamical binary latent variable model [35] handles style within a tracking application. However, although it displayed some generalisation properties across either subjects or activities, no successful experiment has been reported combining both style variation sources at the same time.

A common characteristic of all these prior models is the limited exploitation of the multi-hypothesis capabilities of particle filter to perform an efficient search in low dimensional spaces. Usually, hypotheses are distributed in the low-dimensional space according to a generally unknown low-order dynamic model associated to a Gaussian noise. Such approach has two main drawbacks. First, since dynamic models are not constrained by the manifold geometry, hypotheses can move in the whole space which may lead to tracking divergence. Second, a simple noise model does not characterise the actual uncertainty inherent in the manifold. In addition to poor tracking performance, it may contribute to further divergence. [37] proposed tracking on the surface of an ideal toroidal manifold to address the first problem. The known geometry of the manifold simplifies the estimation of a dynamic model and the propagation on the surface of the torus. However, although this prevents divergence, usage of a generic noise model does not tackle the second problem and reduces tracking performance. Moreover, the generation of the torus manifold requires previous knowledge about the geometry of the motion. Finally, inclusion of several people or styles may require the design of a very different geometry.

II. Generalised Laplacian Eigenmaps

Generalised Laplacian Eigenmaps is a DR method that combines temporal and stylistic information as integral part of its objective function. This is achieved by introducing two types of complementary constraints into the Laplacian Eigenmaps framework.

Our approach requires both the generation of a low dimensional space and the mapping functions to map that space. Mapping methods, where DR and mapping are optimised simultaneously, suffer from high complexity. This leads to lengthy training time and problems of convergence [15], [30] which, in practice, make them unsuitable to deal with large datasets. Since a large amount of data is required in order to cover different inter and intra-subject styles, only an embedded-based method equipped with mapping functions is suitable.

Among these methodologies, Laplacian Eigenmaps is the most appropriate, since it provides a mathematical framework where new constraints can easily be introduced [9], [34], [44]. The insertion of these new constraints, modelled as connectivity graphs, allows extending the preservation of certain properties in the low dimensional space. Thus, not only locality, as intended in the classical LE, but also other interesting properties such as continuity and temporal sequentiality are preserved. Although Isomap shares some of these properties [33], [70], an LE extension (TLE) has already demonstrated better performance when dealing with time series [15].

A. Graph construction

A set of P time series, $Y = \{Y^1, ..., Y^P\} = \{y_k^p\}, \forall k \in [1, Kp], \forall p \in [1, P]$, is defined as a set of P sequences of data points with variable length K_p that occur sequentially in time, where y_k^p is a data point distributed in a high dimensional space ($y_k^p \in \mathbb{R}^N$). Given Y, LE is able to discover its low dimensional representation, $Z = \{Z^1, ...Z^P\} = m_k^p$ with ($m_k^p \in \mathbb{R}^n$), where $n << N$, which preserves the local structure of the original data by solving the Eigen value decomposition problem [25]:

$$L \cdot Z = \lambda \cdot D \cdot Z \tag{1}$$

where L is the Laplacian matrix and D is the corresponding diagonal matrix with entries $D_{ii} = \sum_{j=1}^{M} G(i, j)$. G is a graph whose connectivity controls directly the similarity in the embedded space [25].

In order to preserve simultaneously the temporal structure and the style variance of the original data, both constraints are expressed explicitly by building neighbourhood graphs between the training samples. In this manner, local style neighbours as well as local temporal neighbours are placed nearby in the LE embedded space without the necessity of enforcing any artificial embedded geometry as in [19]. Similarly to [15], two types of neighbourhoods are automatically defined in GLE $\forall y_k^p \in Y^p$:

- Temporal neighbourhood T_k: it ensures temporal continuity on the manifold. The 2τ closest points are defined as the τ-previous and the τ-next points in the time series Y^p.

$$\begin{aligned} T_k &\in \{y^p_{k^-}, ..., y^p_k, ..., y^p_{k^+}\} \\ k^- &= \max(1, k-\tau), \\ k^+ &= \min(k+\tau, K_p) \end{aligned} \tag{2}$$

- Stylistic neighbourhood S_k: based on local geometry, it ensures stylistic continuity between training instances which are close in style by establishing correspondences between repetitions of a given instance in the training set. First, a temporal neighbourhood T_k is defined around the point y^p_k and used as reference. Then, a time series similarity measure is applied between the reference neighbourhood and all the time series in the training set by means of a sliding window.

The score of each comparison is stored in a similarity vector, where an average s-connected filter is applied for non-minima suppression. All fragments $R^h_k, h \in [1, r_k]$, with similarity greater than b standard deviations from the average of the similarity vector are considered as the r_k-th repetition fragments of the temporal neighbourhood T_k. Finally, stylistic neighbours $R^h_k(l)$ are selected as the closest points to y^p_k inside each repetition fragment R^h_k, so the stylistic neighbourhood is defined as follows:

$$S_k \in \{R^1_k(l^1_*), ..., y^p_k, ..., R^{r_k}_k(l^{r_k}_*)\} \tag{3}$$

where

$$l^h_* = \arg\min_{l \in R^h_k} \|y^p_k - R^h_k(l)\| \tag{4}$$

Different similarity measures, such as Edit Distance with Real Penalty, Longest Common Subsequence and Edit Distance on Real Sequence [45], can be applied to detect and align repetitions. In our framework, Dynamic Time Warping [46] has been chosen since, in addition to its simplicity and effectiveness [47], it does not require the two series fragments to have the same sampling frequency. This property is essential when dealing with multi-style time series, where time length, sampling and periodicity are some of the features affected by style variations.

Both neighbourhoods may be understood as constraints (Eq. 7) and modelled as connectivity graphs using the LE formalism (Eq. 5 and Eq. 6).

$$G_T(i,j) = \begin{cases} e^{-\|y_i - y_j\|^2} & i,j \in T_k \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

$$G_S(i,j) = \begin{cases} e^{-\|y_i - y_j\|^2} & i,j \in S_k \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

$$L_T = D_T - G_T \quad L_S = D_S - G_S \tag{7}$$

A manifold which includes temporal-stylistic coherence in its structure is generated by introducing these constraints with an appropriate balance. We propose to ponderate the balance between temporal and stylistic variabilities by introducing a weighting factor β. Since frame rate may be assumed to be fixed for a given set of time series, this factor applied to the stylistic graph increases its importance taken the temporal variance as reference. Low values of β discard the stylistic variations in benefit of the temporal continuity, whereas high values discard temporal information by considering mainly style.

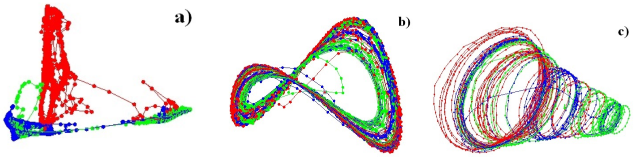


Figure 2. 3D Manifolds created with MoCap data from 3 individuals (red, green and blue) performing each 3 variations of an activity, i.e. walking, fast walking and running. a) LE. b) Temporal LE. c) GLE (automatic, $\beta = 3.05$).

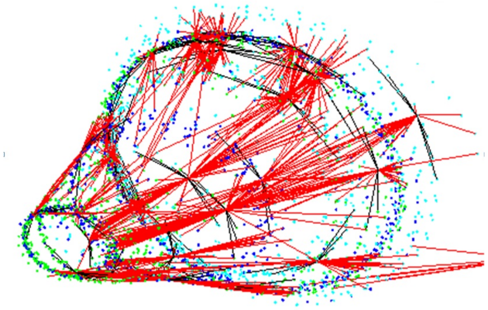


Figure 3. GLE manifold created with Mocap data from 3 individuals (cyan, green and blue) performing each 3 variations of an activity, i.e. walking, fast walking and running. The learned G_T temporal (black) and G_s stylistic connectivity (red) are displayed for a subset of randomly selected points

Once the graph balance has been calculated, graphs are combined as a weighted addition of their Laplacian matrices. The embedded space Z of dimension n is spanned by the eigenvectors given by the n smallest nonzero eigenvalues λ. They are obtained from the solution of the generalised eigenvalue problem [25], which is deduced by minimising the objective function:

$$\arg\min Z^T \cdot (L_T + \beta \cdot L_S) \cdot Z \tag{8}$$

subject to $Z^T \cdot (D_T + \beta \cdot D_S) \cdot Z = I$ where I is the identity matrix.

Under this formulation, LE, could be seen as a special case of GLE where $\beta = \infty$. Similarly, TLE [15] can also be considered as a particular case of GLE where $\beta = 1$. The visual comparison between different LE-based methods highlights the influence of the temporal and stylistic elements, see Figure 2. The internal structures of the GLE manifold and the connectivity given by the temporal and stylistic neighbours, as expressed by the graphs G_T and G_S respectively, are illustrated in Figure 3 where walking, fast walking and running activities are modelled within a single low dimensional space.

B. Automatic estimation of graph balance

The balance weightβ controls the importance given to each graph, i.e. temporal and stylistic, during optimisation. Consequently, the choice of this parameter is essential. Figure 4 illustrates the effect of that choice. We suggest an intuitive formula to provide automatically the appropriate balance between temporal and stylistic information. It is based on the normalisation of data variations along the time and style dimensions. By selecting and normalising both Laplacian matrices using the highest eigenvalues, the balance weight is

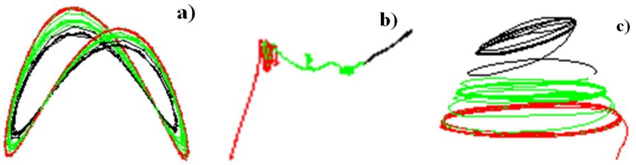


Figure 4. GLE embedded spaces using 1 person and 3 styles of motion: walking (red), walking fast (green) and running (black). Manifolds were generated using: a) $\beta = 1$, b) $\beta = 10$ and c) $\beta = 3$.

calculated using the following equation:

$$\beta = \frac{\max \lambda_{L_T}}{\max \lambda_{L_S}} \tag{9}$$

where the eigenvalues are calculated by performing the eigen-decomposition for each graph individually (Eq. 1).

C. Mapping functions

Spectral methods such as LE allow unsupervised generation of embedded spaces, but they do not provide explicitly any mapping mechanism between the low and high dimensional spaces. This issue has been tackled very effectively by Radial Basis Function Networks (RBFN) [15], [16], [50]. Projection functions are produced by training direct ϕ and inverse ϕ' sets of functions between high and low dimensional spaces.

$$\phi : \mathbb{R}^N \rightarrow \mathbb{R}^n \text{ and } \phi' : \mathbb{R}^n \rightarrow \mathbb{R}^N \tag{10}$$

In this paper, we modify the standard RBFN learning process in order to deal with two of its weaknesses. First, the type of mapping activation functions should be selected to fit the manifold geometry. However, since this information is usually not known, assumptions regarding its geometry have to be made. Second, the number of functions that compose each set is a parameter which needs to be set by the user. Incorrect estimation of these produces poor performance of the mapping components by under or over fitting. Given that variance on style is uncorrelated with temporal variance, standard spherical functions based on Euclidean distance are not able to model the space adequately. Consequently, multi-dimensional Gaussian activation functions ϕ_j (Eq. 10) are more suitable since they can assign different variance values to each dimension (see Figure 5); as demonstrated in the experimental section.

$$\phi_j = e^{(-(X-\mu_j)^T \cdot \Sigma_j^{-1} \cdot (X-\mu_j))} \tag{11}$$

for $j = 1, ..., n_g$, where X is the input feature vector, n_g the number of Gaussian functions to be discovered and μ_j and Σ_j the mean and covariance respectively of each Gaussian function. Under this definition, standard spherical functions are seen as a simplification of Eq. 11, where Σ_j is a scalar multiple of the identity matrix.

Since more suitable characterisation of the activation functions provides higher flexibility about the number of coefficients n_g to be learned, the choice of this parameter is less critical. In our framework, it is estimated automatically by applying the Figueiredo-Jain Gaussian Mixture Model (GMM) parameters automatic estimation (FJ) [51]. In addition, we apply Expectation Maximisation (EM) instead of k-means [15], [50] to determine the mapping components since it has the ability to derive elliptical clusters, instead of spherical ones.

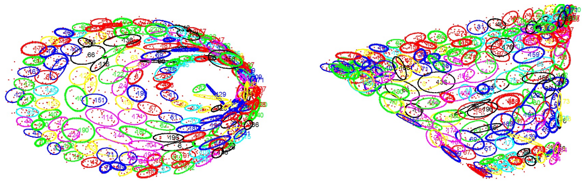


Figure 5. RBFN components, learned for $n_g = 200$, projected on the 3D mocap manifold (perpendicular views).

III. Graph-based Particle Filter

In this section, we introduce our probabilistic tracking method based on particle filter that integrates manifold priors for robust and multi-style pose estimation. The priors are learned by applying a DR method, but in contrast to previous approaches, our technique uses these priors as integral part of the prediction. Thus, this prior embedding supports our particle filter in two ways. First, it provides a propagation model that contemplates simultaneously prediction in style and time. Second, it automatically produces a suitable data-driven noise model in the manifold which simplifies the filter configuration. This prevents divergence towards invalid poses in the low dimensional space by ensuring motion in the vicinity of the manifold.

In this work, we focus on learning the priors using GLE. This provides a propagation model with both temporal and stylistic constraints, where temporal constraints are also employed to provide a style-specific dynamic model. The embedding of the prior is consistent with the nature of the GLE spectral method since it relies on graph information derived during the learning of the manifold. However, and in order to demonstrate the applicability of our tracking methodology to any graph-based DR, e.g. Isomap and LE, experiments are also performed using other DR techniques.

A. Methodology

Using the estimated manifold, the initial pose is set on its surface. Then, particles must be distributed and propagated. Traditionally, this is achieved by applying a low-order dynamic model and a Gaussian noise around that prediction [1], [6]. In such scheme, tracking performance relies directly on the characterisation of the noise function. Since there is no hard constraint associated to the manifold, the estimated distribution of particles could diverge outside the training space and produce unrealistic hypotheses, as Figure 6a illustrates. Although models based on inverse kinematics [52] may provide more accurate hypotheses, they cannot prevent divergence under poor observations.

We propose to update dynamically the process noise using information provided by the GLE prior model. Specifically, a customised noise estimate is obtained for each point of the continuous low dimensional space by representing the RBFN functions as a GMM $\{\pi_j, \mu_j, \Sigma_j\}, j = 1...n_g$. Since these particular functions (Eq.10) are more suitable than traditional spherical functions for modelling and mapping the manifold, they will not only lead to lower reconstruction error when projecting our hypotheses from the latent space to the 3D skeleton space, but also provide accurate modelling of the area

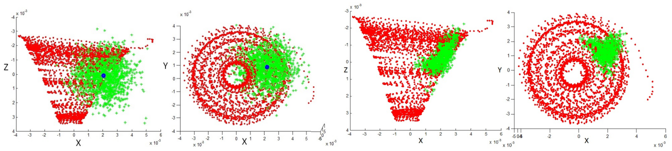


Figure 6. Particle distribution on the manifold by: a) Gaussian noise addition modelled by Eq. 11, b) Graph-based propagation defined by Eq. 13, for the same noise variance

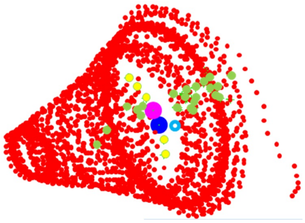


Figure 7. Graph-based triple resampling process. The particle x_{t-1}^i (cyan) is resampled and projected to the manifold with a probability given by Eq. 12. The resulting point from the second resampling m_k (dark blue) is projected forward in time (magenta) to one of the temporal neighbours (yellow). Finally, the third resampling chooses the final hypothesis x_t^i among the stylistic neighbours (green).

around the manifold to produce valid hypotheses as defined by the training set. The noise covariance, at any point x in the low dimensional space, is modelled as the covariance of a subset N_{sg} of the Gaussians set N_g belonging to the GMM which correspond to that point according to the Mahalanobis distance and a certain threshold ϱ.

$$\Sigma_{Noise}(x) = \sum_{i \in N_{sg(x)}} \Sigma_i \tag{12}$$

$$\begin{gathered} N_{sg(x)} = \{\forall j, (x-\mu_j)^T \cdot \sigma_j^{-1} \cdot (x-\mu_j) < \varrho\} \\ N_{sg}(x) \subseteq N_g \end{gathered} \tag{13}$$

Although, Eq. 11 provides a data-driven noise model that is coherent with the manifold and mapping function, it does not ensure that the system will not diverge when only poor observations are extracted from a few consecutive frames. An improved strategy against divergence is achieved by integrating in the particle propagation those constraints provided by the neighbourhood graphs. The differences in terms of noise modelling between the standard particle filter and the proposed propagation schemes are illustrated in Figure 6. In LE-based methodologies, including GLE, connectivity graphs regulate the proximity and locality of the poses on the manifold. Therefore, this connectivity information is very valuable to propagate and predict plausible hypotheses. This is exploited by replacing the traditional deterministic propagation and prediction steps of particle filters by a stochastic propagation based on a triple resampling process, depicted in Figure 7.

First, particles are resampled to propagate valid hypotheses according to their observation weight in the previous time step (Alg.1, 1a-1c), as in traditional PF.

In the second resampling stage, temporal prediction is based on the temporal graph G_T, which replaces the function of conventional dynamic models. It allows tuning the dynamic model and deals with differences of temporal resolution between the test and training data. In this resampling, each particle x_t^i is associated to training points in the manifold m_k^p with a probability proportional to their Euclidean distance d_E.

$$p(m_k^p|x_t^i) \propto e^{(-d_E(x_t^i,m_k^p)/2\sigma^2)} \tag{14}$$

Only one manifold point is randomly selected for each particle. Its corresponding χ-th temporal neighbour $m_{k+\chi}^p \in T_k$ is then used as temporal prediction (Alg.1, 1d-1h) where $\chi \in \mathbb{N}$. χ is a parameter that enables the temporal prediction, where the standard value $\chi = 1$ implies moving to the next neighbour and therefore predicting actually the next pose in time. $\chi = 0$ is equivalent to applying a zero order model. Higher values of χ introduce a shift allowing dealing with test data generated at a lower frame rate than training data's. Therefore, training may be performed using the highest available temporal resolution and the resulted manifold may be applied to application data with lower frame rates, as long as the value χ is tuned appropriately according to the ratio between the two frame rates.

In the third resampling stage, particles are projected in the style dimension based on the stylistic graph G_S. Again, resampling is repeated for each particle and only one sample per particle is selected. All the stylistic neighbours $m_{k'}^{p'} \in S_{k+\chi}$ associated to the temporal prediction $m_{k+\chi}^p$ of the resulting particle x_t^i from the previous stage are taken into account. Their weights are given by their values into the stylistic graph G_S (Alg.1, 1i-1l). Finally, Gaussian noise $p(x_t^i|m_{k'}^{p'}) \sim N(0, \Sigma_{Noise}(m_{k'}^{p'}))$, as estimated by Eq. 12, is added to the final set of particles in order to allow some degree of flexibility around the training manifold.

This triple resampling strategy (see Algorithm 1) provides a stochastic propagation and prediction scheme, coherent with the probabilistic PF framework, which allows moving on the manifold surface. Conceptually, given a previous position of a particle x_{t-1}^i, the prediction x_{t-1}^i is:

$$\begin{aligned} p(x_t^i|x_{t-1}^i) \propto p(x_t^i|m_{k'}^{p'}) \cdot G_s(m_{k+\chi}^p, m_{k'}^{p'}) \\ \cdot G_T(m_k^p, m_{k+\chi}^p) \cdot p(m_k^p|x_(t-1)^i) \end{aligned} \tag{15}$$

where the first term $p(x_t^i|m_{k'}^{p'})$ adds the noise model and the three following ones $p(m_k^p|x_t^i)$, G_T and G_S correspond to the triple resampling process. Although the second and the third resampling could be merged by using $G = G_T + \beta G_s$, considering them individually allows a higher flexibility on the choice of different dynamics.

This methodology is designed to reduce the probability of diverging, increase robustness, improve recovery after divergence and facilitate the prediction in time and style by using the implicit information stored in the connectivity graphs. Although the complexity of the propagation algorithm increases due to its probabilistic procedure, the added computation time to the whole particle filter framework is almost negligible. This is due to the fact that the most expensive part of the algorithm is the evaluation of the likelihood function, whose processing time only depends on the number of hypotheses, which is not affected by the additional resampling process.

IV. Experimental Validation

The proposed modelling technique GLE and the Graph based tracking framework (GbPF) are validated using different types of datasets, described in section IV-A. Section IV-B

Algorithm 1 Particle filter with GLE priors and graph-based propagation

```
Given a set of particles $\{x_{t-1}^i, \omega_{t-1}^i\}_{i=1}^N$ which represents the posterior probability of $p(x_(t-1)|z_(t-1))$ at time $t-1$, and a prior manifold $\{\{m_k^p\}_{k=1}^{K_p}\}_{p=1}^P$
Select $N$ samples from the set $x_{t-1}^i$ with probability $\omega_{t-1}^i$:
Calculate the normalised cumulative probability $cx_{t-1}^n = \frac{\sum_{i=1}^n \omega_{t-1}^i}{\sum_{i=1}^N \omega_{t-1}^i}$
Generate a uniformly distributed random number $r \in [0,1]$ and find the smallest $j$ for which $cx_{t-1}^j \geq r$
Set ${x'}_{t-1}^i = x_{t-1}^j$
Generate $M$ samples $\hat{x}_{t-1}^k$ associated to manifold points $m_k^p$ with a probability $\pi_{t-1}^k \propto e^{-\|{x'}_{t-1}^i - m_k^p\|/2\sigma^2}$ where $\sigma$ is a normalisation factor
Calculate the normalised cumulative probability $c\pi_{t-1}^n = \frac{\sum_{k=1}^n \pi_{t-1}^k}{\sum_{k=1}^M \pi_{t-1}^k}$
Generate a uniformly distributed random number $r' \in [0,1]$ and find the smallest $j$ for which $c\pi_{t-1}^j \geq r'$
Set ${x''}_{t-1}^i = \hat{x}_{t-1}^j$
Propagate ${x''}_{t-1}^i$ to the next time step ${x''}_t^i$ according to the $\chi$-th temporal neighbour in the manifold given by $G_T({x''}_{t-1}^i, {x''}_t^i)$
Generate $B \leq M$ samples $\tilde{x}_t^k$ associated to the manifold points $m_{k'}^{p'}$ with a probability $\rho_t^k = G_s({x''}_t^i, m_{k'}^{p'})$
Calculate the normalised cumulative probability $c\rho_t^n = \frac{\sum_{k=1}^n \rho_t^k}{\sum_{k=1}^B \rho_t^k}$
Generate a uniformly distributed random number $r'' \in [0,1]$ and find the smallest $j$ for which $c\rho_t^j \geq r''$
Set ${x'''}_t^i = \tilde{x}_t^j$
Add noise, $x_t^i = {x'''}_t^i + w_t^i$ where $w_t^i\ N(0, \Sigma_{Noise}({x'''}_t^i))$
Evaluate likelihood function $\omega_t^i \sim f(x_t^i, \phi, I_t)$ over the input image $I_t$
Estimate the mean state of the set $x_t^i$, $E[x_t] = \sum_{i=1}^N \omega_t^i \cdot x_t^i$, and its high dimensional representation $E[y_t] = \phi(E[x_t])$
```

focuses on GLE validation. First, the automatic parameter selection and the mapping methodology are empirically justified. Then, the manifold geometry of the embedded space is discussed and GLE is applied to model learning. Finally, our approach is evaluated quantitatively against relevant methodologies for two computer vision applications, i.e. pose estimation and image segmentation. In section 4.3, Graph-based particle filter is evaluated within an articulated tracking framework relying on GLE and other DR-based priors

A. Experimental setup

Although GLE is a general purpose DR technique, it is particularly suitable for the analysis of human motion data: they form time series whose stylistic variations should be well represented in GLE's continuous space. Since two embedded dimensions are sufficient to represent the temporality of many human activities [15], [17], we propose to add a third dimension so that style variability can be expressed in our model. Parameter b was set to 1.5 in all our experiments.

In order to illustrate the independence of our methodology regarding data, the validation process is conducted using two types of motion sequences, i.e. MoCap and video datasets. Since existing MoCap databases (e.g. [53]) do not contain many stylistic variations for a given activity, we introduce a new MoCap dataset, called "walking2running" (see Figure 8).

Figure 8. From left to right: Walking2running dataset, CMU Mobo, INRIA Ixmas and HumanEva II

It was recorded using an optical MoCap system "Qualysis Track Manager", with a frequency of 120Hz. In each sequence, the subject performed three varieties of "bipedal locomotion": slow walking (2 miles/hour), fast walking (4 miles/hour) and running (6 miles per hour). Transitions between these three locomotion modes were also captured. These actions were performed on a treadmill to allow speed control and ensure consistency between subjects. In our experiments 6 subjects (5 males and 1 female in the age range 24-41) produced 6 sequences of 4800-9000 frames each. 3D skeleton data are represented by quaternions of 13 joint angles.

Image based validations were conducted using three standard databases. The first one, Mobo Dataset [54], provides 6 synchronised video sequences of 9 people walking (Figure 8). Images of the background are also provided. As previously, the data were captured in a controlled environment: a treadmill allowed recording people walking at 2 and 2.8 miles/hour 15% of the dataset was manually annotated to facilitate quantitative evaluation. The second dataset, IXMAS [55], provides 13 day-live actions, such as kicking, sitting and crossing arms, performed 3 times by 11 subjects and observed by 5 synchronised cameras. The provided 2D foreground masks are used as input data in our experiments. Finally, the third dataset, HumanEVA II [64], has been chosen since it has been used widely in the community to evaluate video-based motion capture [64]. It offers a framework which allows comparison of pose estimates to ground truth that was derived by a marker-based Motion Capture system. This dataset is composed of 4 synchronised views of two subjects performing various locomotion activities, i.e. walking, running and balancing.

B. GLE validation

1) Beta weight validation: Eq. 9 is validated using both walking2running MoCap and MoBo video datasets. Figure 9(top) shows, for a range of β values, how the balance between temporal and stylistic variance affects the manifold representation of an activity. In order to quantify the relationship between stylistic and temporal variance, we propose a resolution based metric which calculates the ratio between the stylistic and temporal resolutions displayed in the neighbourhood graphs associated with the low dimensional space. More specifically, we define the Resolution Ratio as the ratio between the average style distance and the average temporal distance between graph neighbours.

$$Res.Ratio = \frac{\sum_{k=1}^K \frac{1}{r_k} \sum_{h=1}^{r_k} \|m_k - R_k^h(l_*^h)\|}{\sum_{k=1}^K \frac{1}{k^+ - k^-} \sum_{t \in T_k, t \neq k} \|m_k - m_t\|} \qquad (16)$$

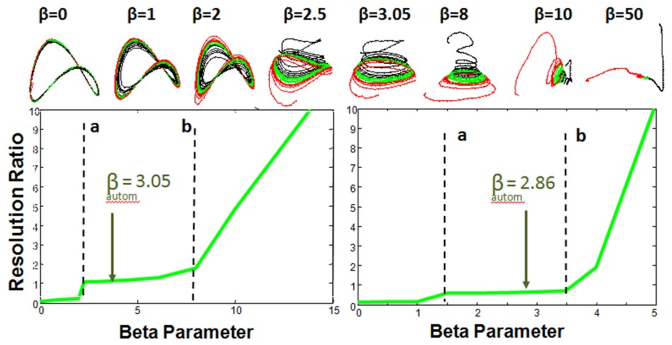


Figure 9. Top: Evolution of the Mocap manifold according to the β parameter. Bottom: Ratio between the style and the temporal resolution for different β values for the Mocap (left) and Mobo (right) manifolds (Automatic estimated values of β are highlighted).

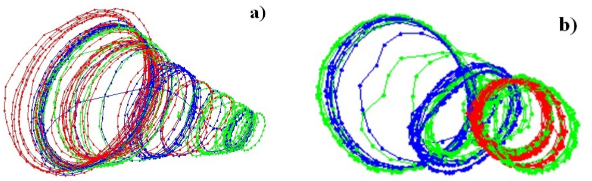


Figure 10. GLE manifolds: a) Mocap data from 3 individuals (red, green and blue) performing each 3 variations of an activity (walking, fast walking and running). b)Using Mobo foreground masks, profile view, 3 subjects (red, green and blue) performing each two variations (walking and fast walking).

where K is the total number of data points m_k in the data set. As discussed in section II-B and IV-A, three different cases are expected according to β values. Lower values, i.e. $\beta \in [0, a]$, suppress style variation and highlight temporal variations of an activity. Higher values, i.e. $\beta \in [b, \infty)$, suppress the temporal variation and discriminate between different styles. Finally, within our range of interest, i.e. $\beta \in [a, b]$, the generated manifolds preserve both the temporal and stylistic variability of an activity. In this range, Figure 9 displays plateaus where the resolution ratio between style and temporal distances is around 1, which ensures similar consideration of both variabilities. Conversely, the other two regions show imbalanced ratios where there is predominance of one mode over the other. Estimates of β, i.e. 3.05 for Mocap data and 2.86 for video data, generated from Eq. 9 fit in the plateau areas.

Figure 10 shows the manifolds generated by GLE using our automatic estimation of graph balance. Although, neither geometry nor axis mapping was imposed on the embedded space, GLE is able to exploit the three dimensions to generate coherent manifolds.

2) Mapping evaluation: Mapping performance was evaluated quantitatively using the walking2running dataset. The training set was composed of 2865 poses belonging to 2 different subjects, whereas 4800 frames from a different sequence was used for testing. Following the methodology proposed in [28], mapping accuracy was measured by evaluating the average distance between a 3D skeleton and the skeleton resulting from its projection into the embedded space and back in the original space. Our mapping approach is compared against other RBFN based mapping using k-means clustering and different types of activation functions, i.e. thin plates, spheres and Gaussians. As shown in Figure 11, multi-dimensional Gaussian activation functions perform best (last three columns). In addition, in the case of EM+FJ, not only EM clustering provides a similar performance to k-means with a considerably lower number of components, but, which is more important, the model-order is estimated automatically. The comparison between the performance for $n_g = 20$ or 200 shows how incorrect selection of this value significantly degrades performance.

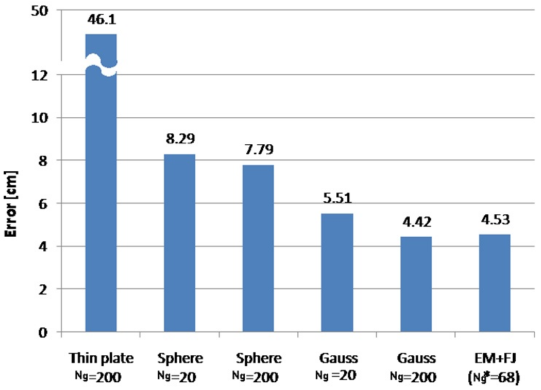


Figure 11. Mapping error for different RBFN learning approaches, using Mocap ”walking2running” dataset.

3) Qualitative analysis of data driven manifold geometry: The truncated conical shapes shown in Figure 10 are coherent with the nature of data where different people walk or run at different speeds. Assuming a constant sampling rate, given the periodic nature of bipedal locomotion, a gait cycle with a constant number of poses is expected. This number of samples per cycle is reduced when increasing the walking speed, although the samples are still temporarily and spatially similar to those of slower cycles. Consequently, in order to maintain the correspondence between similar poses, a reduction of the diameter of the cycles is expected, as depicted in Figure 12b. In a hypothetical situation where a subject were able to run at the speed of the capture rate (i.e. 120fps), only one pose would be captured by cycle, degenerating the cycle into a singular point (see Figure 12b). Therefore, locomotion styles up to this speed would be modelled by a manifold with a conical shape. If the subject were able to run even faster, an inverted cone would appear above that vertex. This explanation can be practically illustrated using sample interpolation so that the number of poses is constant for every step whatever the speed. As expected, a cylindrical structure is obtained as the resulting manifold geometry (see Figure 12c & Figure 13). It can be argued that this conical data driven geometry is, therefore, more coherent with the data than the cylindrical manifold shape imposed in LL-GPLVM (Figure 12a). Furthermore, unlike in a cylindrical geometry, in a conical one, the temporal distance between two connected frames is constant even when speed varies. Since we pursue to apply tracking on the manifold surface, a constant time factor simplifies drastically the definition of dynamic models. Finally, we illustrate the coherence of the GLE manifold by comparing poses which map at similar locations on the embedded space. We focus on the analysis of crossing points between different time series involving various individuals performing variations of the same bipedal locomotion (for instance, subject one -

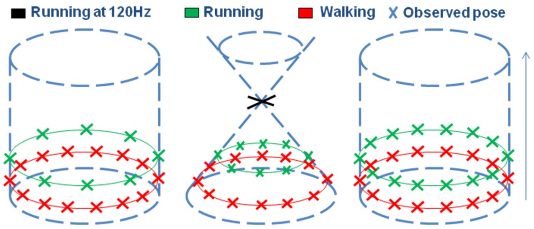


Figure 12. Manifold shapes for bipedal locomotion assuming style variation along the vertical axis: a) LL-GPLVM, b) GLE and c) GLE with resampled data.

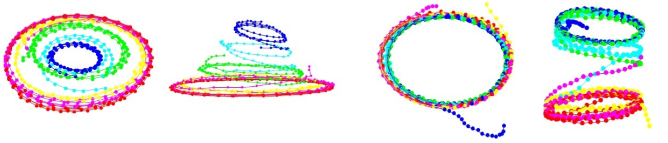


Figure 13. GLE manifolds obtained with either raw data (left) or resampled sequences (right) containing a constant number of poses per cycle, independently of the speed. Right figures show the projection on the first two dimension, while left ones show the first and third dimensions.

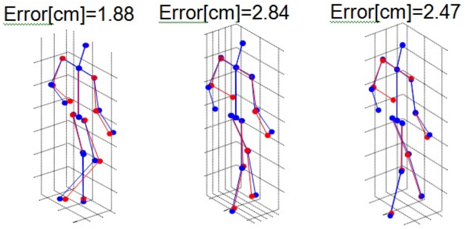


Figure 14. Reconstructed skeletons for 3 crossing points between subject 1 waking fast (red) and subject 2 running (blue).

red in Figure 14- is walking fast while subject two - blue in Figure 14- is running). On Figure 14, the reconstructed skeletons show that both poses represent similar stages of the locomotion cycle. Quantitatively, the average error per joint [64] between poses over the 255 crossing points that appear in our manifold is 2.4cm, which indicates that points of local convergence between different time series are captured by the manifold.

GLE has also been applied to non-cyclic activities. The experiment was conducted using the IXMAS dataset [55] which involves 12 different people performing different activities: “sitting dow”, “kicking”,“scratching the head”, “checking the watch” and “getting up”. For each sequence, salient points are calculated from the provided silhouette [56] as input vectors for GLE. Consistent results are obtained where stylistic variation is preserved in both cyclic (Figure 10) and non cyclic activities (Figure 15). Temporal continuity is represented in two of the dimensions (similar to TLE) whereas style is expressed along the third dimension. Activities that end with the same pose as they start (scratching, checking watch, kicking and cyclic motions) are connected in a circular shape into the temporal dimensions similarly to cyclic motions, while the other activities follow a rather linear and non-connected structure (sitting down, getting up).

The fact that the two first dimensions represent temporal information, whereas the third one expresses style is demonstrated in Figure 16 and in virtual motion videos provided in supplementary material. Note that they show that the manifold could be also applied for realistic human motion synthesis based on mixtures of styles [36].

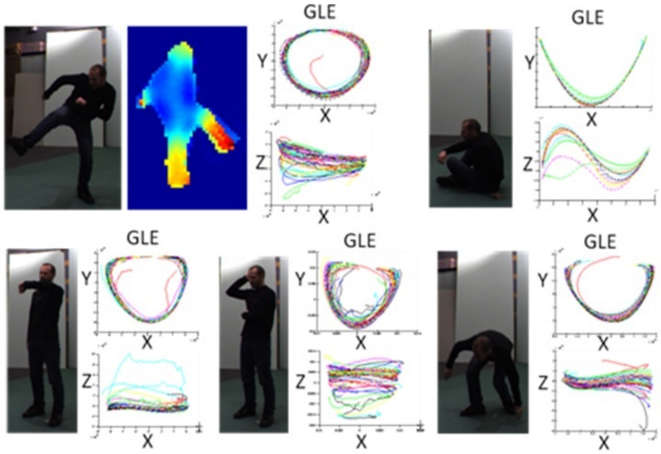


Figure 15. Manifolds obtained using IXMAS dataset for non cyclic activities (kicking, sitting down, checking watch, scratching the head and getting up). Different colours represent diferent subjects. An example of the salient point image used as input is given for the action “kicking”.

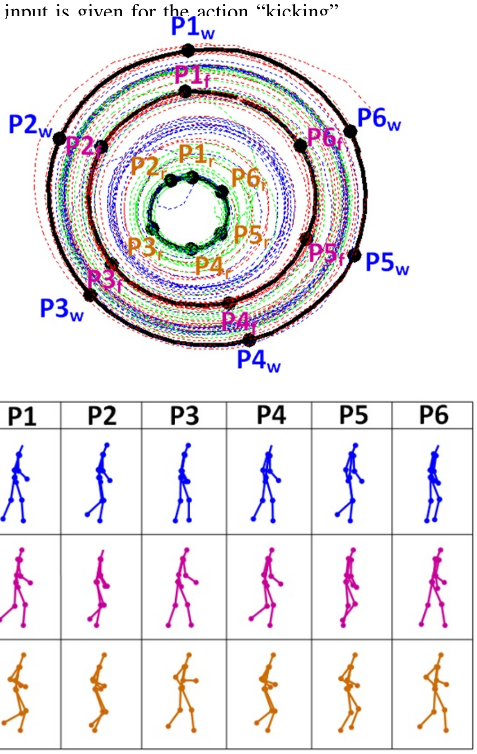


Figure 16. 3D poses projected from a manifold modelling different gait speeds.

4) Quantitative analysis of the reconstruction error: A comparative analysis is conducted to contrast GLE with six standard non-linear DR algorithms. Following the methodology evaluation proposed in [15], prior motion models are used to refine estimated poses provided by imperfect estimation systems, such as tracking frameworks [57], human detectors and inaccurate MoCap capture systems [58]. More specifically, a skeleton estimate is projected into the embedded space where it is associated to its nearest neighbour, whose projection is returned as the refined skeleton.

In order to simulate realistic pose estimates, a range of Gaussian noises $w \sim N(0,4)$, $w \sim N(0,16)$ and $w \sim N(0,64)$ is added to actual poses from the “walking2running” dataset. This results in an average error of 3.3cm, 6.4cm and

Table I
REFINING ERROR [CM].

Method	Error [cm]			
	No noise	*N(0,4)*	*N(0,16)*	*N(0,64)*
Random selection	8.5			
BCGPLVM	9.4	9.5	9.5	9.6
GPDM	7.3	7.2	7.4	7.5
Isomap	7.2	7.7	7.8	7.6
LLE	6.8	7.2	7.4	7.6
LE	8.8	9.3	9.7	8.9
TLE 2D	6.1	6.8	7.7	9.2
TLE 3D	5.6	6.1	7.9	9.6
GLE	**5.4**	**5.9**	**6.7**	**7.3**

12.7cm, respectively. The average error which would result from selecting a random pose from the training dataset is 8.5cm. The influence of the mapping learning process is removed from spectral based frameworks by applying the same mapping methodology to all those techniques ($L = 200$ Gaussians), with the exception of GPLVM-based approaches, where the mapping is an intrinsic part of the method. The training set is composed of 2408 poses belonging to 3 different subjects, whereas 3359 frames from another 3 different subjects (including a female) are used for testing. Without counting stylistic variations between successive steps, each set comprises 15 different styles: 3 people x (3 types of locomotion + 2 transitions). As shown in Table I, GLE performs the best in all experiments, even under very noisy conditions, where most methods perform worst than random selection. Error of more than 5.4cm in the ideal case, i.e. without noise, is explained by inaccuracy resulting from a combination of mapping error and high variability in the dataset. Due to the complexity of BCGPLVM optimisation process and the relatively large size of the dataset, the method fails to converge which leads to large errors.

This experiment also validates our decision of basing our stylistic approach on a low-computational cost embedded technique which computation time is at least one order of magnitude lower that the mapping-based approaches. Whereas GLE was able to learn its lower space in 47min, BCGPLVM and GPDM required 27h and 8h respectively for a MATLAB implementation on a Quad Core 3Ghz with 4 GB RAM. This is in line with the known limitations of mapping based approaches which suffer from high computational complexity [15].

Figure 17 shows the geometries of the embedded spaces obtained for different DR techniques. Most of them fail to represent effectively temporal and stylistic variations in their geometries. Inter-subject variability dominates above the other variations with GPDM, while temporality is lost in favour of large stylistic variations in the cases of LE and Isomap. On the other hand, although TLE is able to preserve temporality, style is completely discarded within its geometry. GLE is the only method able to preserve both temporal and stylistic variations. This can be observed along the third dimension where the inter-subject and intra-subject variability is represented in a common style space, while temporality is preserved in the other 2 dimensions.

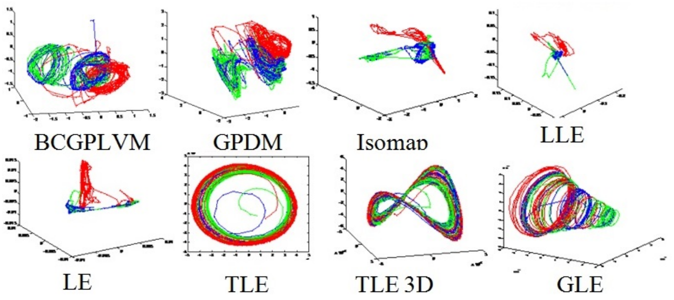


Figure 17. Embedded spaces generated by DR techniques, using MoCap ”walking2running” dataset. GLE produces the only geometry where inter-person variability and locomotion style are preserved.

Table II
SILHOUETTE REFINEMENT VALIDATION.

Methods	TPR	TNR	ACC
BcK. Subtraction + GLE	83.4	97.5	95.3
Adaptive motion detection [24]	61.3	95.5	90.8
Active Shape models [59]	67.9	94.8	91.2

5) Quantitative analysis of silhouette-based motion segmentation: A second quantitative analysis is performed through its application to silhouette-based human segmentation. In a training stage, GLE embedded space is created using binary silhouettes as input data. In the testing stage, extracted image-based silhouette is projected onto the manifold, refined to the nearest neighbour (NN) and projected back in order to obtain a refined blob. The GLE-based segmentation system is compared against two well-known methodologies: an advanced motion segmentation pipeline based on adaptive threshold, shadow removal and morphological operators [24] and an approach based on Active Shape Models (ASM) [59], which uses prior information to segment the shapes.

Our experiments with the MoBo dataset [54] involve 12 sequences of 51 frames each, comprising profile views of 3 different people walking following 2 stylistic variations, i.e. walking and walking fast. For each sequence, binary silhouettes were manually extracted to train the systems and to provide ground truth for quantitative evaluation of motion segmentation. The training and testing sets are composed of 6 different sequences each.

The standard metrics used to evaluate the refined silhouettes against the ground truth are sensitivity or true positive rate (TPR), specificity or True Negative Rate (TNR) and accuracy (ACC) [60].

$$TPR = T_p/(T_p + F_n) \quad TNR = T_n/(F_p + T_n) \tag{17}$$

$$ACC = (T_p + T_n)/(T_p + F_n + F_p + T_n) \tag{18}$$

where T_p and T_n are the numbers of true positive and true negative pixels respectively.

As depicted in Table II and Figure 18, the GLE based system produces more accurate blobs than the advanced motion detector and ASM. These results demonstrate that GLE can complement conventional image segmentation techniques, such as background substraction, when applied to human segmentation.

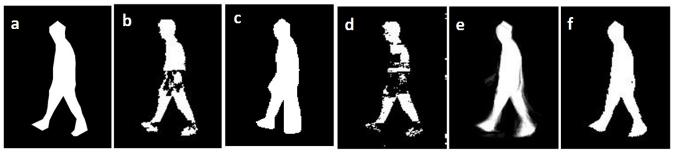


Figure 18. a) Ground truth, b) Silhouette extracted using [24], c) Silhouette estimated using [61], d) Binary image extracted by background subtraction, e) weighted response from GLE and f) refined silhouette.

C. GbPF validation

In this section, GLE Graph-based Particle filter is validated by comparing its performance with other state-of-the-art methodologies, i.e. conventional particle filter [62], annealed particle filter [20], and particle filter using GPLVM as a prior [42]. Although PF and APF do not include any training, they incorporate kinematic constraints based on human morphology to discard invalid hypotheses.

The proposed tracking framework is validated using HumanEVA [64] since it is widely accepted by the scientific community, provides numerical validation without access to groundtruth and is available for multi-style sequences (walking-running can be considered as different styles of bipedal locomotive activity). This point is especially relevant for us given our goal of validating a tracking system able to cope with inter-person intra-activity variability.

In order to demonstrate the generality of the framework and how it is able to infer effective models of human poses from a training set, we train the priors with a completely different set of sequences, the “walking2running” introduced in section IV-A). It is important to note that this dataset was created using a different MoCap system with different marker configurations and a different joint model than HumanEva. Sequences S2_Combo_1 and S4_Combo_1 from HumanEva II were used as test sequence. In sequence S4_Combo_1, frames 298 to 331 are ignored as accurate groundtruth is not available, as reported in [22].

The state vector x_t containing the parameters to be estimated by the particle filter is defined as:

$$x_t = \{x, y, z, \theta, \varphi, \vartheta, l_1, l_2, l_3\} \tag{19}$$

where x, y and z are the 3D coordinates of the base of the spinal cord, θ,φ and ϑ are the global rotation angles of the body regarding a fix 3D reference and l_1,l_2,l_3 are the coordinates of the 3D human configuration in the low dimensional space. Parameter χ has been set to 1, since training and test data display identical frame rates.

1500 particles were used in all the experiments based on PF, and 300 particles in 5 layers, which have the equivalent time complexity, for those relying on APF. The four synchronised cameras that composed the video dataset were employed to facilitate the comparison with previous approaches [22], [63], [64], but experiments with fewer cameras were also performed. Two sets of experiments were conducted using different likelihood functions, where the observation is modelled as either edges plus standard asymmetric silhouettes (E+S) (Figure 19) or bidirectional silhouettes (BS) [64] (Table III).

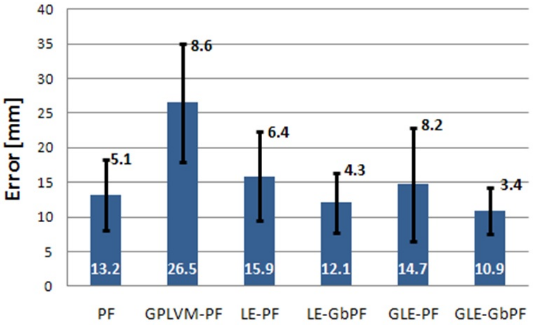


Figure 19. Performance comparison on HumanEVA II for walking sequences. Standard deviation is given as error bar. Results are given by computing frames [1-437] on S4_Combo_1 and [1:415] on S2_Combo_1. E+S observations from 4 cameras were used.

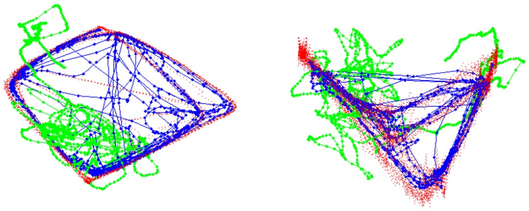


Figure 20. Tracking results on the LE manifold for S4_Combo_1 (HumanEva II) sequence. The training samples are represented in red, while dark blue corresponds to Graph-based particle filter estimations and green circles to conventional PF. The right figure shows the projection on the first two dimensions, while the left one displays the first and third dimensions.

Figure 19 reports the accuracy obtained by GLE-GbPF compared with alternative PF frameworks when using E+S observation. Note that, as reported in [64], methods are unable to track the subject over the full length of the sequence. Therefore, only the first section of each sequence is used in this experiment. First, the figure reveals that the inclusion of DR-based priors does not necessarily improve accuracy especially when this prior is not able to represent properly the stylistic variations of the test subject (i.e. GPLVM-PF), but also if they do not properly constrain the search space (i.e. GLE-PF).

Second, as proved by comparing the results between the tracking paradigms based on either conventional (i.e. LE-PF and GLE-PF) and Graph-constrained (i.e. LE-GbPF and GLE-GbPF) propagations, an adequate prediction schema for restricting the search space, such as Graph-based PF, improves tracking performance. This is illustrated in Figure 20 where Graph-constraints prevent divergence from the trained space even when using the limited prior model provided by LE.

In terms of the computational time required to evaluate the same number of particles, our schema doubles the PF’s cost due to the mapping functions and conversion from quaternions to cartesian coordinates. However, this cost is one order of magnitude lower than GPLVM-PF whose mapping functions are much more expensive.

Table III provides a quantitative comparison between our proposal and the state of the art, i.e. APF, using a more advanced observation model, i.e. BS, exploiting between 2 and 4 camera views. Results with a single camera are not reported because none of the methods managed to track a subject through the whole sequence, i.e. average errors were above 50 cm.

Accuracy of GLE-GbPF appears to be relatively good and independent of the number of cameras, if more than one view is available. This is due to additional information provided by GLE and its exploitation through Graph based propagation,

Table III
PERFORMANCE COMPARISON ON HUMANEVA II
Standard deviation is given between brackets. Results are given for frames corresponding to walking and running using the BS observation likelihood function and a variable numbers of cameras (C)

Error	S4_Combo_1	S2_Combo_1	Average
[cm]	[1:822]	[1:755]	
APF 4C	7.6 (1.5)	8.1 (1.4)	7.9 (1.5)
GLE-GbPF 4C	9.4 (1.7)	9.3 (1.5)	9.4 (1.6)
APF 3C	10.8(4.9)	8.8(3.2)	9.7(3.7)
GLE-GbPF 3C	9.9(4.4)	9.6(1.3)	9.8(2.8)
APF 2C	14.25(8.3)	10.1(3.1)	12.2(6.6)
GLE-GbPF 2C	9.7(2.1)	9.5(1.6)	9.6(1.9)

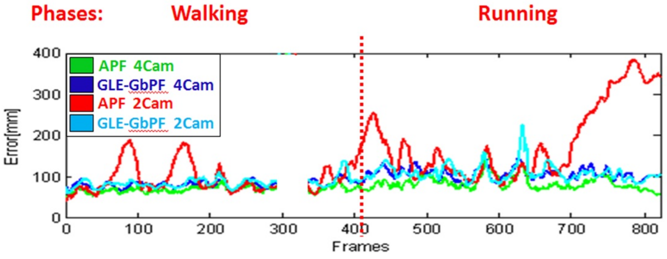


Figure 21. Numerical comparison for Graph-based Particle Filter and APF for S4_Combo_1 (HumanEva II) sequence using 4 (APF in green, GLE-GbPF in blue) and 2 cameras (APF in red, GLE-GbPF in cyan).

which helps to maintain the same level of performance when observation quality degrades. APF demonstrates a very different behaviour when the number of camera views decreases. Its performance is seriously affected when the scenario becomes underconstrained (see Figure 21).

In the ideal scenario, when views from the 4 cameras are available, APF seems to remain the state-of-the-art method. Clearly, APF combines efficient search in the pose space and effective constraints based on morphological knowledge. However, the fact that those constraints were learned, unlike in our case, from a dataset very similar to the testing one, i.e. HumanEvaI [64] where, for example, characters also perform their activity along a circular path, may suggest that APF's tracking task was easier than GLE-GbPF's.

In order to test further the potential of our framework, an additional monocular experiment was conducted. Since a single view showing a subject moving either towards or away from the camera hardly provides any information regarding individual limbs as illustrated by the failure of any system to process the HumanEVA II sequences, experiments were repeated using only short sub-sequences (i.e. 60 frames) where individuals walked in a plane almost parallel to the camera plane. Performances for 1 to 4 cameras are shown on Figure 22. These results confirm the superiority of GLE-GbPF over APF when observation quality is degraded. GLE-GbPF's usage of priors and efficient hypothesis propagation helps to reduce observation ambiguity.

Performance of GLE-GbPF is further illustrated in Figure 23 where pose estimates are projected in frames captured by camera 1 in both HumanEva sequences. Poses are correctly estimated during the walking and running phases, where the main error comes from poor estimation of the global rotation and translation parameters, as shown in Figure 23a third row, fourth column or Figure 23b second row, first column. It can also be noticed that accuracy in the estimation of the arms is usually lower than for legs' during the running phase. This is due to the biomechanical differences observed between actual running and hopping on a treadmill, which impacts more arm movement.

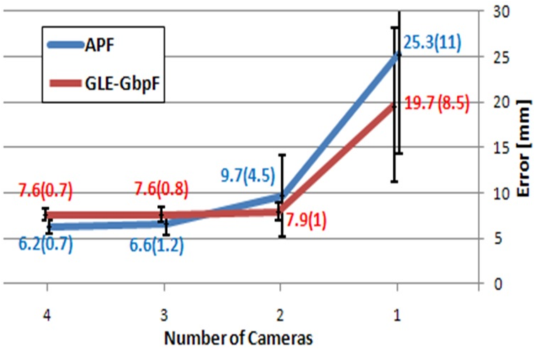


Figure 22. Numerical comparison between GLE-GbPF(blue) and APF(red) for SubSeq1&2 and varying number of cameras

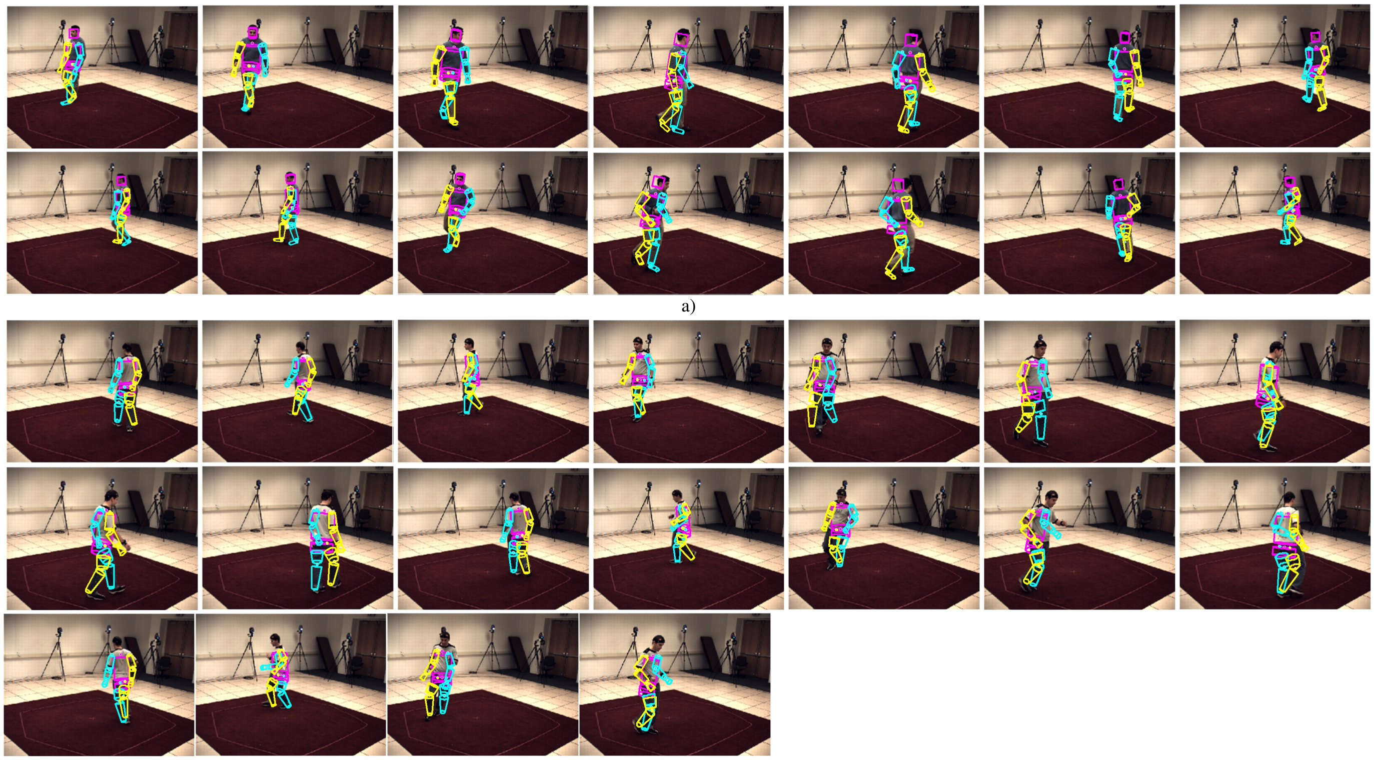
As conclusion, our combined methodology shows excellent generality and robustness properties outperforming the state of the art in underconstrained environments, which are likely to be more representative of realistic scenarios. In addition since such performance is achieved using totally unrelated training and testing tests, our methodology appears specially indicated for those applications where a subject/environment specific training cannot be assumed.

V. CONCLUSION

This paper presents GLE, a novel dimensionality reduction approach designed to address stylistic variations in time series. There are three main benefits of the method: a) the capacity of preserving style while reducing dimensionality in an unsupervised way, b) the generation of a continuous space which simultaneously models temporal and stylistic variations and c) the data-driven discovery of a manifold's geometry. GLE is based on the idea of weighted combination of temporal and stylistic neighbourhood graphs within the LE paradigm. The automatic estimation of the graph weight and an EM-GMM-based RBFN mapping scheme ensure the efficiency of the method. Experimental validation has shown qualitatively and quantitatively that GLE outperforms existing dimensionality reduction methods, since it is able to simultaneously cope with stylistic variations both from different people and activities.

We also introduce, GbPF, a novel methodology conceived for efficient tracking in low dimensional space derived from a spectral DR method. The strengths of our approach are a propagation scheme which facilitates the prediction in time and style, and a noise model coherent with the manifold. Tracking is constrained by the manifold surface, which prevents divergence, increases robustness and the probability of recovering after failure.

Finally, we propose a human pose tracker designed by combining GLE and GbPF, which displays state-of-the-art performance in underconstrained scenarios. As such, it extends prior-based tracking to new subjects, scenarios and environmental conditions which differ from training data.



a)

b)

Figure 23. Results for Graph-based Particle Filter for HumanEva II S2_Combo_1 (a) and S4_Combo_1 (b) sequence using 4 cameras and bi-directional silhouettes as observation. Frames: 1 to 700 and 822 respectively, every 50.

References

[1] R. Urtasun, D. J. Fleet and P. Fua, “Gaussian process dynamical models for 3D people tracking”, in *CVPR*, 2006.

[2] A. Elgammal and C. Lee, ”Body pose tracking from uncalibrated camera using supervised manifold learning,” in NIPS EHuM Workshop, 2006.

[3] R. Li, M. H. Yang, S. Sclaroff and T. P. Tian, ”Monocular tracking of 3D humanmotion with a coordinated mixture of factor analyzers,” in ECCV 2, 2006.

[4] W. Pan and L. Torresani, ”Unsupervised hierarchical modeling of locomotion styles,” in ICML, 2009.

[5] G. Taylor and G. Hinton, ”Factored conditional restricted Boltzmann Machines for modeling motion style,” in ICML, 2009.

[6] Z. Lu, M. Carreira-Perpian and C. Sminchisescu, ”People Tracking with the Laplacian Eigenmaps Latent Variable Model,” Advances in Neural Information Processing Systems, vol. 20, pp. 1705–1712, 2008.

[7] C. Sminchisescu and A. Jepson, ”Generative modeling for continuous non-linearly embedded visual inference,” in ICML, 2004.

[8] T. Matsubara, S. Hyon and J. Morimoto, ”Learning parametric dynamic movement primitives from multiple demonstrations,” ICONIP, vol. 1, pp. 347-354, 2010.

[9] A. Elgammal and C. Lee, ”Inferring 3d body pose from silhouettes using activity manifold learning,” in CVPR, 2004.

[10] K. Grochow, S. Martin, A. Hertzmann and Z. Popov, ”Style-based inverse kinematics,” in SIGGRAPH, 2004.

[11] R. Poppe, ”Evaluating Example-based Pose Estimation: Experiments on the HumanEva Sets,” in CVPR EHuM2, 2007.

[12] M. Vasilescu, ”Human motion signatures: analysis, synthesis, recognition,” ICPR, vol. 3, pp. 456- 460, 2002.

[13] J. M. Wang, D. J. Fleet and A. Hertzmann, ”Multifactor Gaussian process models for style-content separation,” in ICML, 2007.

[14] A. Safonova, J. K. Hodgins and N. S. Pollard, ”Synthesizing physically realistic human motion in low dimensional behavior-specific spaces,” in SIGGRAPH, 2004.

[15] M. Lewandowski, J. Martnez del Rincon, D. Makris and J. C. Nebel, ”Temporal extension of laplacian eigenmaps for unsupervised dimensionality reduction of time series,” in ICPR, 2010.

[16] A. Elgammal and C. S. Lee, ”Separating style and content on a nonlinear manifold,” in CVPR, 2004.

[17] J. Wang, D. Fleet and A. Hertzmann, ”Gaussian process dynamical models,” in NISP 18, 2006.

[18] H. Sidenbladh, M. J. Black and L. Sigal, ”Implicit probabilistic models of human motion for synthesis and tracking,” in ECCV 1, 2002.

[19] R. Urtasun, D. J. Fleet and N. Lawrence, ”Modeling human locomotion with topologically constrained latent variable models,” in HMUMCA Workshop, 2007.

[20] J. Deutscher, A. Blake and I. Reid, ”Articulated Body Motion Capture by Annealed Particle Filtering,” in CVPR, 2000.

[21] J. MacCormick and M. Isard, ”Partitioned sampling, articulated objects, and interface-quality hand tracking,” in ECCV 2, 2000.

[22] J. Darby, B. Li and N. Costen, ”Tracking human pose with multiple activity models,” Pattern Recognition, vol. 43, no. 9, pp. 3042-3058, 2010.

[23] T. Jaeggli, E. Koller-Meier and L. Van Gool, ”Multi-Activity Tracking in LLE Body Pose Space,” in 2nd Workshop on HUMAN MOTION Understanding, Modeling, Capture and Animation, ICCV, 2007.

[24] C. Orrite, J. Martinez, E. Herrero and G. Rogez, ”2D Silhouette and 3D skeletal models for human detection and tracking,” ICPR, vol. 4, pp. 244-247, 2004.

[25] M. Belkin and P. Nivogi, ”Laplacian eigenmaps and spectral techniques for embedding and clustering,” in NISP 14, 2001.

[26] J. Tenenbaum, V. Silva and J. Langford, ”A global geometric framework for nonlinear dimensionality reduction,” Science, vol. 290, no. 5500, p. 2319-2323, 2000.

[27] S. T. Roweis and L. K. Saul, ”Nonlinear dimensionality reduction by locally linear embedding,” Science, vol. 290, pp. 2323-2326, 2000.

[28] M. Lewandowski, D. Makris and J. Nebel, ”Automatic configuration of spectral dimensionality reduction methods,” Pattern Recognition Letters, vol. 31, no. 12, pp. 1720-1727, 2010.

[29] B. Schlkopf, A. J. Smola and K.-R. M, ”Kernel Principal Component Analysis,” in ICANN, 1997.

[30] N. Lawrence., ”Gaussian process latent variable models for visualisation of high dimensional data,” in NISP 16, 2004.

[31] M. Carreira-Perpiñan and Z. Lu, ”The Laplacian Eigenmaps Latent Variable Model,” JMLR W&P, vol. 2, pp. 59-66, 2007.

[32] S. G¨nter, N. Schraudolph and S. Vishwanathan, ”Fast Iterative Kernel Principal Component Analysis,” Journal of Machine Learning Research, vol. 8, p. 1893-1918, 2007.

[33] O. Jenkins and M. Mataric, ”A spatio-temporal extension to isomap nonlinear dimension reduction,” in ICML, 2004.

[34] N. Lawrence and J. Quinonero-Candela, ”Local distance preservation in the GP-LVM through back constraints,” in ICML, 2006.

[35] G. Taylor, L. Sigal, D. Fleet and G. Hinton, ”Dynamical binary latent variable models for 3D human pose tracking,” in CVPR, 2010.

[36] J. Min and H. Liu, ”Synthesis and editing of personalized stylistic human motion,” in SIGGRAPH i3D, 2010.

[37] A. Elgammal and C. Lee, ”Tracking People on a Torus,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 31, no. 3, pp. 520-538, 2009.

[38] H. Sidenbladh, M. Black and D. Fleet, ”Stochastic tracking of 3d human figures using 2d image motion,” in IEEE European Conference on Computer Vision, 2000.

[39] S. Wachter and H. Nagel, ”Tracking persons in monocular image sequences.,” Computer Vision and Image Understanding, vol. 74, no. 3, p. 174-192, 1999.

[40] N. R. Howe, M. E. Leventon and W. T. Freeman, ”Bayesian reconstruction of 3D human motion from single-camera video,” in NIPS 12, 2000.

[41] M. Brand, ”Shadow puppetry,” in ICCV, 1999.

[42] R. Urtasun, D. J. Fleet, A. Hertzmann and P. Fua, ”Priors for people tracking from small training sets,” in ICCV, 2005.

[43] G. W. Taylor, G. E. Hinton and S. Roweis, ”Modeling human motion using binary latent variables,” in NISP 19, 2007.

[44] C. Chen, L. Zhang, J. Bu, C. Wang and W. Chen, ”Constrained laplacian eigenmap for dimensionality reduction,” Neurocomputing, vol. 73, no. 4-6, pp. 951-958, 2009.

[45] M. Morse and J. Patel, ”An efficient and accurate method for evaluating time series similarity,” in SIGMOD ’07, 2007.

[46] L. Rabiner and B. H. Juang, Fundamentals of Speech Recognition, I. Prentice-Hall, Ed., 1993.

[47] D. M. Gavrila and L. S. Davis, ”Towards 3-d model-based tracking and recognition of human movement: a multi-view approach,” in IEEE International Workshop on Automatic Face- and Gesture-Recognition, 1995.

[48] M. Muller, Information Retrieval for Music and Motion, Springer-Verlag, 2007.

[49] M. Lewandowski, Advanced non linear dimensionality reduction methods for multidimensional time series: applications to human motion analysis, PhD Thesis ed., Kingston University, 2011.

[50] A. Elgammal and C. Lee, ”Nonlinear manifold learning for dynamic shape and dynamic appearance,” CVIU, vol. 106, no. 1, pp. 31-46, 2007.

[51] M. Figueiredo and A. Jain, ”Unsupervised learning on finite mixture models,” IEEE TPAMI, vol. 24, no. 3, pp. 381-396, 2002.

[52] S. Hauberg and K. Pedersen, ”Predicting Articulated Human Motion from Spatial Processes,” International Journal of Computer Vision, 2011.

[53] The Carnegie Mellon, ”The Carnegie Mellon University Graphics Lab Motion Capture Database,” March 2013. [Online]. Available: http://mocap.cs.cmu.edu.

[54] R. Gross and J. Shi, ”The CMU Motion of Body (MoBo) Database,” tech. report CMU-RI-TR-01-18, Robotics Institute, Carnegie Mellon University, June, 2001

[55] Inria, March 2013. [Online]. Available: http://4drepository.inrialpes.fr/public.

[56] L. Gorelick, M. Blank, E. Shechtman and M. Irani, ”Actions as space-time shapes,” IEEE TPAMI, vol. 29, pp. 2247-2253, 2007.

[57] J. Martinez, D. Makris, C. Orrite and J. Nebel, ”Tracking Human Position and Lower Body Parts Using Kalman and Particle Filters Constrained by Human Biomechanics,” IEEE Transactions on Systems, Man, and Cybernetics Part B, vol. 41, no. 1, pp. 26-37, 2011.

[58] P. Kuo, D. Makris and J. Nebel, ”Integration of bottom-up/top-down approaches for 2D pose estimation using probabilistic Gaussian modelling,” Computer Vision and Image Understanding, vol. 115, no. 2, pp. 242-255, 2011.

[59] T. Cootes, D. Cooper, C. Taylor and J. Graham, ”A trainable method of parametric shape description,” Image and Vision Computing, vol. 10, no. 5, p. 289-294, 1992.

[60] J. Neyman and E. Pearson, ”The testing of statistical hypotheses in relation to probabilities a priori,” Joint Statistical Papers, Cambridge University press, p. 186-202, 1933.

[61] A. Blake and M. Isard, Active contours, Springer, 1998.

[62] M. Isard and A. Blake, ”CONDENSATION - conditional density propagation for visual tracking,” Int. J. Computer Vision, vol. 29, no. 1, pp. 5-28, 1998.

[63] J. Darby, B. Li and N. Costen, ”Behaviour based particle filtering for human articulated motion tracking,” in ICPR, 2008.

[64] L. Sigal, A. Balan and M. J. Black, ”HumanEva: Synchronized Video and Motion Capture Dataset and Baseline Algorithm for Evaluation of Articulated Human Motion,” International Journal of Computer Vision, vol. 87, no. 1-2, 2010.

[65] M. Du and L. Guan, ”Du, M., Guan, L.: Monocular human motion tracking with the DE-MC particle filter,” in Int. Conf. on Acoustics, Speech, and Signal Processing, 2006.

[66] J. J. Pantrigo, A. Snchez, K. Gianikellis and A. S. Montemayor, ”Combining Particle Filter and Population-based Metaheuristics for Visual Articulated Motion Tracking,” Electronic Letters on Computer Vision and Image Analysis, vol. 5, no. 3, pp. 68-83, 2005.

[67] V. John, E. Trucco and S. J. McKenna, ”Markerless Human Motion Capture using Charting and Manifold Constrained Particle Swarm Optimisation,” in BMVC Postgraduate Workshop, 2010.

[68] V. Krger, J. Andersen and T. Prehn, ”Probabilistic Model-Based Background Subtraction,” in SCIA, 2005.

[69] M. Sivabalakrishnan and D. Manjula, ”RBF Approach to Background Modelling for Background Subtraction in Video Objects,” IJCS, vol. 1, no. 1, pp. 35-42, 2010.

[70] X. Geng, D. Zhan and Z. Zhou, ”Supervised nonlinear dimensionality reduction for visualization and classification”, in IEEE Transactions on systems, man, and cybernetics-B, vol.35, pp.1098-1107 ,2005.