Graph-based Particle Filter for Human Tracking with Stylistic Variations

Jesús Martínez del Rincón
http://cism.kingston.ac.uk/people/researcher/1214
Jean-Christophe Nebel
http://cism.kingston.ac.uk/people/academic/423
Dimitrios Makris
http://cism.kingston.ac.uk/people/academic/326

Digital Imaging Research Centre
Kingston University
UK

Abstract

In this paper, we propose an integrated particle filter-based pose tracking framework which combines priors able to model human motions keeping stylistic variations, reducing the probability of divergence and facilitating the recovering after failure. A novel unsupervised dimensionality reduction technique, Generalised Laplacian Eigenmaps (GLE), generates compact and coherent continuous spaces which explicitly express style. The proposed particle filter embeds the GLE manifold to take advantage of its geometry into the propagation and hypothesis generation stage. The method is validated using standard HumanEva 2 dataset.

1 Introduction

Articulated human tracking is one of the most active areas in computer vision due to its numerous applications such as video surveillance, gesture analysis, human computer interfaces and computer animation. However, it still remains as a major challenge due to the high complexity and dimensionality of the human pose space, which have a clearly negative impact on existing trackers, reducing their reliability at reconstructing human-like poses and making impossible recovering from failure.

Such difficulties have led to the development of approaches that address the size of the solution space, either using efficient search strategies such as annealing [1] and space partition [2] or by reducing its dimensionality [20-26]. Since the computational cost of search strategies increases with space dimensionality, many dimensionality reduction methods (DR) have been developed and explored as prior models for articulated human tracking. They are particularly relevant when dealing with the human pose space which, although it appears high-dimensional in its traditional individual angular parameterisation, has in fact a significantly smaller intrinsic dimensionality [3,4,5]. However, these processes may result in a loss of generality by compressing important information such as style, intra-activity variance and inter-subject variability.

In this paper we propose to address those natural limitations of model priors for articulated motion tracking. By employing a novel DR methodology able to preserve not only temporal information but also the stylistic variation among people, Generalised Laplacian Eigenmaps (GLE), we offer a tracking framework robust and sufficiently general so that it can be applied to different scenarios and actors. In addition, we suggest a novel and integrated tracking scheme, completely coherent with the prior formation process, which outperforms traditional methodologies only based of stochastic searching schemes in a lower dimensional space. As a result, this new scheme improves search

BMVC 2011 http://dx.doi.org/10.5244/C.25.105

efficiency, reduces risks of divergence and increases the probability of recovering after failure.

1.1 Related work

A low dimensional representation not only has to provide a compact and analytically tractable space suitable for search, but also must be sufficiently general to capture human pose variations. Since linear methodologies are not able to cope simultaneously with both requirements, this led to the development of many non linear models, such as mapping-based (Gaussian process latent variable model GPLVM [6]) and embedded-based approaches (Laplacian Eigenmaps LE [7], Isomap [8] and Local Linear Embedding [9]).

The exploitation of non-linear DR techniques for tracking in a lower-dimensional space requires locality in the low dimensional space, i.e. nearby regions in high dimensional space must be mapped to nearby regions in low dimensional space. If this property is not available, artificially high values of the noise model and complex non-linear dynamic models are required to deal with the absence of continuity inside the space. Several techniques, such as STIsomap [10], back constraint GPLVM (BC-GPLVM) [11], Gaussian process dynamical model (GPDM) [5] and Temporal Laplacian Eigenmaps (TLE) [4], have attempted to address this issue by introducing a temporal constrain to ensure smooth transition in the latent space. Although they succeeded in improving tracking performance for a given activity, they failed to represent stylistic variations, such as different people performing the same activity or the same person performing different variations of an activity. Although a few approaches have been suggested to deal with stylistic variations [12,13,14,15], none of them has been fully validated within a pose tracking framework.

Regarding the usage of prior models specifically for human articulated tracking, many different approaches have been proposed in the past [16,17,18,19,20]. However, the inclusion of manifolds produced by DR techniques has now become the most popular. Howe et al. [21] proposed Gaussian mixture representations of short human motion fragments in the high dimensional space integrated into a Bayesian MAP framework. Brand [22] and Sidenbladh et al. [23] also modelled the human pose manifold with a Gaussian mixture in combination with an HMM to infer the mixture component index. However, all these approaches model the priors by using linear and Gaussian methods, which are not adequate to describe the complexity of the human motion space. More recently, Sminchisescu and Jepson [24] proposed to associate the embedding space produced with a non-linear spectral method with a low-dimensional probabilistic model based on a simply parametric latent density (Gaussian mixture). Urtasun et al. [25] used a dynamic MAP estimation framework based on a more advanced prior learning methodology, i.e. GPLVM, and subsequently [26] extended their framework using GPDM to learn a latent space with associated dynamics. Li et al. [27] proposed a similar approach based on a different DR technique, LLC, where its coordinated mixture of factor analyzers are integrated within a particle filtering framework. However, the absence of dynamics makes it less accurate than GPDM. Finally, Taylor et al. [28] also learnt a binary latent space with dynamics (using an energy-based model) but applied it to motion synthesis, instead of tracking.

As a common characteristic, all these previous methods exploits the multi-hypothesis capabilities of particle filter to perform an efficient search in low dimensional spaces, where hypotheses are distributed in the low-dimensional space according to a generally unknown low-order dynamic model associated to a Gaussian noise. However, such approach does not prevent divergence when the noise is not constrain by manifold

geometry since hypotheses can move freely in the whole space instead of being constrained to remain in the vicinity of training points.

2 Methodology

In this section, we introduce our probabilistic tracking framework based on particle filter that integrates motion priors for robust and multi-style pose estimation. The priors are learned by applying GLE whose capacity to produce general manifolds allows preserving both temporal continuity and stylistic variations. The embedding of the prior is consistent with the nature of the GLE spectral method since it relies on graph information derived during training. This prior embedding supports a specialised particle filter in two ways. First it provides a propagation model with both temporal (dynamic model) and stylistic constraints. Secondly it provides automatically a suitable process noise model in the manifold created from training data. This prevents divergence towards invalid poses in the low dimensional space by ensuring moving in the vicinity of the manifold.

2.1 Prior model learning: Generalised Laplacian Eigenmaps (GLE)

Given a set of data points, $Y = \{y^k\}\ \forall k \in [1,M]$, distributed in a high dimensional space ($y^k \in \mathbb{R}^N$), LE is able to discover its low dimensional representation, $Z = \{m^k\}$ with ($m^k \in \mathbb{R}^n$), where $n << N$, which preserves the local structure of the original data by ensuring:

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

where L is the Laplacian matrix and D is the corresponding diagonal matrix with entries $D_{kk} = \sum_{j=1}^{M} G(k,j)$. G is a graph whose connectivity controls directly the similarity in the embedded space [7].

Since the LE framework only aims at preserving the local structure of each data point, in the case of time series, the produced embedded space may conserve neither the original temporal structure nor the style variance present in the training data. To address this, we proposed to express both the temporal structure and the style variance of the original data, 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 need of enforcing any artificial embedded geometry as in [12].

Similarly to [4], two types of neighbourhoods are automatically defined in GLE for each data point m^k:

- Temporal neighbourhood T_k: it ensures temporal continuity on the manifold. The $2t$ closest points:
$$T_k \in \{ m^{k-t}, \ldots, m^k, \ldots, m^{k+t} \} \tag{2}$$
are defined as the t-previous and the t-next points in the time series.
- Stylistic neighbourhood S_k: based on local geometry, it ensures stylistic continuity between training instances which are close in style. First, a temporal neighbourhood is defined around the point m^k. Then, Dynamic Time Warping (DTW) [29] is applied over a sliding window through the entire training set to detect and temporally align r_k repetitions of the temporal neighbourhood, $R_k^h\ \forall h \in [1, r_k]$. Finally, stylistic neighbours $R_k^h(l)\ \forall l \in [1, r_k]$ are selected as the closest points to m_k inside each repetition R_k^h.
$$S_k \in \{R_k^1(1), \ldots, m^k, \ldots, R_k^{r_k}(r_k)\} \tag{3}$$

Both neighbourhoods may be understood as constraints (Eq. 6) and modelled as graphs (Eq. 4 and 5).

$$G_T(k,j) = \begin{cases} e^{-\|y^k - y^j\|^2} & k,j \in T_k \\ 0 & otherwise \end{cases} \quad (4)$$

$$G_S(k,j) = \begin{cases} e^{-\|y^k - y^j\|^2} & k,j \in S_k \\ 0 & otherwise \end{cases} \quad (5)$$

$$L_T = D_T - G_T \qquad L_S = D_S - G_S \quad (6)$$

A manifold which includes temporal-stylistic coherence in its structure is generated by introducing these constraints with an appropriate balance β. The embedded space Z is spanned by the eigenvectors given by the n smallest nonzero eigenvalues λ where n is the number of resulting dimensions. They are obtained from the solution of the generalised eigenvalue problem [7], which is deduced by minimising the objective function:

$$\arg\min Z^T \cdot (L_T + \beta \cdot L_S) \cdot Z \quad (7)$$

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 $\beta = \infty$. A visual comparison between different LE-based methods and the influence of the temporal and stylistic is depicted in Figure 1. The internal structures of the GLE manifold and the connectivity given by the temporal and stylistic neighbours is shown in Figure 2.

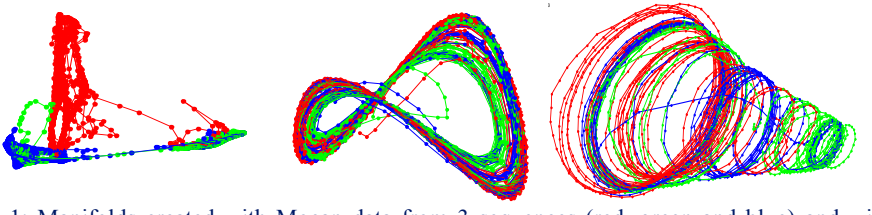

Figure 1: Manifolds created with Mocap data from 3 sequences (red, green and blue) and with 3 variations of an activity per sequence (walking, fast walking and running). Left: LE. Middle: Temporal LE. Right: GLE.

Since spectral methodologies such as LE do not provide explicitly any mapping mechanism between the low and high dimensional spaces, Radial Basis Function Networks (RBFN) are used to tackle this issue because they proved their effectiveness [4,13,30]. Direct ϕ and inverse functions ϕ' between high and low dimensional spaces are trained to provide projection functions.

$$\phi: \mathbb{R}^D \rightarrow \mathbb{R}^d \text{ and } \phi': \mathbb{R}^d \rightarrow \mathbb{R}^D. \quad (8)$$

2.2 Graph-based propagation and prediction for particle filter

Once the manifold has been created and a point on its surface has been selected as the initial pose, particles must be distributed and propagated. Traditionally, this is achieved by applying a low-order dynamic model and a Gaussian noise around that prediction [25,31]. 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.

We propose to tune the process noise on information provided by the GLE prior model. Specifically, a customised noise estimate is obtained for each point of the continuous low dimensional space by considering the RBFN functions as a Gaussian Mixture Model (GMM).

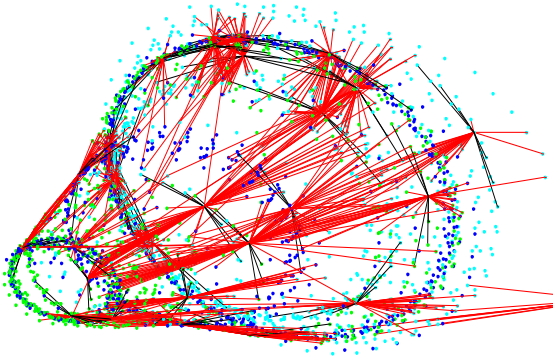

Figure 2: GLE manifold created with Mocap data from 3 sequences (cyan, green and blue) and with 3 variations of an activity per sequence (walking, fast walking and running). The learned temporal (black) and stylistic connectivity (red) given by the graphs G_T and G_S respectively, for a subset of randomly selected points.

First we propose the usage of multi-dimensional Gaussian activation functions ϕ_j (Eq. 8) in the RBFN.

$$\phi_j = e^{-(X-\mu_j)^T \cdot \Sigma_j^{-1} \cdot (X-\mu_j)} \tag{9}$$

for $j=1, \ldots, ng$, where X is the input feature vector and ng the number of Gaussians to be used. These particular functions are more suitable than traditional spherical functions for modelling and mapping the manifold given the intrinsic multidimensionality of a multi-style space. They will not only lower reconstruction error when projecting our hypotheses from the latent space to the 3D skeleton space, but also provide accurate modelling of the area around the manifold to produce valid hypotheses according to the training set. The noise covariance can be modelled, at any point in the low dimensional space x, as the covariance of a subset Nsg of the Gaussian set Ng belonging to the GMM which correspond to that point according to the Mahalanobis distance and a certain threshold τ.

$$\boldsymbol{\Sigma_{Noise}(x)} = \sum_{\boldsymbol{i \in N_{sg}}} \boldsymbol{\Sigma_i} \qquad \boldsymbol{N_{sg}} = \left\{ \forall \boldsymbol{j},\ (\boldsymbol{x}-\boldsymbol{\mu_j})^{\boldsymbol{T}} \cdot \boldsymbol{\Sigma_j^{-1}} \cdot (\boldsymbol{x}-\boldsymbol{\mu_j}) < \tau \right\}, \boldsymbol{N_{sg}} \subseteq \boldsymbol{N_g} \tag{10}$$

Although, this equation could provide a satisfactory noise level (as we will see in the result section) 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. This is addressed by integrating a combined searching and dynamical model into the manifold. This will be achieved by using the constraints used during the creation of the embedded space for particle propagation. In this manner, we propose an integrated methodology coherent with the manifold.

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 achieved by replacing the traditional deterministic propagation and prediction steps of particle filters by a stochastic propagation based on a triple resampling process.

Thus, firstly, particles are resampled to propagate valid hypotheses in the standard way according to their observation weight in the previous time step (Alg.1, 1a-1c).

The second resampling stage projects particles in time based on temporal graph G_T. It will associate each particle x_t^i to training points in the manifold m^k with a probability proportional to their Euclidean distance.

$$p\left(m^k \middle| x_t^i\right) \propto \exp\left(-d_{Euc}(x_t^i, m^k)/2\sigma^2\right) \tag{11}$$

Only one manifold point is randomly selected for each particle. Its corresponding temporal neighbour $\boldsymbol{m^{k+1} \in T_k}$ is then used as temporal prediction (Alg.1, 1d-1h).

The third resampling stage projects particles in the style dimension based on the stylistic graph $\boldsymbol{G_S}$. Again, the resampling is repeated for each particle and only one sample per particle is selected. All the stylistic neighbours $\boldsymbol{S_i}$associated to the temporal prediction of the resulting particle from the previous resampling are taken into account. Their probability is given by their values into the stylistic graph $\boldsymbol{G_S}$ (Alg.1, 1i-1l). Finally, Gaussian noise $\boldsymbol{p(x_t^i|m^s) \sim N(0, \Sigma_{Noise}(m^s))}$, as estimated by Eq.10, is added to the final set of particles in order to allow some degree of flexibility around the training manifold (Alg.1, 2).

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

Given a set of particles $\{\boldsymbol{x_{t-1}^i, \omega_{t-1}^i}\}_{i=1}^N$ which represents the posterior probability of $\boldsymbol{p(x_{t-1}|z_{t-1})}$ at time t-1, and a prior manifold $\{\boldsymbol{m^k}\}_{k=1}^M$

1. Select M samples from the set $\boldsymbol{x_{t-1}^i}$ with probability $\boldsymbol{\omega_{t-1}^i}$:
   a. Calculate the normalised cumulative probability $\boldsymbol{cx_{t-1}^n = \frac{\sum_{i=1}^n \omega_{t-1}^i}{\sum_{i=1}^N \omega_{t-1}^i}}$
   b. Generate a uniformly distributed random number $\boldsymbol{r \in [0,1]}$ and find the smallest j for which $\boldsymbol{cx_{t-1}^j \geq r}$
   c. Set $\boldsymbol{x'^i_{t-1} = x_{t-1}^j}$
   d. Generate M samples $\boldsymbol{\hat{x}_{t-1}^k}$ associated to manifold points $\boldsymbol{m^k}$ with a probability $\boldsymbol{\pi_{t-1}^k \propto \mathbf{exp}\,(-d_{Euc}(x'^i_{t-1}, m^k)/2\sigma^2)}$ where σ is a normalisation factor
   e. Calculate the normalised cumulative probability $\boldsymbol{c\pi_{t-1}^n = \frac{\sum_{k=1}^n \pi_{t-1}^k}{\sum_{k=1}^M \pi_{t-1}^k}}$
   f. Generate a uniformly distributed random number $\boldsymbol{r \in [0,1]}$ and find the smallest j for which $\boldsymbol{c\pi_{t-1}^j \geq r}$
   g. Set $\boldsymbol{x''^i_{t-1} = \hat{x}_{t-1}^j}$
   h. Propagate $\boldsymbol{x''^i_{t-1}}$ to the next time step $\boldsymbol{x''^i_t}$ according to the next temporal neighbour in the manifold given by $\boldsymbol{G_T(x''^i_{t-1}, x''^i_t)}$
   i. Generate $\boldsymbol{B \leq M}$ samples $\boldsymbol{\tilde{x}_t^k}$ associated to the manifold points $\boldsymbol{m^k}$ with a probability $\boldsymbol{\rho_t^k = G_S(\tilde{x}_t^k, m^k)}$
   j. Calculate the normalised cumulative probability $\boldsymbol{c\rho_t^n = \frac{\sum_{k=1}^n \rho_t^k}{\sum_{k=1}^B \rho_t^k}}$
   k. Generate a uniformly distributed random number $\boldsymbol{r \in [0,1]}$ and find the smallest j for which $\boldsymbol{c\rho_t^j \geq r}$
   l. Set $\boldsymbol{x'''^i_t = \tilde{x}_t^j}$
2. Noise addition $\boldsymbol{x_t^i = x'''^i_t + w_t^i}$ where $\boldsymbol{w_t^i \sim N(0, \Sigma_{Noise}(x'''^i_t))}$
3. Likelihood function evaluation $\boldsymbol{\omega_t^i \sim f(x_t^i, \phi, I_t)}$ over the input image $\boldsymbol{I_t}$
4. Estimate the mean state of the set $\boldsymbol{x_t^i}$ in the high dimensional space, $\boldsymbol{E[\phi(x_t)] = \sum_{i=1}^N \omega_t^i \cdot \phi(x_t^i)}$

Thank to this triple resampling strategy, see Alg. 1, we provide 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 $\boldsymbol{x_{t-1}^i}$, the prediction $\boldsymbol{x_t^i}$ is:

$$p(x_t^i|x_{t-1}^i) \propto p(x_t^i|m^s) \cdot G_S(m^s, m^{k+1}) \cdot G_T(m^k, m^{k+1}) \cdot p(m^k|x_{t-1}^i) \quad (12)$$

This methodology reduces the probability of diverging, increases the robustness and possibility of recovering after failure and facilitates the prediction in time and style by using the manifold information recorded in the connectivity graphs and its mapping functions.

Although the complexity of the algorithm increases due to this probabilistic procedure, the added computation time to the whole 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, and the number of hypotheses H to evaluate does not increase with the probabilistic procedure. The complexity introduced for the second resampling is O($H*P$) where H is the number of particles and P is the number of training points. In the third resampling, the maximum theoretical complexity would be O($H*P$) if the graph was fully connected. However, in practice, the percentage of connectivity c is around 1%, which leads to a complexity O($c*H*P$).

3 Experimental Results

The proposed algorithm is validated using a standard and well-known framework for articulated human pose estimation and tracking, HumanEVA [32]. This is achieved by integrating our priors as well as the propagation strategy into the APF [1] baseline algorithm provided by the dataset authors. The HumanEva evaluation framework has been chosen due to its acceptance among the scientific community, the numerical validation provided by the system without access to the grountruth and the availability of multi-style sequences (walking-running-balancing 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 effectively the intrinsic human pose from a training set to apply it in a different scenario, we train the priors with a completely different set of sequences. With this purpose, we use our new MoCap dataset, called "walking2running". 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/hour), as well as transitions between these three locomotion modes. The actions are performed on a treadmill to allow speed control. In our experiments 3 subjects produced 3 sequences of 4800-9000 frames each. 3D skeleton data are represented by quaternions of 13 joint angles.

The state vector $\boldsymbol{x_t}$ containing the parameters to be estimated by the particle filter is defined as $\boldsymbol{x_t} = \{\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\varphi}, \boldsymbol{\vartheta}, \boldsymbol{l_1}, \boldsymbol{l_2}, \boldsymbol{l_3}\}$, 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 $\boldsymbol{l_1}, \boldsymbol{l_2}, \boldsymbol{l_3}$ are the coordinated of the 3D human configuration in the low dimensional space. As setup parameters 1500 particles were used in all the experiments for the version based on PF, and 100 particles in 5 layers for APF. The four synchronised cameras that composed the video dataset were employed.

A comparative analysis is conducted to contrast GLE Graph-based Particle filter with other methodologies from the state of the art, such as conventional particle filter, annealed particle filter, or particle filter using GPLVM as a prior. Sequence S4_Combo_1 – the most complex sequence of the dataset - was used as test sequence in order to demonstrate the contribution of graph-based propagation when using a zero order dynamic model and the noise estimation given by Eq. 10.

Error [cm]	S4 _Combo_1	
E+S PF	14.1 (6.5)	
E+S APF	14.5 (9)	
E+S GPLVM-PF	17.58 (10.1)	
E+S GLE-PF	13.7 (9.5)	
E+S GLE-GbPF	12.2 (7.5)	
BS PF*	11.9 (8.1)	13.8 (9.1)*-
BS GLE-GbPF*	11.4 (7.8)	12.6 (6.4)*

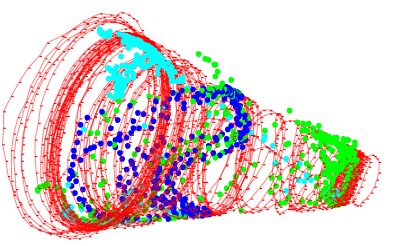

Figure 3: Left. Performance comparison on HumanEVA II. Standard deviation is given between brackets. Results are given for frames [1-437] (before divergence of all the methods) in the E+S and BS experiments and for [1-830] (walking and running) in the BS experiments (indicated with an *). Right. Results on the manifold for Graph-based Particle Filter for S4_Combo_1 (HumanEva II) sequence using bi-directional silhouettes. Dark blue corresponds to walking (frames 1-370), green to running (371-830) and cyan to balancing (frames 831-1257).

We can observe in the table in Figure 3 left the improvement in accuracy achieved by using priors in conjunction with particle filter. However, this effect is small or even detrimental and does not represents a competitive advantage, especially when this prior is not able to represent properly the stylistic variations of the test subject (see GPLVM-PF). The inclusion of the graph-based propagation model shows a much clearer improvement which, in combination with the capacity of GLE for representing stylistic variations, is able to cope with the running phase, the walking phase and their transitions.

This comparative study was made by using the default observation based on edges plus silhouettes (E+S) as likelihood function. For this observation, results reported in the state of the art [32] shows an average error of 14cm, which matches with the results we obtained with a conventional particle filter. However, they are outperformed by our new method which 14% (or 2cm) more accurate. As reported in [32], 'E+S' is unable to track the subject over the full length of the sequence. In this experiment, trackers diverge after the frame 437 for all tested methodology. On the other hand [32]demonstrated that the bidirectional silhouettes (BS) was suitable for this task and able to cope with fast running motions. Using this observation, our method not only processes the whole sequence, but also remains superior to particle filter. The quality of the new observation is highlighted by the fact that on the first 437 frames, it is in average 1cm more accurate than E+S when using our method.

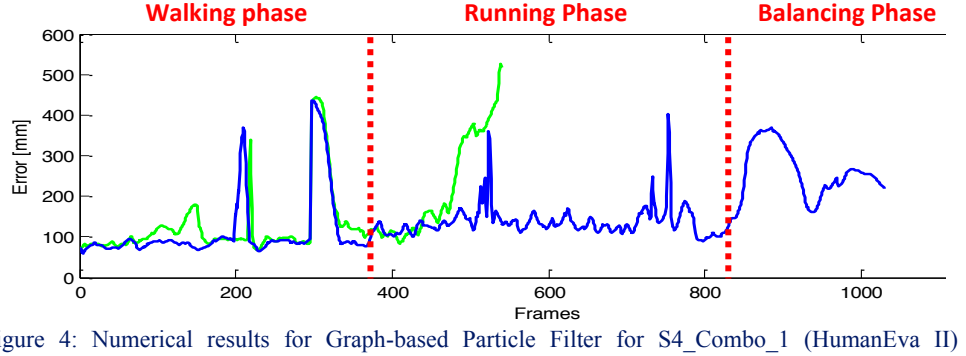

Figure 4: Numerical results for Graph-based Particle Filter for S4_Combo_1 (HumanEva II) sequence using Edges+Silhouettes (green) and bi-directional silhouettes (blue) as observation.

Figure 4 shows the error per frame using our methodology (GLE-GbPF) using both observation schemes. It shows that graph-based propagation provides added robustness

preventing divergence from the manifold (except when observation is very poor), which allows recovering from large errors due to poor estimations of the global position and rotation (predicted by simply adding Gaussian noise to their previous values) in frames 200, 514 and 732. Figure 5 shows the volumetric representation of the final estimation for few frames seen by camera 1. Finally, Figure 3 right shows the estimation provided by particle filter within our framework on the low dimensional space. A colour code has been used to classify the activity performed by the subject. Thus, in dark blue, we can see the poses theoretically corresponding to walking and how they are placed on the area of the manifold corresponding to walking poses base of the cone in most of the cases. Similarly, the running poses in green are also situated in the area corresponding to running in the training set (top of the cone). The balancing poses, although estimated wrongly due to their absence in the training set, they are also placed on the surface of the cone.

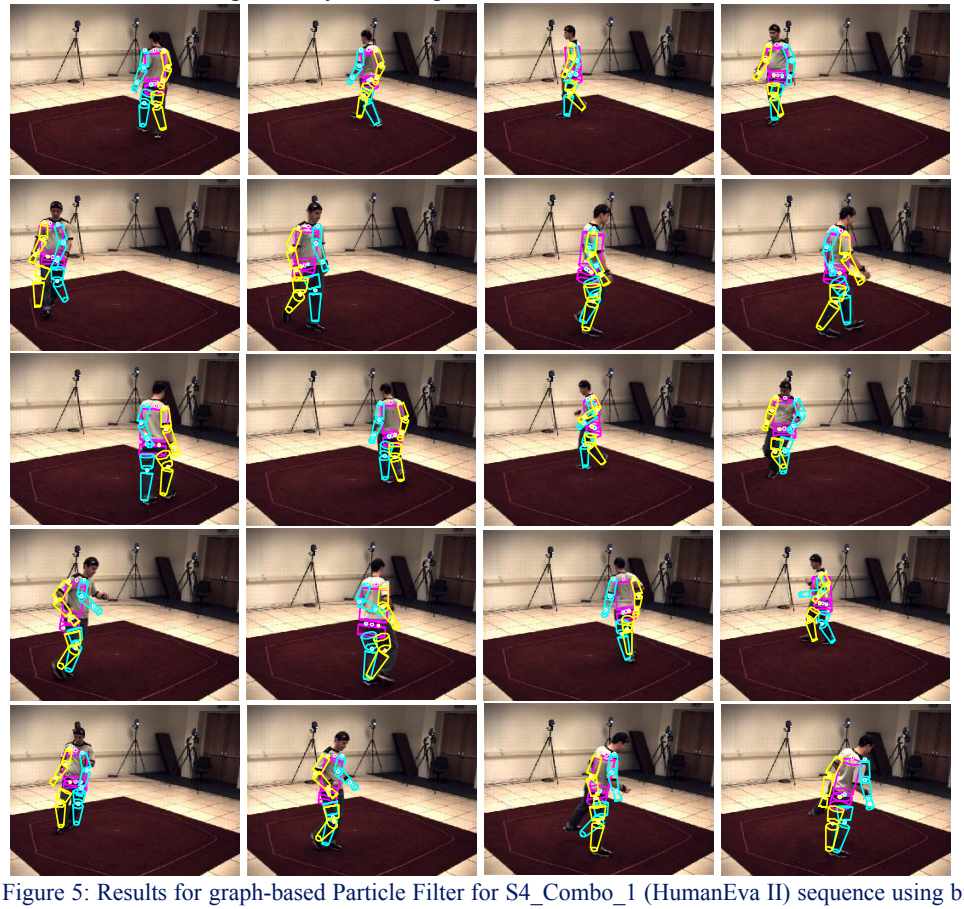

Figure 5: Results for graph-based Particle Filter for S4_Combo_1 (HumanEva II) sequence using bi-directional silhouettes as observation. Frames: 1 to 900, every 50.

4 Conclusions

In this paper, we introduce a novel tracking framework based on particle filter which integrates priors based on a dimensionality reduction and a graph-based propagation scheme. Our system is able to successfully track different stylistic variations thank to the usage of a new DR technique, GLE, for modelling the space of activity. These GLE-based

priors are capable of representing not only variations in the execution of a family of activities (walking, running) but also those due to the individual particularities among subjects. This allows tracking of new subjects, scenarios and environmental conditions which are present in different datasets. In addition, the graph-based particle filter ensures a coherent propagation and prediction of particles which follow the training data by moving on the manifold surface, avoiding divergence, increasing the robustness and the probability of recovering after failure and facilitating the prediction in time and style. As future work, we plan to extend the methodology for activities of different nature in order to cope with complex scenarios of activities.

References

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

[2] J. MacCormick and M. Isard, "Partitioned sampling, articulated objects, and interface-quality hand tracking," in *ECCV 2*, 2000, pp. 3-19.

[3] A. Safonova, J. K. Hodgins, and N. S. Pollard, "Synthesizing physically realistic human motion in low dimensional behavior-specific spaces," in *SIGGRAPH*, 2004, p. 514 – 521.

[4] M. Lewandowski, J. Martínez del Rincón, D. Makris, and J. C. Nebel, "Temporal extension of laplacian eigenmaps for unsupervised dimensionality reduction of time series," in *ICPR*, 2010, pp. 161-164.

[5] J. Wang, D. Fleet, and A. Hertzmann, "Gaussian process dynamical models," in *NISP 18*, 2006, p. 1441–1448.

[6] N. Lawrence., "Gaussian process latent variable models for visualisation of high dimensional data," in *NISP 16*, 2004, pp. 329-336.

[7] M. Belkin, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *NISP 14*, 2001, p. 585–591.

[8] J. Tenenbaum, V. Silva, and J. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, p. 2319–2323, 2000.

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

[10] O. Jenkins and M. Matarić, "A spatio-temporal extension to isomap nonlinear dimension reduction," in *ICML*, 2004, p. 441–448.

[11] N. Lawrence and J. Quinonero-Candela, "Local distance preservation in the GP-LVM through back constraints," in *ICML*, 2006, p. 513–520.

[12] R. Urtasun, D. J. Fleet, and N. Lawrence, "Modeling human locomotion with topologically constrained latent variable models," in *HMUMCA Workshop*, 2007, pp. 104-118.

[13] A. Elgammal and C. S. Lee, "Separating style and content on a nonlinear manifold," in *CVPR*, 2004, pp. 478-485.

[14] W. Pan and L. Torresani, "Unsupervised hierarchical modeling of locomotion styles," in *ICML*, 2009, pp. 785-792.

[15] J. M. Wang, D. J. Fleet, and A. Hertzmann, "Multifactor Gaussian process models for style-content separation," in *ICML*, 2007, pp. 975-982.

[16] 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, p. 205–208.

[17] J. J. Pantrigo, A. Sánchez, 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.

[18] J. Darby, B. Li, and N. Costen, "Behaviour based particle filtering for human articulated motion tracking," in *ICPR*, 2008, pp. 1-4.

[19] 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.

[20] 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.

[21] N. R. Howe, M. E. Leventon, and W. T. Freeman, "Bayesian reconstruction of 3D human motion from single-camera video," in *NIPS 12*, 2000, p. 820–826.

[22] M. Brand, "Shadow puppetry," in *ICCV*, 1999, p. 1237–1244.

[23] H. Sidenbladh, M. J. Black, and L. Sigal, "Implicit probabilistic models of human motion for synthesis and tracking," in *ECCV 1*, 2002, p. 784–800.

[24] C. Sminchisescu and A. Jepson, "Generative modeling for continuous non-linearly embedded visual inference," in *ICML*, 2004, p. 759–766.

[25] R. Urtasun, D. J. Fleet, A. Hertzmann, and P. Fua, "Priors for people tracking from small training sets," in *ICCV*, 2005, p. 403–410.

[26] R. Urtasun, D. J. Fleet, and P. Fua, "Gaussian process dynamical models for 3D people tracking," in *CVPR*, 2006, p. 238–245.

[27] 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, p. 137–150.

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

[29] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Inc. Prentice-Hall, Ed., 1993.

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

[31] Z. Lu, M.A. Carreira-Perpiñan, and C. Sminchisescu, "People Tracking with the Laplacian Eigenmaps Latent Variable Model," *Advances in Neural Information Processing Systems*, vol. 20, pp. 1705--1712, 2008.

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