HUMAN POSE TRACKING IN LOW DIMENSIONAL SPACE ENHANCED BY LIMB CORRECTION

Alexandros Moutzouris, Jesus Martinez-del-Rincon, Michal Lewandowski, Jean-Christophe Nebel, Dimitrios Makris

Digital Imaging Research Centre, Kingston University, UK

ABSTRACT

This paper proposes a two-level 3D human pose tracking method for a specific action captured by several cameras. The generation of pose estimates relies on fitting a 3D articulated model on a Visual Hull generated from the input images. First, an initial pose estimate is constrained by a low dimensional manifold learnt by Temporal Laplacian Eigenmaps. Then, an improved global pose is calculated by refining individual limb poses. The validation of our method uses a public standard dataset and demonstrates its accurate and computational efficiency.

Index Terms— 3D Pose Tracking, Dimensionality Reduction, Temporal Laplacian Eigenmaps, Visual Hull.

1. INTRODUCTION

Since many applications, such as human computer interaction and visual surveillance, rely on the analysis of human motion, the development of human tracking systems is a highly active research area. Tracking exploits the temporal coherence of video sequences to estimate pose parameters over time. However, due to the complexity of human actions, the detection of each body part separately is a challenging task.

A common tracking approach is particle filter and specific variants have been used with human data [1]. For example, the Annealed Particle Filter (APF) [2] replicates the annealing procedure into a particle filter framework in order to better search the pose space. However, because of the high dimensionality of the human pose space, trackers tend to provide sub-optimal solutions.

In order to deal with this issue, dimensionality reduction algorithms, such as Isomap [3], Local Linear Embedding [4] and Gaussian Process Latent Variable Model (GPLVM) [5], have been exploited to find low dimensional representations. Their application to human articulated motion data has been particularly prolific, especially when a single activity is assumed [6]. Urtasun et al. proposed a 3D human tracking pipeline based on Gaussian Process Dynamical Models (GPDMs) to learn a low dimensional space of human poses [7]. Raskin et al. [8] used GPDM to create a low-dimensional space where a Hierarchical-APF operates to generate particles. The main drawback of using low dimensional manifolds in tracking frameworks is that they do not provide satisfactory solutions when the actual poses are away from the manifold [9] due to unseen stylistic variations of the activity.

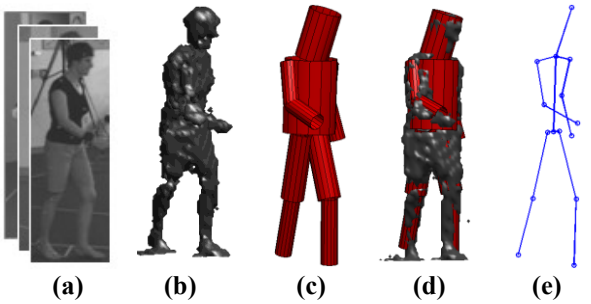

Fig.1. a) Input images, **b)** computed visual hull H , **c)** volumetric model M , **d)** fitted model to visual hull, **e)** extracted skeleton.

In this paper we present a two-level 3D pose tracking approach whose multiple cameras input is used to extract a 3D articulated human model. We call this method Manifold Projection – Limb Correction (MPLC). It is based on a deterministic search instead of particle filter to avoid the increase of computational cost due to the propagation of particles across levels [8] and deals with the problem of stylistic variations of human activity by using a refinement process.

In a first level, 3D human poses are constrained on a low dimensional activity manifold by optimizing a full-body likelihood function. Since accurate tracking requires a temporally smooth and consistent data model, the proposed constraining manifold is generated by Temporal Laplacian Eigenmaps (TLE) [10], which aims at preservation of the temporal topology present in high dimensional spaces. In the second level, individual limb poses are refined by optimizing a likelihood function for each limb separately. The validation of our method shows higher accuracy than a standard particle filter approach of the same computational cost.

2. METHODOLOGY

2.1. Tracking constrained by spectral manifold

The first stage of our approach is based on the projection of human poses on a low dimensional manifold, using a spectral dimensionality reduction method and its associated Radial Basis Function Network (RBFN) mapping [11]. We call this method Manifold Projection (MP). In this work, we do not deal with the problem of global tracking, therefore we assume that global rotation and translation are given for every frame.

First, Visual Hull is created from the silhouettes of the input images, using a shape-from-silhouette 3D reconstruction technique. We adopt the Bounding Edge [12] method for its estimation. A Visual Hull H_j at time t_j is described as a set of 3D voxel points (see Fig.1 (b)).

In order to extract a 3D pose from the Visual Hull, we use a 3D articulated human model M and a likelihood function f to evaluate fitting between M and H. Model M has L limbs and it is visualized by a skeleton and a volumetric representation (Fig. 1 (c) and (e)). The volumetric representation is described as the set of 3D voxel points associated to body parts, i.e. legs, arms, torso and head, which are represented by cylinders. We define function f as the overlap between M and H:

$$f(M,H) = |M \cap H| . \quad (1)$$

We assume that the training data $Y = \{y_j, j = 1,...,n\}$, $y_j \in \mathbb{R}$ corresponds to n consecutive frames of a specific action A, distributed on a manifold in a high dimensional space. A manifold $X = \{x_j, j = 1,...,n\}$, $x_j \in \mathbb{R}$ with $d < D$ is created using the TLE dimensionality reduction method. TLE has been chosen since it has been proved more accurate in the particular case of human articulated models than others dimensionality reduction methods [10]. In order to provide generative abilities to unseen examples, a RBFN is learned to provide projection functions between high and low dimensional spaces $\phi : \mathbb{R}$ $\mathbb{R}$ and $\phi' : \mathbb{R}$ $\mathbb{R}$.

For the current frame i, the MP method consists of five steps (Fig.2).

Step 1: The model of the previous frame M_{i-1} is projected to the low dimension space $\mathbb{R}$ where P_{i-1} is the projection point, $P_{i-1} = \phi(M_{i-1})$.

Step 2: The closest point in the manifold X to point P_{i-1}, Q_i, is estimated.

Step 3: A sample of K points is selected from a neighborhood W_i of point Q_i on the manifold X:

$$W_i' = \{Q_{i,k}', k = 1,...,K\}, Q_{i,k}' \in \mathbb{R} . \quad (2)$$

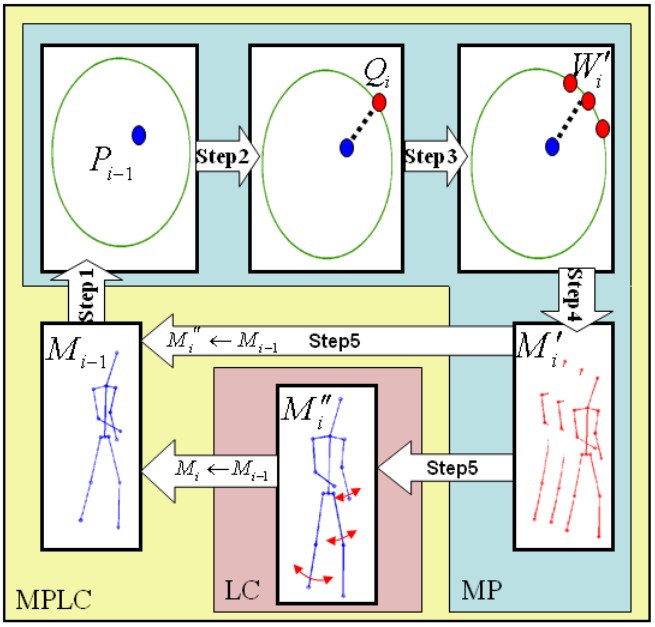

Fig.2. Flowchart of MP, LC and MPLC pipelines.

Step 4: All points of W_i' are back-projected to the high dimension space $\mathbb{R}$. Let

$$M_i' = \{M_{i,k}', k = 1,...,K\} \quad (3)$$

be the set of candidate model representations in $\mathbb{R}$.

Step 5: Finally, every M_i' is compared with the Visual Hull H_i using the observation function f. The best pose M_i'' is chosen by maximizing the function $f(M_{i,k}', H_i)$, i.e.

$$M_i'' = \{M_{i,k}', k : \arg\max_r f(M_{i,r}', H_i)\} . \quad (4)$$

2.2. Limb Error Detection and Correction

Since the manifold representation is constrained by the training data, there may be some discrepancy between the observed limbs and the manifold poses because of stylistic variations intrinsic to every subject. Therefore, the previous process needs to be refined to deal with this issue.

The second stage of our approach applies Limb Correction (LC) for those limbs with significant error. The input of the LC stage is the 3D model M_i''. Every limb $N^z, z = 1,...,L$ of the model may be rotated by a rotation matrix R^z, and let $N^z(R^z)$ be the limb after the rotation.

First, voxel points of the Visual Hull H_i that falls into the torso T are removed, and let H_i be the new Visual Hull, $H_i \leftarrow H_i - T$.

Then, we identify in a hierarchical order which limb N^j would benefit from a refinement process. This is achieved by quantifying the relative overlap between the Visual Hull H_i and the limb of interest:

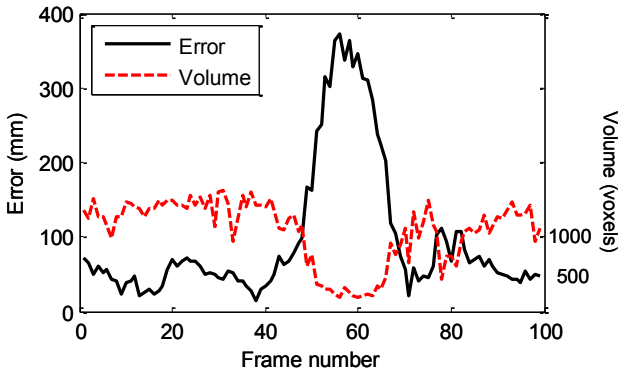

Fig.3. Poses error and intersection metric per frame. **Black:** Error of a limb (arm) using the method MP. **Red:** Volume of the intersection of the Visual Hull and the limb.

$$\frac{f\left(N^j, H_i\right)}{\left|N^j\right|} < h \ . \tag{5}$$

This metric is justified by Fig. 3, which demonstrates the inverse relationship between the overlapping volumes and the limb error.

If the relative overlap is higher than a threshold h, then we search in a range of joint angles that correspond to rotation matrices $R^b, b = 1,...,B$. We select the best pose $N^j(R)$ by maximizing the function $f(N^j(R^b), H_i)$,

$$N^j(R) = \max_k f(N^j(R^b), H_i) \ . \tag{6}$$

Before assessing the next limb, we remove the voxel points of the Visual Hull H_i that fall into the limb $N^j(R)$, and update the Visual Hull H_i:

$$H_i \leftarrow H_i - N^j(R) \ . \tag{7}$$

Finally, the pose solution M_i is derived after considering the new limb poses $N^j(R)$ that have been estimated by the limb correction process.

3. EXPERIMENTAL RESULTS

In this section we present the results of our method and we compare it with an equivalent, i.e. of similar computational cost, particle filter approach (PF) [13]. It is applied on the 2D low dimensional space, where n particles are back-projected on the high dimensional space to be evaluated by the likelihood function (3).

The Image & MOCAP Synchronized Dataset (IMSD) [14] and HumanEva Dataset [1] have been used for our experiments. The dimensionality of the pose space is reduced to 2, which is its intrinsic value has shown in [10]. Our training set contains 1121 frames of the S3 walking sequence in trial 3 from HumanEva I. The IMSD (walking) is used for testing.

A standard metric proposed by Sigal [1] is applied for quantitative evaluation: for each of the 15 points of the skeleton representation ($L = 10$) the error is calculated as the Euclidean distance between the point of the Skeleton Model and the corresponding point of the ground truth.

Parameters K and n are set to ensure similar computational times for each methodology. In MP method we use $K = 1$ to $K = 25$ with a step of 3 in equation (2). In equation (5) for LC method we use $h = 50\%$, and we choose an area ± 10 degrees, with a step of one, for every joint angle. In PF method we use n particles from 5 to 50, with a step of 5.

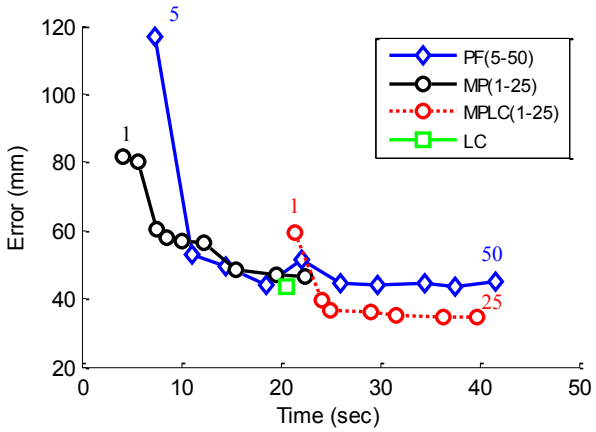

Fig.4. Comparison of average errors for 100 frames according to the average computational time for each frame.

Fig. 4 represents the average error for 100 frames as a function of the average computational time for each frame, for PF, MP, MPLC and LC methods. By fixing the processing time we can obtain a direct comparison between MPLC and PF. For instance, the average computational time of MPLC and PF methods is approximately 30sec/frame. The corresponding average error for PF is 44mm (standard deviation $\sigma = 12$ mm) while MPLC's is 35mm ($\sigma = 10$ mm). As we can see, the MPLC method produces better results than PF for the same computational cost.

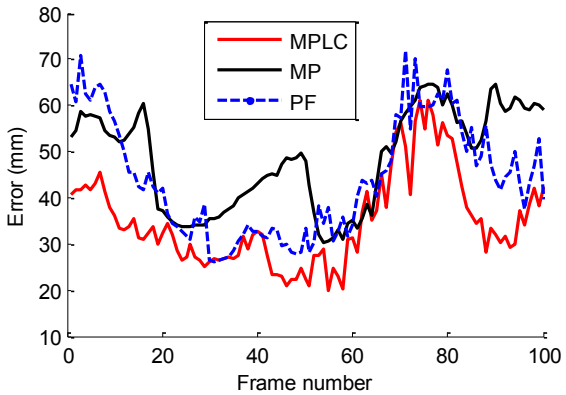

Fig.5. Average error per frame for 100 frames processed by methods MPLC, MP and PF.

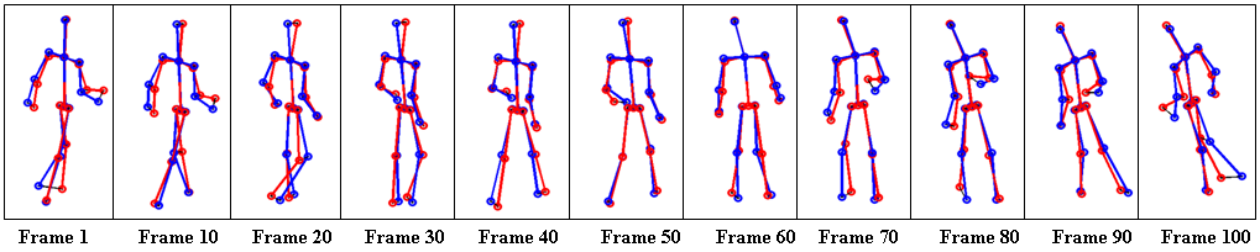

Fig.6. Skeleton models for **Red:** ground truth and for **Blue:** our method **(**MPLC15)

Fig. 5 displays the average error for every frame, as a function of the frame number, for the MP (black), and MPLC (red) methods with $K=15$ and for PF method (blue) with 35 particles. MPLC and MP average computational time are approximately 30sec/frame and 15 sec/frame respectively. Whereas MP's accuracy is 48mm ($\sigma=10$ mm), MPLC's is only 35mm ($\sigma=10$ mm). In Fig. 6 we compare skeleton models generated by MPLC method with $K=15$, to ground truth.

We can conclude that the inclusion of the LC extension provides a significantly advantage in terms of accuracy with an acceptable computational load increase. The LC extension could also be applied into a two step PF framework. However, in practice, a particle filter would be much more computationally expensive, as all particles would have to be propagated (and then evaluated) from the PF to the LC stage. In all cases, our proposed method (MPLC) outperforms the particle filter, assuming the same computational time.

4. CONCLUSION

Particle filters are popular techniques for human tracking. Those methods are computationally expensive because of the large number of particles that they use. Moreover, they tend to provide sub optimal solutions due to the high dimensionality of the human pose space. Consequently, usage of dimensionality reduction methods is a very attractive pre-processing step. However, they exclude the style of activities.

To deal with those problems, we propose a two stage human tracking method. First, an initial pose is estimated on a low-dimensional manifold space. We use a deterministic search instead of a particle filter to avoid the high computational cost of particle evaluation. Secondly, we deal with the problem of stylistic variations of human activity by refining each limb individually.

This study demonstrates that our method estimates better solutions than a particle filter of similar computational cost.

In future work, we will integrate in our framework a module which calculates the character's global position. In addition, we will extend our system to handle multiple activities.

5. REFERENCES

[1] L. Sigal, A.O. 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 (IJCV), Springer, vol. 87, no. 1, pp. 4-27, 2010.

[2] J. Deutscher, A. Blake and I. Reid, "Articulated body motion capture by annealed particle filtering," CVPR, vol 2, pp. 126-133, 2000.

[3] J.B. Tenenbaum, V. Silva and J.C. Langford, "A global geometric framework for nonlinear dimensionality reduction," Science, vol. 290, no. 5500, pp. 2319-2323, 2000.

[4] S.T. Roweis and L.K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," Science, vol. 290, no. 5500, pp. 2323-2326, 2000.

[5] N.D. Lawrence, "Gaussian process latent variable models for visualisation of high dimensional data," NIPS 16, pp. 329-336, MIT Press, 2004.

[6] R. Poppe, "Vision-based human motion analysis: An overview," Computer Vision and Image Understanding, Elsevier, vol. 108, no. 1-2, pp 4-18, 2007.

[7] R. Urtasun, D.J. Fleet and P. Fua, "3D people tracking with Gaussian process dynamical models," IEEE, CVPR, vol. 1, pp. 238-245, 2006.

[8] L. Raskin, M. Rudzsky and E. Rivlin, "3D human body-part tracking and action classification using a hierarchical body model," in Proceedings of the British Machine Vision Conference, 2009.

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

[10] M. Lewandowski, J. Martinez-del-Rincon, D. Makris and J-C. Nebel, "Temporal Extension of Laplacian Eigenmaps for Unsupervised Dimensionality Reduction of Time Series," International Conference on Pattern Recognition (ICPR), 2010.

[11] M. Lewandowski, D. Makris and J-C. Nebel, "Automatic Configuration of Spectral Dimensionality Reduction Methods for 3D Human Pose Estimation," Workshop on Visual Surveillance, 2009.

[12] K.M. Cheung, S. Baker and T. Kanade, "Shape-from-silhouette across time part i:Theory and algorithms," IJCV, vol. 62, pp. 221-247, 2005.

[13] M. Isard and A. Blake, "Condensation—conditional density propagation for visual tracking," IJCV, Springer, vol. 29, no. 1, pp. 5-28, 1998.

[14] Brown University, "Image & MOCAP Synchronized Dataset(v.1.0)," http://www.cs.brown.edu/~ls/Software/index.html, Accessed 20/1/2011.