3D Human Motion Tracking Using Dynamic Probabilistic Latent Semantic Analysis

Kooksang Moon and Vladimir Pavlović
Department of Computer Science, Rutgers University
Piscataway, NJ 08854
{ksmoon, vladimir}@cs.rutgers.edu

Abstract

We propose a generative statistical approach to human motion modeling and tracking that utilizes probabilistic latent semantic (PLSA) models to describe the mapping of image features to 3D human pose estimates. PLSA has been successfully used to model the co-occurrence of dyadic data on problems such as image annotation where image features are mapped to word categories via latent variable semantics. We apply the PLSA approach to motion tracking by extending it to a sequential setting where the latent variables describe intrinsic motion semantics linking human figure appearance to 3D pose estimates. This dynamic PLSA (DPLSA) approach is in contrast to many current methods that directly learn the often high-dimensional image-to-pose mappings and utilize subspace projections as a constraint on the pose space alone. As a consequence, such mappings may often exhibit increased computational complexity and insufficient generalization performance. We demonstrate the utility of the proposed model on the synthetic dataset and the task of 3D human motion tracking in monocular image sequences with arbitrary camera views. Our experiments show that the proposed approach can produce accurate pose estimates at a fraction of the computational cost of alternative subspace tracking methods.

1. Introduction

Estimating 3D body pose from 2D monocular image is a fundamental problem for many applications ranging from surveillance to advanced human-machine interfaces. However, the shape variation of 2D images caused by changes in pose, camera setting, and view points makes this estimation a challenging problem. Computational approaches to pose estimation in these settings are often characterized by complex algorithms and a tradeoff between the estimation accuracy and computational efficiency. In this paper we propose the low-dimensional embedding method for 3D pose estimation that exhibits both high accuracy, tractable estimation, and invariance to viewing direction.

3D human pose estimation from monocular 2D images can be formulated as the task of matching an image of the tracked subject to the most likely 3D pose. To learn such a mapping one needs to deal with a dyadic set of high dimensional objects—the poses, y and the image features, z. Because of the high dimensionality of the two spaces learning a direct mapping $z \rightarrow y$ often results in complex models with poor generalization properties. One way to solve this problem is to map the two high dimensional vectors to a lower dimensional subspace x: $x \rightarrow z$ and $x \rightarrow y$ [3, 10]. However, in these approaches, the correlation between the pose and the image feature is weakened by learning the two mappings independently and the temporal relationship is ignored during the embedding procedure.

Our approach to pose estimation is inspired by probabilistic latent semantic analysis (PLSA) that is often used in domains such as the computational language modeling to correlate complex processes. In this paper we extended PLSA to account for the dynamic nature of sequential data. We choose the Gaussian Process Latent Variable model (GPLVM) to form a pair of mapping functions between the latent variable and the two objects. A GPLVM framework is particularly suited for this model because the dynamic nature of sequence can be directly integrated into the embedding procedure in a probabilistic manner [25, 12, 23]. The two generative nonlinear embedding models and the marginal dynamics results in a new hybrid model called the dynamic PLSA (DPLSA). DPLSA models the semantical relationship between a sequence of 3D poses and a sequence of image features via the shared latent dynamic space.

This paper is organized as follows. We first define the DPLSA model that utilizes the marginal dynamic prior to learn the latent space of sequential data. We then propose the new framework for human motion modeling based on the DPLSA model and suggest learning and inference methods in this specific modeling context. The framework can be directly extended for multiple view points by using the mixture model in the space of the latent variables and the image features. The utility of the the new framework is examined thorough a set of experiments of tracking 3D human figure motion from synthetic and real image sequences.

2. Related Work

Dyadic data refers to a domain with two sets of objects in which data is measured on pairs of units. One of the popular approaches for learning from this kind of data is the latent semantic analysis (LSA) that was devised for document indexing. Deerwester *et al.* [2] considered the term-document association data and used singular-value decomposition to decompose document matrix into a set of orthogonal matrices. LSA has been applied to a wide range of problems such as information retrieval and natural language processing [14, 1].

Probabilistic Latent Semantic Analysis (PLSA) [5] is a generalization of LSA to probabilistic settings. The main purpose of LSA and PLSA is to reveal semantical relations between the data entities by mapping the high dimensional data such text documents to a lower dimensional representation called latent semantic space. Some exemplary application areas of PLSA in computer vision include image annotation [11] and image category recognition [4, 18].

Latent space approach for high-dimensional data has been applied to human motion tracking problems in the past. Various dimensionality reduction techniques such as Principal Components Analysis (PCA), isometric feature mapping (Isomap), Local linear (LLE) and spectral embedding have been successfully used in human tracking [16, 3, 19].

Recently, the GPLVM that produces a continuous mapping between the latent space and the high dimensional data in a probabilistic manner [8] was used for human motion tracking. Tian *et al.* [22] use a GPLVM to estimate the 2D upper body pose from the 2D silhouette features. Urtasun *et al.* [24] exploit the SGPLVM for 3D people tracking. Wang *et al.* [25] introduced Gaussian Process Dynamic Model (GPDM) that utilizes the dynamic priors for embedding and the GPDM is effectively used for 3D human motion tracking [23]. In [12] marginal AR prior for GPLVM embedding is proposed and utilized for 3D human pose estimation from the synthetic and real image sequences. Lawrence and Moore [9] propose the extension of GPLVM using a hierarchical model in which the conditional independency between human body parts is exploited with low dimensional non-linear manifolds. However, these approaches utilize only the pose in latent space estimation and as a consequence, the optimized latent space cannot guarantee the proper dependency between the poses and the image observations in regression setting.

Shon *et al.* [15] propose a shared latent structure model that utilizes the latent space that links corresponding pairs of observations from the multiple different spaces and apply it for image synthesis and robotic imitation of human actions. Although their model also utilizes GPLVM as the embedding model, their applications are limited to non-sequential cases and the linkage between two observations is explicit(*e.g.*image-image or pose-pose). The shared latent structure model using GPLVM is utilized for pose estimation in [13]. This work focuses on the semi-supervised regression learning and makes use of unlabeled data (only pose or image) to regularize the regression model. In contrast, our work, using a statistical foundation of PLSA, focuses on the computational advantages of the shared latent space. In addition, it explicitly considers the latent dynamics and the multi-view setting ignored in [13].

3. Dynamic PLSA with GPLVM

The starting point of our framework design is the symmetric parameterization of Probabilistic Latent Semantic Analysis [5]. In this setting the co-occurrence data $y \in Y$ and $z \in Z$ are associated via an unobserved latent variable $x \in X$:

$$P(y, z) = \sum_{x \in X} P(x)P(y|x)P(z|x). \tag{1}$$

With a conditional independence assumption, the joint probability over data can be easily computed by marginalizing over the latent variable. We extend the idea to the case in which the two sets of objects, Y and Z are sequences and the latent variable x_t is only associated with the dyadic pair (y_t, z_t) at time t. And we solve the dual problem by marginalizing the parameters in the conditional probability models instead of marginalization of Z.

Consider the sequence of length T of M-dimensional vectors, $Y = [y_1 y_2 ... y_T]$, where y_t is a human pose (*e.g.*joint angles) at time t. The corresponding sequence $Z = [z_1 z_2 ... z_T]$ represent the sequence of N-dimensional image features observed for the given poses. The key idea of our Dynamic Probabilistic Latent Semantic Analysis (DPLSA) model is that the correlation between the pose Y and the image feature Z can be modeled using a latent-variable model where two mappings between the latent variable X and Y and between X and Z are defined using a Gaussian Process latent variable model of [8]. In other words, X can be regarded as the intrinsic subspace that Y and Z jointly share. The graphical representation of DPLSA for human motion modeling is depicted in Fig. 1.

We assume that sequence $X \in \Re^{D \times T}$ of length T is generated by possibly nonlinear dynamics modeled as a known

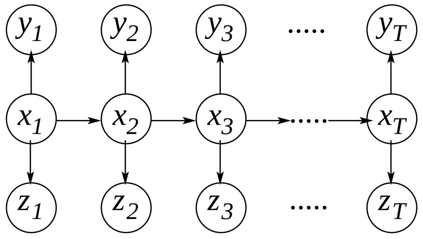

Figure 1. Graphical model of DPLSA.

mapping ϕ parameterized by parameter γ_x [25, 23] such as

$$x_t = A_1\phi_{t-1}(x_{t-1}|\gamma_{x,t-1}) + A_2\phi_{t-2}(x_{t-2}|\gamma_{x,t-2}) + \ldots + w_t. \tag{2}$$

Then the 1st order nonlinear dynamics is characterized by the kernel matrix

$$K_{xx} = \phi(X_\Delta|\gamma_x)\phi(X_\Delta|\gamma_x)^T + \alpha^{-1}I. \tag{3}$$

The model can further be generalized to higher order dynamics.

The mapping from X to Y is a generative model defined using a GPLVM [8]. We assume that the relationship between the latent variable and the pose is nonlinear with additive noise, v_t a zero-mean Gaussian noise with covariance $\beta_y^{-1}I$:

$$y_t = Cf(x_t|\gamma_y) + v_t. \tag{4}$$

C represents a linear mapping matrix and $f(\cdot)$ is a nonlinear mapping function with a hyperparameter γ_y. By choosing the simple prior of a unit covariance , zero mean Gaussian distribution on the element c_{ij} in C and x_t, marginalization of C results in a mapping:

$$P(Y|X, \beta_y) \sim |K_{yx}|^{-M/2} \exp\left\{-\frac{1}{2}tr\{K_{yx}^{-1}YY^T\}\right\} \tag{5}$$

where

$$K_{yx}(X, X) = f(X|\gamma_y)f(X|\gamma_y)^T + \beta_y^{-1}I. \tag{6}$$

Similarly, the mapping from the latent variable X into the image feature Z can be defined by

$$z_t = Dg(x_t|\gamma_z) + u_t. \tag{7}$$

The marginal distribution of this mapping becomes

$$P(Z|X, \beta_z) \sim |K_{zx}|^{-N/2} \exp\left\{-\frac{1}{2}tr\{K_{zx}^{-1}ZZ^T\}\right\} \tag{8}$$

where

$$K_{zx}(X, X) = g(X|\gamma_z)g(X|\gamma_z)^T + \beta_z^{-1}I. \tag{9}$$

Notice that the kernel functions and parameters are different in the two mappings from the common latent variable sequence X to Y and to Z.

The joint distribution of all co-occurrence data and all intrinsic sequence in a $\mathcal{Y} \times \mathcal{Z} \times \mathcal{X}$ space is finally modeled as

$$\begin{aligned} &P(X, Y, Z|\theta_x, \theta_y, \theta_z) = \\ &\qquad P(X|\theta_x)P(Y|X, \theta_y)P(Z|X, \theta_z) \end{aligned} \tag{10}$$

where $\theta_y \equiv \{\beta_y, \gamma_y\}$ and $\theta_z \equiv \{\beta_z, \gamma_z\}$ represent the sets of hyperparameters in the two mapping functions from X to Y and from X to Z. θ_x represents a set of hyperparameters in the dynamic model (*e.g.*α for a linear model and α, γ_x for a nonlinear model).

3.1. Human Motion Modeling Using Dynamic PLSA

In human motion modeling, one's goal is to recover two important aspects of human motion from image feature: (1) 3D posture of the human figure in each image and (2) an intrinsic representation of the motion. Given a sequence of image features Z, the joint *conditional* model of the pose sequence Y and the corresponding embedded sequence X can be expressed as

$$\begin{aligned} &P(X, Y|Z, \theta_x, \theta_y, \theta_z) \propto \\ &\qquad P(X|\theta_x)P(Y|X, \theta_y)P(Z|X, \theta_z). \end{aligned} \tag{11}$$

Notice that the two mapping processes $P(Y|X)$ and $P(Z|X)$ have different noise models which can account for different factors (*e.g.*motion capture noise for the pose and camera noise for the image) that influence one but not the other process.

3.2. Learning

Human motion model is parameterized using a set of hyperparameters θ_x, θ_y and θ_z, and the choice of kernel functions, K_{yx} and K_{zx}. Given both the sequence of poses and the corresponding image features, the learning task is to infer the subspace sequence X in the marginal dynamics space and the hyperparameters. Using the Bayes rule and (11) the joint likelihood is in the form

$$\begin{aligned} &P(X, Y, Z, \theta_x, \theta_y, \theta_z) = P(X|\theta_x)P(Y|X, \theta_y) \\ &\qquad P(Z|X, \theta_z)P(\theta_x)P(\theta_y)P(\theta_z). \end{aligned} \tag{12}$$

To mitigate the overfitting problem, we utilize priors over the hyperparameters [25, 23, 15] such as $P(\theta_x) \propto \alpha^{-1}$ (or $\alpha^{-1}\gamma_x^{-1}$), $P(\theta_y) \propto \beta_y^{-1}\gamma_y^{-1}$ and $P(\theta_z) \propto \beta_z^{-1}\gamma_y^{-1}$.

The task of estimating the mode X^* and the hyperparameters, $\{\theta_x^*, \theta_y^*, \theta_z^*\}$ can then be formulated as the ML/MAP estimation problem

$$\{X^*, \theta_x^*, \theta_y^*, \theta_z^*\} = \arg \max_{X,\theta_x,\theta_y,\theta_z} \{\log P(X|\theta_x) \\ + \log P(Y|X, \theta_y) + \log P(Z|X, \theta_z)\} \quad (13)$$

which can be achieved using generalized gradient optimization such as CG, SCG or BFG. The task's nonconvex objective can give rise to point-based estimates of the posterior $P(X|Y, Z)$ that can be obtained by starting the optimization process from different initial points.

3.3. Inference and Tracking

Having learned the DPLSA on training data Y and Z, the motion model can be used effectively in inference and tracking. Because we have two conditionally independent GPs, estimating current pose (distribution) y_t and estimating current point x_t in the embedded space can be decoupled. Given image features z_t in frame t the optimal point estimate x_t^* is the result of the following nonlinear optimization

$$x_t^* = \arg \max_{x_t} P(x_t|x_{t-1}, \theta_x)P(z_t|x_t, \theta_z). \quad (14)$$

Due to GP nature of dependencies, the second term assumes conditional Gaussian form, however its dependency on x_t is nonlinear [8] even with linear motion models in x. As a result, the tracking posterior $P(x_t|z_t, z_{t-1}, \ldots)$ may become highly multimodal. We utilize a particle-based tracker for our final pose estimation during tracking. However, because the search space is the low dimensional embedding space, only a small number of particles ($<$ 20, empirical result) is sufficient for tracking allowing us to effectively avoid the computational problems associated with sampling in high dimensional spaces.

A sketch of this procedure using particle filter based on the sequential importance sampling algorithm with N_P particles and weights ($w^{(i)}, i = 1, ..., N_P$) is shown below.

```
Input  : Image $z_t$, Human motion model e.g.(10)
         and prior point estimates
         $(w_{t-1}^{(i)}, x_{t-1}^{(i)}, y_{t-1}^{(i)})|Z_{0..t-1}, i = 1, ..., N_P$.
Output: Current intrinsic state estimates
         $(w_t^{(i)}, x_t^{(i)})|Z_{0..t}, i = 1, ..., N_P$
1) Draw the initial estimates $x_t^{(i)} \sim p(x_t|x_{t-1}^{(i)}, \theta_x)$.
2) Find optimal estimates $x_t^{(i)}$ using nonlinear
optimization in (14).
3) Find point weights
$w_t^{(i)} \sim P(x_t^{(i)}|x_{t-1}^{(i)}, \theta_x)P(z_t^{(i)}|x_t^{(i)}, \theta_z)$.
```

Algorithm 1: Particle filter in human motion tracking.

Finally, because the mapping from X to Y is a GP function, we can easily compute the distribution of poses y_t for each particle $x_t^{(i)}$ by using the well known result from GP theory: $P(y_t|x_t^{(i)}) \sim \mathcal{N}(\mu^{(i)}, {\sigma^{(i)}}^2 I)$.

$$\begin{aligned} \mu^{(i)} &= \mu_Y + Y^T K_{yx}(X, X)^{-1} K_{yx}(X, x_t^{(i)}) \quad (15) \\ {\sigma^{(i)}}^2 &= K_{yx}(x_t^{(i)}, x_t^{(i)}) \\ &\quad - K_{yx}(X, x_t^{(i)})^T K_{yx}(X, X)^{-1} K_{yx}(X, x_t^{(i)}) \quad (16) \end{aligned}$$

where μ_Y is the mean of training set. The distribution of poses at time t is thus approximated by a Gaussian mixture model. The mode of this distribution can be selected as the final pose estimate.

4. Mixture Models for Unknown View

The image feature for the specific pose can vary according to a camera view point and orientation of the person with respect to the imaging plane. In a dynamic PLSA framework, the view point factor R can be easily combined into the generative model $P(Z|X)$ that represents the image formation process.

$$P(X, Y, Z, R|\theta_x) = P(X|\theta_x)P(Y|X)P(Z|X, R)P(R). \quad (17)$$

While the continuous representation of R is possible, learning such a representation from a finite set of view samples may be infeasible in practice. As an alternative, we use a quantized set of view points and suggest a mixture model,

$$P(Z|X, \beta_z, \gamma_r) = \sum_{r=1}^{S} P(Z|X, R = r, \beta_z^r, \gamma_z^r)P(R = r) \quad (18)$$

where S denotes the number of views. Note that all the kernel parameters (β_z^r, γ_z^r) can be potentially different for different r.

4.1. Learning

Collecting enough training data for a large set of view points can be a tedious task. Instead, by using realistic synthetic data generated by 3D rendering software which allows us to simulate the realistic humanoid model and render textured images from a desired point of view, one can build a large training set for multi-view human motion model. In this setting one can simultaneously use all views to jointly estimate a complete set of DPLSA parameters as well as the latent space X. Given the pairs of the pose and the corresponding image feature with view point, learning the complete mixture models reduces to joint optimization of

$$P(X, Y, Z_1, Z_2, ..., Z_S, R_1, ..., R_S) = P(X)P(Y|X)$$
$$\prod_s P(Z_s|X, R_s = s)P(R_s = s). \quad (19)$$

where S is the number of quantized views. The optimization of X and model parameters is a straightforward generalization of the method described in Sec. 3.2.

4.2. Inference and Tracking

The presence of an unknown viewing direction during tracking necessitates its estimation in addition to that of the latent state x_t. This joint estimation of x_t and R can be accomplished by directly extending the particle tracking method of Sec. 3.3. This approach is reminiscent of [3] in that it maintains the multiple view-based manifolds representing the various mappings caused from different view points.

5. Experiments

5.1. Synthetic Data

In our first experiment we demonstrate the advantage of our DPLSA framework on the subspace selection problem. We also compare the predictive ability of DPLSA when estimating the sequence Y from the observation Z.

We generate a set of synthetic sequences using the following model: intrinsic motion X is generated with 2 periodic functions in the $\Re^{T\times 2}$ space. The sequences are then mapped to a higher dimensional space of Y (in $\Re^{T\times 7}$) through a mapping which is a linear combination of nonlinear features $x_1{}^2, x_1, x_2, x_2{}^2$. Z (in $\Re^{T\times 3}$) is finally generated by mapping Y into a non-linear lower observation space in a similar manner. Examples of the three sequences are depicted in Fig. 2. This model is reminiscent of the generative process that may reasonably model the mapping from intrinsic human motion to image appearance/features.

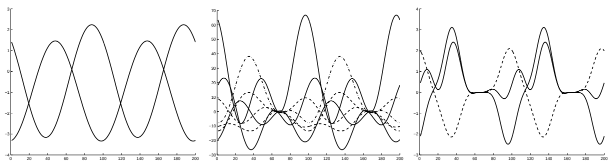

Figure 2. A example of synthetic sequences. Left: X in the intrinsic subspace. Middle: Y generated from X Right: Z generated from Y. See text for detail.

We apply three different motion modeling approaches to model the relationship between the intrinsic motion X, the 3D "pose" space Y and the "image feature" space Z. The first method (Model 1) is the manifold mapping of [3] which learns the embedding space using LLE or Isomap based on the observation Z and optimizes the mapping between X and Y using Generalized RBF interpolation. The second approach (Model 2) is the human motion modeling using Marginal Nonlinear Dynamic System (MNDS) [12], a model that attempts to closely approximate the data generation process. Model 3 is our proposed DPLSA approach described in Sec. 3.1. During the learning process, initial embedding is estimated using probabilistic PCA. Initial kernel hyperparameter values were typically set to 1, except for the dynamic models were the variances were initially assigned values of 0.01.

We evaluate predictive accuracy of models in inferring Y from Z. We generate 25 sequences using the procedure described above. We generate the testing Z by adding white noise to the training sequence and infer Y from this Z. Table 1 shows individual mean square error (MSE) rates of predicting all 7 dimensions in Y. All values are normalized with respect to the total variance in the true Y. The results demonstrate that the DPLSA model outperforms both the LLE-based model as well as the MNDS. We attribute the somewhat surprising result when compared to Model 2 to the sensitivity of this model to estimates of the initial parameters of $Z \rightarrow Y$ mapping. This problem can be mitigated by careful manual selection of the initial parameters, a typically burdensome task. However, another crucial advantage of our DPLSA model over Model 2 is the computational cost in inferring Y. For instance, the mean number of iterations of scaled CG optimization is 72.09 for Model 3 and 431.82 for Model 2. This advantage will be further exemplified in the next set of experiment.

5.2. Synthetic Human Motion Data

5.2.1. Single view point. In a controlled study we conducted experiments using a database of motion capture data for a 59 d.o.f body model from the CMU Graphics Lab Motion Capture Database [6] and synthetic video sequences. We used 5 walking sequences from 3 different subjects and 4 running sequences from 2 different subjects. The models were trained and tested on different subjects to emphasize the robustness of the approach to changes in the motion style. Initial model values, prior to learning updates, were set in the manner described in Sec. 5.1. We exclude 6 joint angles that exhibit very small variances but very noisy (*e.g.*clavicle and finger). The human figure images are rendered on Maya using the software generously provided by the authors of [20, 21]. Following this, we extract the silhouette images to compute the 10-dimensional Alt moment

Table 1. MSE rates of predicting Y from Z.

Model	$\bar{e}_{y1}$	$\bar{e}_{y2}$	$\bar{e}_{y3}$	$\bar{e}_{y4}$	$\bar{e}_{y5}$	$\bar{e}_{y6}$	$\bar{e}_{y7}$	$\sum \bar{e}_{y_i}$
LLE+GRBF	0.06	0.14	0.08	0.06	0.27	0.16	0.04	0.81
MNDS	0.02	0.06	0.06	0.06	0.13	0.10	0.04	0.47
DPLSA	0.03	0.06	0.03	0.03	0.10	0.08	0.02	**0.34**

image features as in [22]. Also 3D latent space is employed for all motion tracking experiments. Fig. 3 depicts the log precisions of $P(Y|X)$ and $P(Z|X)$ on the 2D projection of latent space learned from 1 cycle of walking sequence. Note that the precisions around the embedded X are different in the two spaces even though the X is commonly shared by both GPLVM models. To evaluate our DPLSA model in human motion tracking, we again compare it to the MNDS model in [12] that utilizes the direct mapping between the poses and the image features. The models are learned from one specific motion sequence and tested on different sequences. Fig. 4 shows the mean error in the 3D joint position estimation and the number of iterations in SCG per frame during the tracking. We use the error metric similar to the one in [17]. The error between estimated pose $\widehat{Y}$ and the ground truth pose Y from motion capture is $E(Y, \widehat{Y}) = \sum_{j=1}^{J} \parallel \widehat{y}_j - y_j \parallel / J$ where y_j is the 3D location of specific joint and J is the number of joints considered in the error matric. We choose 9 joints which have a wide motion range (*e.g.*throx and right & left wrist, humerus, femur and tibia). The height of human figure in this virtual space is 28 and the error unit can be relatively computed (*e.g.*when the height of man is $175cm$, the error unit is $175/28 \approx 6.25cm$). When the model was learned from 1 walking sequence and tested on 4 other sequences, the average error was 1.91 for MNDS tracking and 1.85 for DPLSA tracking. Results of full pose estimation for one running sequence are depicted in Fig. 5. The models achieve similarly a good accuracy in pose estimation. The distinct difference between the two models, however, is exhibited in the computational complexity of the learning and inference stages as shown in Fig. 4. The DPLSA model, on average, requires 1/5 of the iterations to achieve the same level of accuracy as the competing model. This can be explained by the presence of complex direct interaction between high dimensional features and pose states in the competing model. In DPLSA such interactions summarized via the low dimensional subspace. As a result, DPLSA representation can lead to potentially more suitable algorithm for real-time tracking.

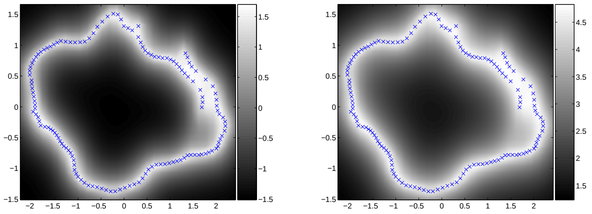

Figure 3. Latent spaces with the grayscale map of log precisions. Left: $P(Y|X)$. Right: $P(Z|X)$.

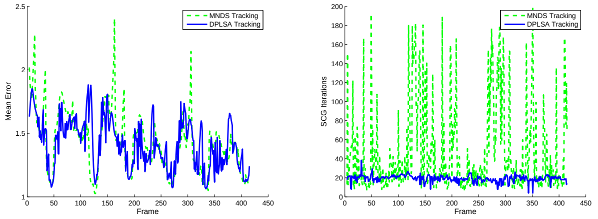

Figure 4. Tracking performance comparison. Left: pose estimation accuracy. Right: mean number of iterations of SCG.

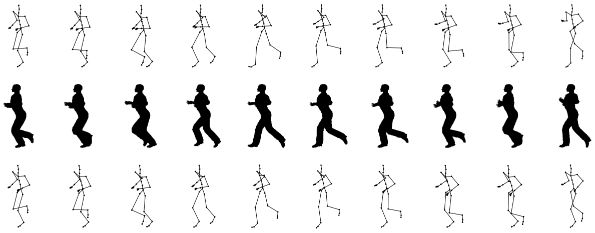

Figure 5. Input silhouettes and 3D reconstructions from a known viewpoint of $\frac{\pi}{2}$. First row: true poses. Second rows: silhouette images. Third row: estimated poses.

5.2.2. Multiple view points. We used one person sequence to learn the mixture models of the 8 different views (view angles = $\frac{\pi}{4}i, i = 1, 2, \ldots, 8$ in clockwise direction, 0 for frontal view) and human motion model. We made testing sequences by picking different motion capture sequences and rendering the images from 8 view points. Fig. 6 shows the one example of 3D tracking results from 2 different view points. In the experiment the view point of input images are unknown and inferred frame by frame during tracking with pose estimation. Although the pose estimation for some ambiguous silhouette is erroneous, our system can track the pose until the end of sequence with the proper view point estimation and 3D reconstructions are matched well to the true poses. Notice that the last two rows depict poses viewed from $3\pi/4$, *i.e.*the subject walking in the direction of the top left corner.

5.3. Real Video Sequence

We applied our method to tracking of real monocular image sequences with fixed view point. We used the sideview sequences from CMU Mobo database [7]. DPLSA model was trained on walking sequences from the Mocap data and tested on the motion sequences from the Mobo set. Fig. 7 shows two example sequences of our tracking result. The lengths of the testing sequence are 300 and 340 frames. Although the frame rates and the walking style are different and there exists noise in the silhouette images, the reconstructed pose sequence depicts a plausible walking motion that agrees with the observed images.

6. Conclusions

We have reformulated the shared latent space approach for learning models from sequential dyadic data. The reformulated generative statistical model is called the Dynamic Probabilistic Latent Semantic Analysis (DPLSA), which extends the successful PLSA formalism to a continuous state estimation problem of mapping sequences of human figure appearances in images to estimates of the 3D figure pose. Our preliminary results indicate that the DPLSA formalism can result in highly accurate trackers that exhibit fractional computational cost of the traditional subspace tracking methods. Moreover, the method is easily amenable to extensions to unknown or multi-view camera tracking tasks.

The proposed model has the potential to handle complex classes of tracking problems, such as the challenging rapidly changing motions, when coupled with multiple model frameworks such as the switching dynamic models. Our future work will address these new directions as well as focus on continuing extensive evaluation of DPLSA on additional motion datasets.

References

[1] J. Bellegarda. Exploiting both local and global constraints for multi-spanstatistical languagemodeling. In *ICASSP*, volume 2, pages 677–680, 1998.

[2] S. C. Deerwester, S. T. Dumais, T. K. Landauer, and G. W. F. andR. A. Harshman. Indexing by latent semantic analysis. *JASIST*, 41(6):391–407, 1990.

[3] A. Elgammal and C.-S. Lee. Inferring 3d body pose from silhouettes using activity manifold learning. In *CVPR*, volume 2, pages 681–688, 2004.

[4] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman. Learning object categories from google's image search. In *ICCV*, pages 1816–1823, 2005.

[5] T. Hofmann. Probabilistic latent semantic analysis. In *Proc. of Uncertainty in A.I,*, Stockhol, 1999.

[6] http://mocap.cs.cmu.edu/.

[7] http://www.hid.ri.cmu.edu/Hid/databases.html.

[8] N. D. Lawrence. Gaussian process latent variable models for visualisation of high dimensionaldata. In *NIPS*. 2004.

[9] N. D. Lawrence and A. J. Moore. Hierarchical gaussian process latent variable models. In *ICML*, pages 481–488. ACM, 2007.

[10] C.-S. Lee and A. Elgammal. Simultaneous inference of view and body pose using torus manifolds. In *ICPR*, 2006.

[11] F. Monay and D. Gatica-Perez. Plsa-based image auto-annotation: constraining the latent space. In *ACM Inter. Conf. on Multimedia*, pages 348–351, 2004.

[12] K. Moon and V. Pavlovic. Impact of dynamics on subspace embedding and tracking of sequences. In *CVPR*, pages 198–205, June 2006.

[13] R. Navaratnam, A. Fitzgibbon, and R. Cipolla. Semi-supervised joint manifold learning for multi-valued regression. In *ICCV*, 2007.

[14] P. W. Peter and S. T. Dumais. Personalized information delivery: an analysis of information filteringmethods. *Commun. ACM*, 35:51–60, December 1992.

[15] A. Shon, K. Grochow, A. Hertzmann, and R. Rao. Learning shared latent structure for image synthesis. *NIPS*, 2005.

[16] H. Sidenbladh, M. J. Black, and D. J. Fleet. Stochastic tracking of 3d human figures using 2d image motion. In *ECCV*, pages 702–718, 2000.

[17] L. Sigal, S. Bhatia, S. Roth, M. J. Black, and M. Isard. Tracking loose-limbed people. In *CVPR*, pages 421–428, 2004.

[18] J. Sivic, B. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman. Discovering object categories in image collections. In *ICCV*, volume 1, pages 370–378, 2005.

[19] C. Sminchisescu and A. Jepson. Generative modeling for continuous non-linearly embedded visual inference. In *ICML*, 2004.

[20] C. Sminchisescu, A. Kanaujia, Z. Li, and D. Metaxas. Discriminative density propagation for 3d human motion estimation. In *CVPR*, pages 390–397, 2005.

[21] C. Sminchisescu, A. Kanujia, Z. Li, and D. Metaxas. Conditional visual tracking in kernel space. In *NIPS*, 2005.

[22] T.-P. Tian, R. Li, and S. Sclaroff. Articulated pose estimation in a learned smooth space of feasible solutions. In *CVPR*, 2005.

[23] R. Urtasun, D. J. Fleet, and P. Fua. 3d people tracking with gaussian process dynamical models. In *CVPR*, pages 238–245, 2006.

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

[25] J. M. Wang, D. J. Fleet, and A. Hertzmann. Gaussian process dynamical models. In *NIPS*. 2005.

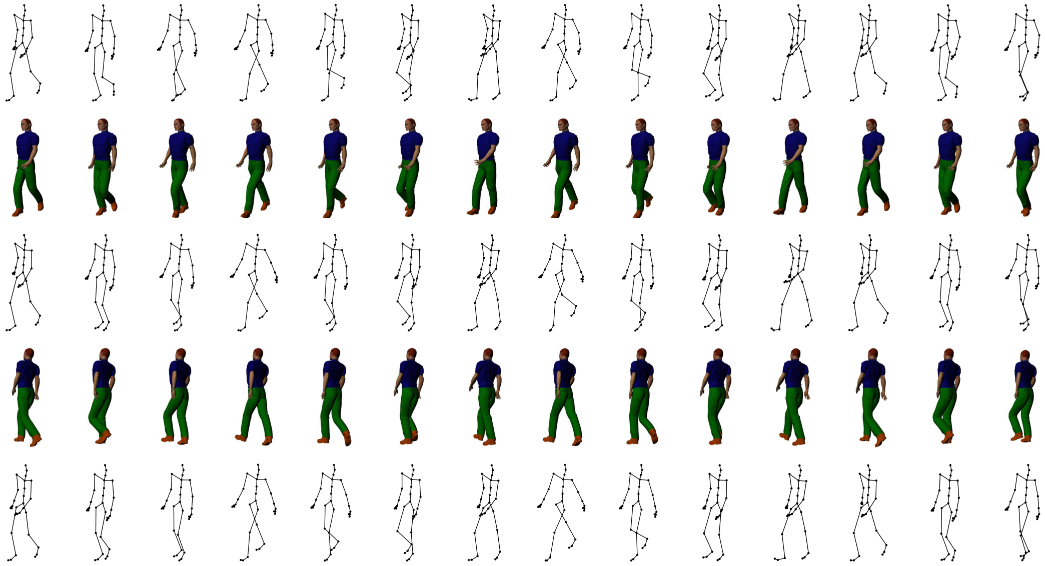

Figure 6. Input images with unknown view point and 3D reconstructions using DPLSA tracking. First row: true pose. Second and third rows: $\frac{\pi}{4}$ view angle. Fourth and fifth rows: $\frac{3\pi}{4}$ view angle.

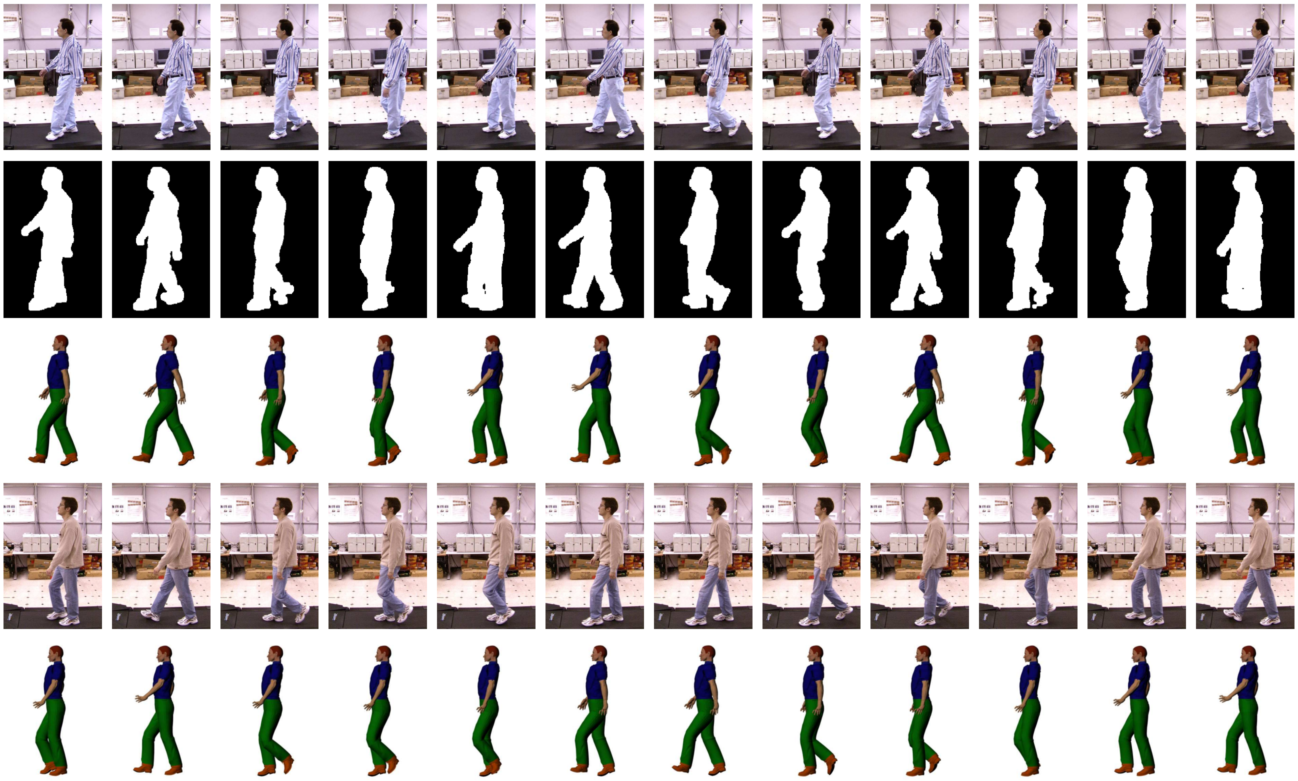

Figure 7. First: Input real walking images of subject 22. Second row: Image silhouettes. Third row: Images of the reconstructed 3D poses. Fourth row: Input real walking images of subject 15. Fifth row: Images of the reconstructed 3D poses.