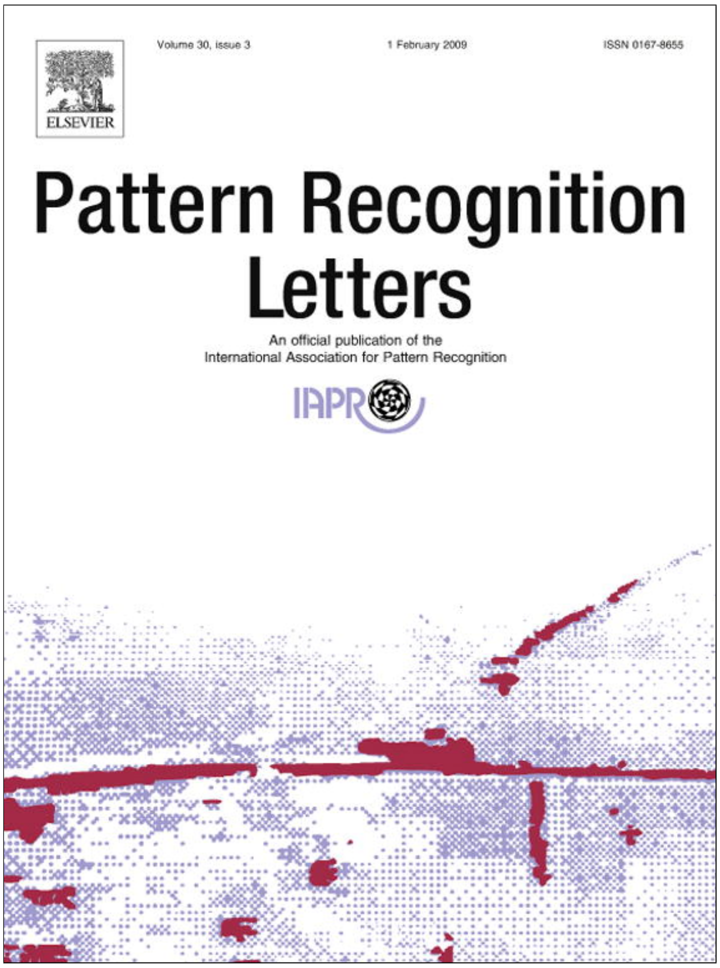

Volume 30, issue 3 1 February 2009 ISSN 0167-8655

ELSEVIER

Pattern Recognition Letters

An official publication of the
International Association for Pattern Recognition

IAPR

Pattern Recognition Letters 30 (2009) 267–274

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Tracking articulated objects by learning intrinsic structure of motion

Xinxiao Wu, Wei Liang *, Yunde Jia

Beijing Laboratory of Intelligent Information Technology, School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, PR China

ARTICLE INFO

Article history:
Received 20 November 2007
Received in revised form 22 September 2008
Available online 26 October 2008

Communicated by B. Kamgar-Parsi

Keywords:
Computer vision
Articulated objects tracking
Non-linear manifold learning

ABSTRACT

In this paper, we propose a novel dimensionality reduction method, temporal neighbor preserving embedding (TNPE), to learn the low-dimensional intrinsic motion manifold of articulated objects. The method simultaneously learns the embedding manifold and the mapping from an image feature space to an embedding space by preserving the local temporal relationship hidden in sequential data points. Then tracking is formulated as the problem of estimating the configuration of an articulated object from the learned central embedding representation. To solve this problem, we combine Bayesian mixture of experts (BME) with Gaussian mixture model (GMM) to establish a probabilistic non-linear mapping from the embedding space to the configuration space. The experimental result on articulated hand and human pose tracking shows an encouraging performance on stability and accuracy.

1. Introduction

Articulated objects tracking has become an important part in human computer interaction. Previous approaches to articulated object tracking are generally classified into two categories: 3D model-based (Stenger et al., 2001; Wu et al., 2001) and learning-based (Rosales and Sclaroff, 2006; Thayananthan et al., 2006; Campos and Murray, 2006). The 3D model-based approaches presuppose an explicit parametric model and require a good initialization. Automatic initialization and searching in a high-dimensional configuration space remain challenging problems in 3D model-based tracking. Alternatively, learning-based approaches directly estimate the configuration from an observable image via the mapping from an image feature space to a configuration space, so they could capture the articulated pose in real time. Moreover, learning-based tracking requires neither explicit 3D model nor prior labeling, which has the potential to solve the automatic initialization and re-initialization problems.

We adopt the learning-based framework for articulated objects tracking. However, a barrier to learning-based method stems from the high-dimensional representation of motion in both image feature space and configuration space, which makes the direct mapping from image feature to configuration extremely complex. Fortunately, articulated motion is highly constrained and its intrinsic structure lies on a compact low-dimensional manifold. Attempting to solve the curse of dimensionality problem, we propose temporal neighbor preserving embedding (TNPE) to learn the embedding manifold space of time-series motion. Different from NPE (He et al., 2005), TNPE preserves the local temporal relation between data points by assuming that the sequentially adjacent points are neighbors, which makes TNPE efficient on finding neighbors and suitable for training large sequential data.

Considering the intrinsic ambiguity in recovering the high-dimensional configuration from the low-dimensional embedding representation, we combine Bayesian mixture of experts (BME) (Xu et al., 1995) with Gaussian mixture model (GMM) to organize the probabilistic multi-valued mapping from the central embedding space to the configuration space. Each expert could be considered as a "local configuration estimator" and the estimates of multiple experts are combined according to their corresponding proportions (gating) in a probabilistic mixture model for global estimation.

The framework of our approach is based on learning the mapping from the image feature space to the embedding space and the mapping from such central embedding space to the configuration space. The remainder of the paper is organized as follows. Section 2 discusses related work. The TNPE method is described in Section 3. The BME with GMM is given in Section 4. Section 5 demonstrates the experiment result. Finally, Section 6 presents conclusions and suggests future work.

2. Related work

In the learning-based framework for articulated hand tracking, Rosales and Sclaroff (2006) proposed a non-linear supervised learning framework to map image features to likely 3D hand poses. Agarwal and Triggs (2006) introduced Relevant Vector Machine

* Corresponding author. Tel./fax: +86 10 6891 4849.
E-mail addresses: wuxinxiao@bit.edu.cn (X. Wu), liangwei@bit.edu.cn (W. Liang), jiayunde@bit.edu.cn (Y. Jia).

0167-8655/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2008.09.014

regression to select sparse highly relevant training examples for recovering the 3D body pose from monocular images. These methods directly estimate the configuration of articulated object from the image feature without exploiting the certain constraints on natural motion.

To learn the high constraints and reduce the high-dimensionality, Wu and Huang (2000) employed PCA to obtain the subspace of hand configuration and predicted the hand state in a linear manifold subspace. Nevertheless, PCA is linear and inadequate to handle the non-linearity in articulated motion. Sminchisescu et al. (2005) adopted kernel PCA (KPCA) to restrict the visual inference to the low-dimensional kernel induced state space. KPCA is non-linear but does not explicitly consider the structure of manifold on which the data may possibly reside. In comparison with PCA and KPCA, TNPE shares many properties of non-linear techniques and could capture the intrinsic local manifold structure of motion. Elgammal and Lee (2004) applied LLE (Roweis and Saul, 2000) or Isomap (Tenenbaum et al., 2000) to establish a low-dimensional embedding of activity manifold for human pose estimation. Wang et al. (2003) used Isomap to obtain a compact object representation for visual tracking. Both LLE and Isomap yield mappings that are defined only on the training data and they remain unclear how to naturally evaluate the maps on novel testing data. Therefore, Elgammal and Lee (2004) and Wang et al. (2003) needed to firstly learn the embedding representation and then the mapping from the high-dimensional image feature space to the low-dimensional embedding space. Different from LLE and Isomap, TNPE simultaneously learns the intrinsic embedding representation and the mapping from the image feature space to the embedding space. Being defined on the whole data space including both training and testing data, TNPE has a good generalization on novel data points. Wang et al. (2008) proposed Gaussian Process Dynamical Models (GPDM) to simultaneously learn the embedding space and the subspace dynamic model for human tracking. The Auto-Regressive (AR) model supposed in GPDM is able to capture some special and simple motions like walking, but inadequate for non-Gaussian motions with large variety in movements. Moreover, it is time consuming to calculate kernel matrix especially in large training dataset. TNPE does not establish dynamic model in manifold space, but preserves temporal continuity between time-series data by introducing temporal neighbors, and takes advantage of such temporal information to improve computation efficiency.

Given the learned central embedding manifold, articulated tracking is formulated as inferring the configuration from its low-dimensional representation. Elgammal and Lee (2004) introduced Generalized Radial Basis Function (GRBF) to facilitate interpolation of intermediate 3D pose so as to recover the configuration from the central embedding manifold. Different from the GRBF mapping of which the output is a single estimated configuration, BME and GMM are combined to model multimodal probabilistic distribution over the output configuration with more accuracy and flexibility. Tian et al. (2005) used Gaussian Process Latent Variable Models (GPLVM) to establish the inverse mapping from the low-dimensional space to the high-dimensional pose space. GPLVM is an extension to probabilistic PCA and its output conforms to the Gaussian distribution. In real tracking, the probability distribution of estimated configuration is non-Gaussian, and so we mix multiple Gaussian modes to approximate the real non-Gaussian distribution of estimated configuration.

3. Temporal neighbor preserving embedding

Due to the curse of high-dimensionality in both image feature and configuration space, it becomes an ill posed problem to directly estimate 3D configurations of articulated objects from its visual images. However, articulated motion is highly constrained and its intrinsic structure lies on a compact low-dimensional manifold. Therefore, learning the low-dimensional manifold is a natural choice to reduce the high dimension by exploiting physical constraints hidden in motion. Motivated by the temporal coherence in time-series data, we propose the temporal neighbor preserving embedding (TNPE) to simultaneously learn the low-dimensional manifold space and the mapping from the image feature space to the embedding space.

For image sequences, the temporal coherence between frames provides useful information about the neighborhood structure and the local geometry of the manifold. We develop temporal NPE using temporal relationship to define the neighbors of each point. NPE is time consuming for computing the full distance matrix to find k-nearest neighbors, and TNPE just defines a temporal window moving at each point to get a set of sequentially ordered points as its neighbors.

Given a sequence from time 1 to T, $X = [x_1, x_2, \ldots, x_T] \in R^{D_f \times T}$, where the tth column x_t is the image vector at time t, we aim at finding a low-dimensional representation written as $Y = [y_1, y_2, \ldots, y_T] \in R^{d \times T}$, where $d \ll D_f$. The mapping from x_t to y_t is formulated as $y_t = A^T x_t$, where A is a $D_f \times d$ transformation matrix. The procedure of TNPE is summarized as follows:

1. Find the k-nearest temporal neighbors $\{x_s | t - k/2 \leqslant s \leqslant t + k/2\}$ of x_t, where s, t index input images.
2. Minimize $\sum_{t=1}^{T} |x_t - \sum_{s=1}^{k} w_{ts} x_s|^2$ to obtain the $T \times T$ weight matrix W, where $w_{ts} = 0$ if x_s is not one of the k-nearest temporal neighbors of x_t and where $\sum_{s=1}^{k} w_{ts} = 1$.
3. Compute the transformation matrix A by solving the generalized eigenvector problem:

$$XMX^T a = \lambda XX^T a, \tag{1}$$

where $M = (I - W)^T (I - W)$ and I is $T \times T$ identity matrix. Let the column vectors $a_0, \ldots, a_{d-1}$ be the solutions of Eq. (1), ordered according to their eigenvalues $\lambda_0 \leqslant \cdots \leqslant \lambda_{d-1}$. Thus the embedding is $x_t \to y_t : y_t = A^T x_t$, where y_t is d-dimension vector and $A = [a_0, a_1, \ldots, a_{d-1}]$ is $D_f \times d$ matrix.

In Step 2, the weight matrix W summarizes the local temporal relating the data points x_t to its neighbors. To preserve such relationship among the manifold data points y_t, we minimize the cost function

$$\begin{aligned} \Phi(Y) &= \sum_t \left(y_t - \sum_s w_{ts} y_s \right)^2 = Tr[Y(I - W)^T (I - W) Y^T] \\ &= Tr[A^T XMX^T A] \end{aligned} \tag{2}$$

with respect to the transformation matrix A, where $M = (I - W)^T (I - W)$.

In order to remove an arbitrary scaling factor in the projection, we enforce the following constraint:

$$\frac{1}{T} \sum_t y_t y_t^T = \frac{1}{T} YY^T = \frac{1}{T} A^T XX^T A = I_d. \tag{3}$$

Eqs. (2) and (3) are both quadratic and the optimal A is determined by solving a generalized eigenvalue problem $XMX^T a = \lambda XX^T a$. The first to the dth smallest generalized eigenvectors form the columns of A.

In Step 1, k is the size of a temporal window providing temporal history for each training data point. The time to find the k-nearest neighbors is $O(D_f N^2)$ for NPE, while it is $O(N)$ for TNPE (N is the number of training data). Therefore, TNPE reduces the computing time obviously, especially for the large training data. We compare TNPE with NPE on computational efficiency in Fig. 1. From the figure, the

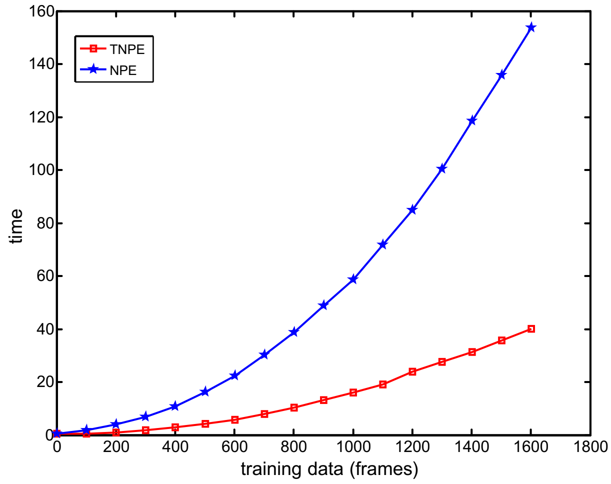

Fig. 1. The efficiency compared with TNPE and NPE.

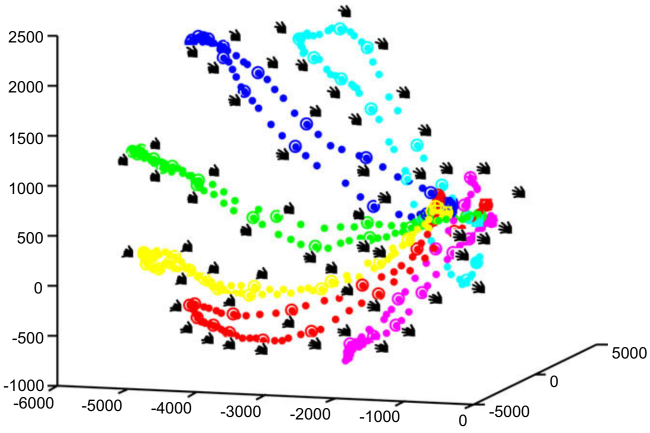

Fig. 2. Low-dimensional representation by TNPE.

time consuming of NPE (represented by pentacle line) is quadratic increasing with the number of training data, while TNPE (shown by square line) is approximately linear increasing. Fig. 2 demonstrates the low-dimensional representation of several different hand motions in which the first three dimensions are selected to be shown. The resulted intrinsic coordinates approximately lie on the different curves indicated by different colors.

4. Mapping from an embedding space to a configuration space

Given an image feature of an articulated object, we could learn its low-dimensional representation in the central embedding space via TNPE described in Section 3. In order to obtain the corresponding configuration, the mapping from the low-dimensional embedding space to the high-dimensional configuration space will be learned. Considering the intrinsic ambiguity hidden in this mapping, we introduce mixture of experts to establish a global non-linear and probabilistic multimodal mapping function. Each expert handles a linear mapping in a local region. For each expert, input is the low-dimensional embedding representation and output is the Gaussian distribution of high-dimensional configuration. According to the corresponding gatings, multiple experts are combined in mixture Gaussian model to attain the global non-linear mapping. The gating of each expert is calculated using the gating-net parameters by Bayesian rule. In our work, the parameters of gating-net are determined by the parameters of GMM which approximates the probabilistic distribution over the embedding space.

4.1. Learning the non-linear mapping by Bayesian mixture of experts

Let $y \in R^d$ be the embedding vector and $z \in R^{D_c}$ the configuration vector, the probabilistic mapping from y to z through mixture of K experts is formulated as

$$P(z|y) = \int_j p(z|y,j)p(j|y)dj = \sum_{j=1}^{K} g_j(y)p_j(z|y), \tag{4}$$

where $g_j(y) = p(j|y) \geqslant 0, \sum_{j=1}^{K} g_j(y) = 1$ and $p_j\,(z|y) = N(z|W_jy + u_j,\psi_j)$. The conditional distribution $p_j(z|y)$ represents a local probabilistic mapping function from y to z in the jth expert. The gating of the jth expert $g_j(y)$ could be considered as a "responsibility" representing how reliable the jth expert's mapping is. For the input y, $g_j(y)$ is actually the posterior probability indicating how probable y is assigned to the partition corresponding to the jth expert. The posterior probability is calculated by

$$g_j(y) = p(j|y) = \frac{\alpha_j p(y|j)}{\sum_{i=1}^{K} \alpha_i p(y|i)}, \tag{5}$$

where $\alpha_j \geqslant 0,\ \sum_{j=1}^{K} \alpha_j = 1$ and $p(y|j) = N(y|v_j, \sum_j)$.

According to Eqs. (4) and (5), the mapping function $p(z|y)$ can be rewritten as

$$p(z|y, \theta_E, \theta_G) = \sum_{j=1}^{K} N(z|W_jy + u_j, \psi_j) \frac{\alpha_j N(y|v_j, \Sigma_j)}{\sum_{i=1}^{K} \alpha_i N(y|v_i, \Sigma_i)} \tag{6}$$

where $\theta_{\mathrm{E}} = \{W_j, u_j, \psi_j | j = 1,2,\ldots,K\}$ and $\theta_G = \{\alpha_j, v_j, \sum_j | j = 1, 2, \ldots, K\}$ are, respectively, the expert mapping parameters and the expert gating-net parameters. In Sections 4.2 and 4.3, we will, respectively, discuss the estimation of θ_G and θ_E.

4.2. Approximating the embedding space probability distribution

In order to estimate the parameters $\theta_G = \{\alpha_j, v_j, \sum_j | j = 1, 2, \ldots, K\}$ of the expert gating-net, we utilize Gaussian mixture model (GMM) to approximate the probability distribution over the embedding space, by assuming that the jth component in GMM corresponds to the partition of the jth expert. The weight of the jth component in GMM is assigned to α_j; the mean vector and covariance matrix of the jth Gaussian component are, respectively, assigned to v_j and $\sum_j$.

4.3. Estimating the mapping parameters by maximum likelihood

We utilize Maximum Likelihood (ML) estimation algorithm to estimate the parameters of mixture experts $\theta_{\mathrm{E}} = \{W_j, u_j, \psi_j | j = 1,2,\ldots,K\}$ using closed-form solution instead of Expectation-Maximization (EM) for computation efficiency.

Given the training data $\{y_i, z_i\}_{i=1}^{N}$, we introduce $\Lambda_i = [\lambda_{i1}, \lambda_{i2}, \ldots, \lambda_{iK}]$ with $\sum_{j=1}^{K} \lambda_{ij} = 1$ to represent the gatings of K experts for the input data y_i. The element λ_{ij} indicates the gating of the jth expert and could be calculated by Eq. (5) given θ_G and y_i. K is the number of mixture experts and N is the number of training data. From Eq. (6), the expected log-likelihood function is given by

$$\begin{aligned}\Phi &= \log \prod_{i=1}^{N} \prod_{j=1}^{K} \Big\{ (2\pi)^{D_c/2} |\psi_j|^{-1/2} \\ &\quad \times \exp\Big\{ -\frac{1}{2} [z_i - u_j - W_j y_i]^T \psi_j^{-1} [z_i - u_j - W_j y_i] \Big\} \Big\}^{\lambda_{ij}} \\ &= C - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{K} \lambda_{ij} \log |\psi_j| - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{K} \lambda_{ij} [z_i - \overline{W_j}\overline{y_i}]^T \psi_j^{-1} [z_i - \overline{W_j}\overline{y_i}],\end{aligned} \tag{7}$$

where $\overline{W_j} = [W_j, u_j]$ and $\overline{y_i} = [y_i^T, 1]^T$.

By maximizing the objective function in Eq. (7), $\overline{W_j}$ is estimated by

$$\overline{W_j} = \left(\sum_{i=1}^{N} \lambda_{ij} z_i \overline{y_i}^T\right)\left(\sum_{i=1}^{N} \lambda_{ij} \overline{y_i}\,\overline{y_i}^T\right)^{-1}. \tag{8}$$

In our experiment, we simplify the covariance matrix ψ_j proportional to identity matrix:

$$\psi_j = \sigma_j^2 I, \quad \sigma_j^2 = \sum_{i=1}^{N} \lambda_{ij}(z_i - \overline{W_j}\,\overline{y_j})^T (z_i - \overline{W_j}\,\overline{y_j}) \Big/ \sum_{i=1}^{N} \lambda_{ij}. \tag{9}$$

5. Experimental result

In this section, we demonstrate our method on both articulated hand and human body tracking. For an image feature, we first obtain its corresponding low-dimensional representation in the embedding space via TNPE, and then estimate the configuration through BME with GMM. The tracking framework is depicted in Fig. 3.

5.1. Articulated hand tracking

5.1.1. Hand image feature and configuration modeling

In articulated hand tracking, image features are represented by silhouette images and the configuration is modeled by joint angles from CyberGlove. Firstly, foreground regions are extracted from background and then the hand silhouette is obtained by removing noise and filling in the regions. To simplify the computation, the hand silhouette is normalized to a size of 60 × 70 pixels, and the feature vector dimensionality is 4200. The dimensionality of the configuration vector is 14 represented by joint angles obtained from 5DT CyberGlove. Each finger is modeled as a planar mechanism with 2 DOF (degrees of freedom) for flexion $(\theta^x_{MP}, \theta_{PIP})$ and 1 DOF for abduction and adduction (θ^Z_{MP}) with palm. The abduction and adduction DOF of the middle finger (θ^Z_{MP}) is set zero and ignored because the middle finger does not naturally make side movements. The joint angle θ_{DIP} is also ignored for a dependency with θ_{PIP} defined by $\theta_{DIP} = (2/3)\theta_{PIP}$ (Lee and Kunii, 1995).

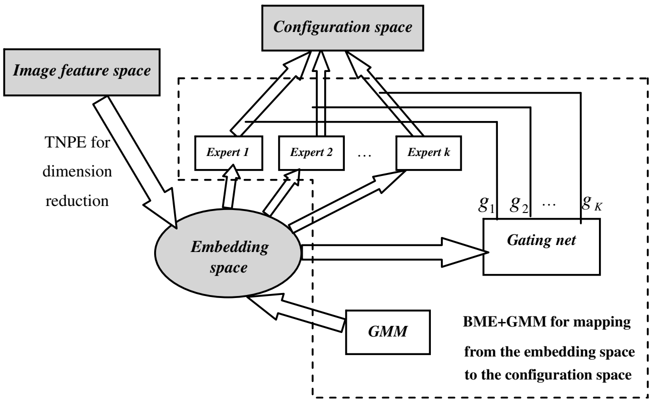

Fig. 3. The tracking framework by manifold learning.

Image sequences from a monocular camera with the corresponding joint angles from CyberGlove construct the training and testing data. We collect nine motion sequences totaling of 900 frames with the corresponding joint angle vectors as training data and another nine sequences of 860 frames with joint angle vectors for testing. Since 5DT CyberGlove only captures local movement that is represented by 14 joint angles, our experiment limits in the local motion of articulated hand.

5.1.2. Result on comparisons and tracking

Fig. 4 demonstrates the comparison on the tracking error of the proposed TNPE–BME method with KPCA–BME, ISOMAP–GRBF and GPDM–BME. Different from the RMS errors (in degrees) used by Agarwal and Triggs (2006), we report the mean (over all angles) tracking error as

$$E(Z', Z) = \frac{1}{m}\sum_{i=1}^{m} \frac{|(z'_i - z_i) \bmod \pm 180°|}{range(z_i)}, \tag{10}$$

where $Z' = [z'_1, z'_2, \ldots, z'_m]^T$ is the estimated joint angle vector and $Z = [z_1, z_2, \ldots, z_m]^T$ is the ground truth from CyberGlove. The $range(z_i)$ represents the range of variation (in the testing data) for the joint angle z_i, i.e. $range(z_i) = |(\max(z_i) - \min(z_i)) \bmod \pm 180°|$. The RMS error varies depending on the range of joint angle, and we use the ratio of the absolute error to the range for an invariant and consistent error measure. Different from KPCA, TNPE can capture the intrinsic motion structure considering temporal continuity between time-ordered poses. In comparison with ISOMAP, TNPE has good generalization on novel data point not included in training dataset and is able to perform better on computing efficiency by taking advantage of temporal neighborhood. GPDM is able to discover non-linear latent embedding space of motion, but it is time consuming for them to calculate kernel matrix especially in large training dataset, while TNPE computes efficiently avoiding calculating distance and kernel matrixes. Table 1 concludes comparison results between TNPE and other related methods on articulated hand tracking. The second column indicates the execution time on the same training dataset (MatlabR2006b code, Dell Pentium D, 3.4 GHz PC), and the

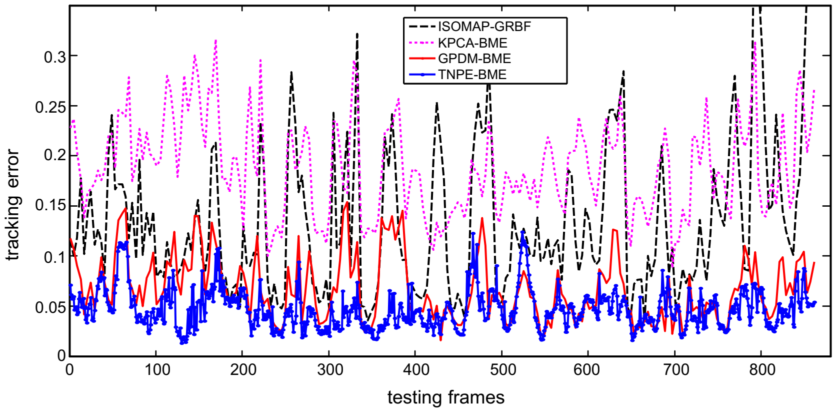

Fig. 4. Comparison on tracking errors of articulated hand over the whole testing data.

Table 1
Comparison results on articulated hand tracking between TNPE and other methods.

Method	Approximate total training time	Mean tracking error	Considering temporal continuity (Yes/No)	Learning mapping function (Yes/No)
KPCA	6 s	9.4°	No	Yes
ISOMAP	54 s	6.8°	No	No
GPDM	83 min	2.8°	Yes	No
TNPE	7 s	2.5°	Yes	Yes

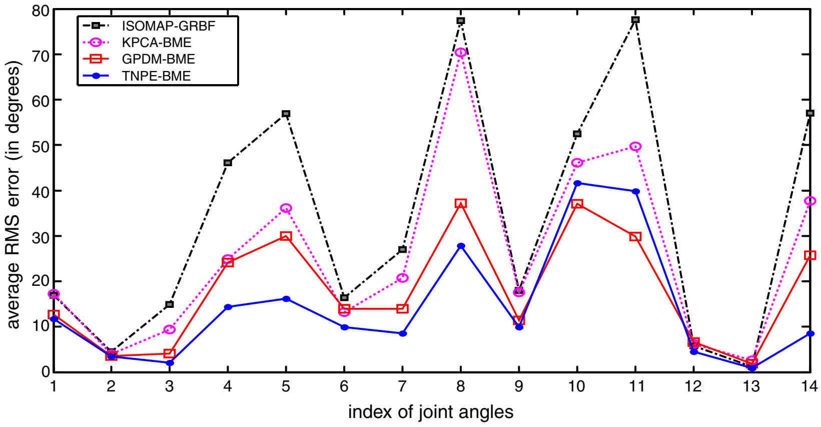

Fig. 5. Articulated hand tracking errors on individual joint angles.

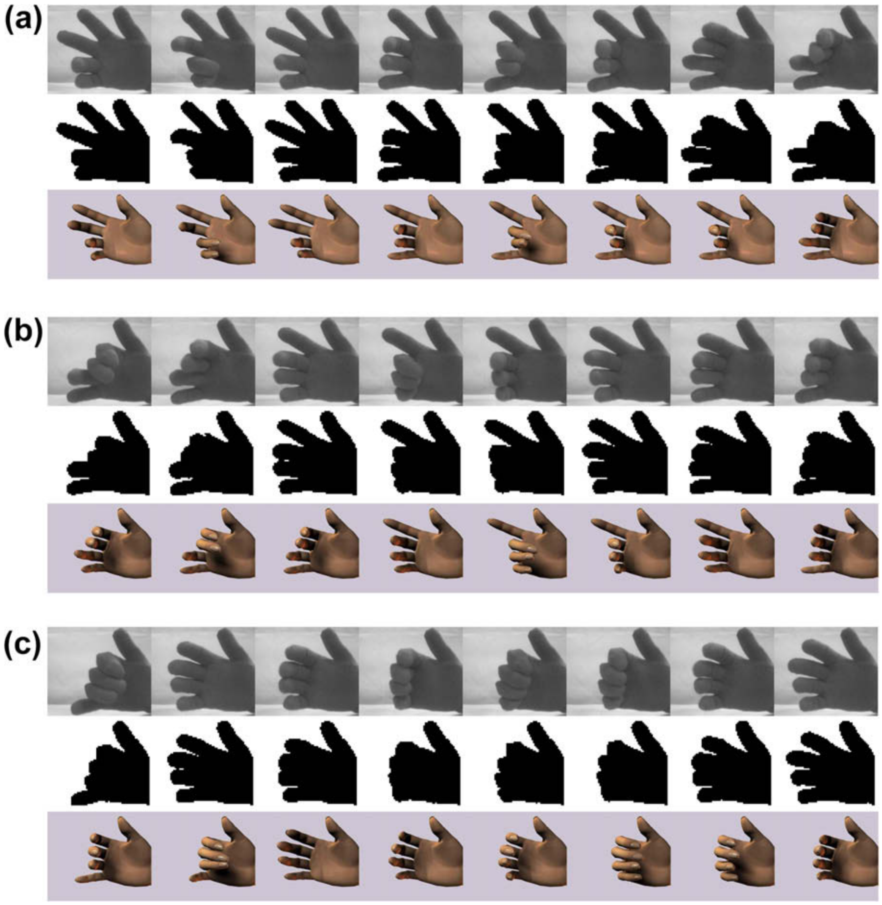

Fig. 6. Articulated hand tracking results.

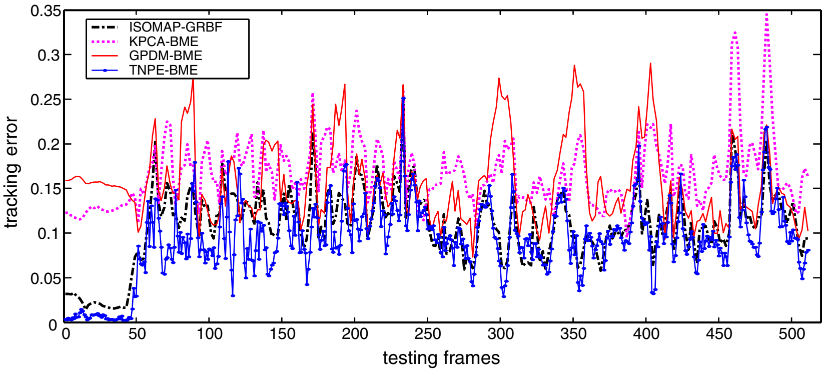

Fig. 7. Comparison on tracking errors of articulated human body over the whole testing data.

Table 2
Comparison results on articulated human body tracking between TNPE and other methods.

Method	Approximate total training time	Mean tracking error	Considering temporal continuity (Yes/No)	Learning mapping function (Yes/No)
KPCA	1 s	17.3°	No	Yes
ISOMAP	12 s	10.9°	No	No
GPDM	20 min	15.1°	Yes	No
TNPE	2 s	9.3°	Yes	Yes

third column indicates the mean tracking error over all the testing frames (in degrees).

The quantitative comparison on the average error of joint angles over the whole testing data is shown in Fig. 5, in which the horizontal axis indexes joint angles and the vertical axis indicates average errors of joint angles. The average error of individual joint angle is calculated by

$$Avg.E(z',z) = \frac{1}{n}\sum_{i=1}^{n}|(z'_i - z_i)\bmod \pm 180^\circ|, \qquad (11)$$

where n is the number of testing data. From Fig. 5, we can learn that TNPE–BME performances better with lower estimation error. However, in all these methods the estimation of θ_{PIP} is less accurate than that of other angles owing to occlusion. This occlusion problem can be alleviated by exploiting more accurate feature descriptor, increasing the training data or introducing multiple camera tracking.

In Fig. 6, we reconstruct the estimated hand configuration by Curious Lab Poser 6. There are three sequences. For each, the top row shows input images, the middle row shows hand silhouettes and the bottom row is reconstructed configuration estimated by TNPE–BME.

5.2. Articulated human body tracking

5.2.1. Human body image feature and configuration modeling

In the second experiment, we test our proposed method on tracking human body in ballet motion. The human body images with the corresponding joint angles are synthesized. The human silhouette is obtained similar to hand and normalized to a size of 49 × 50 pixels and the dimensionality of the feature vector is 2450. The dimensionality of configuration vector is 42 represented by 3 (X-rotation, Y-rotation, Z-rotation) angles for each of 14 joints. We collect totaling of 520 frames with the corresponding joint

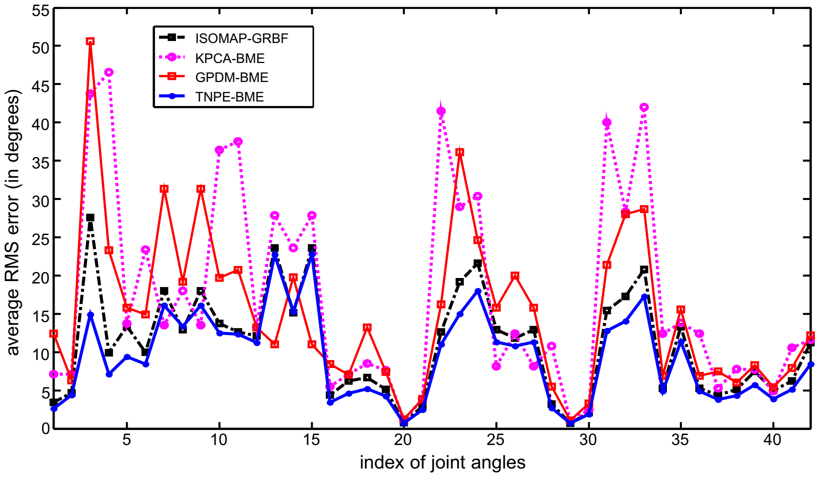

Fig. 8. Articulated human body tracking errors on individual joint angles.

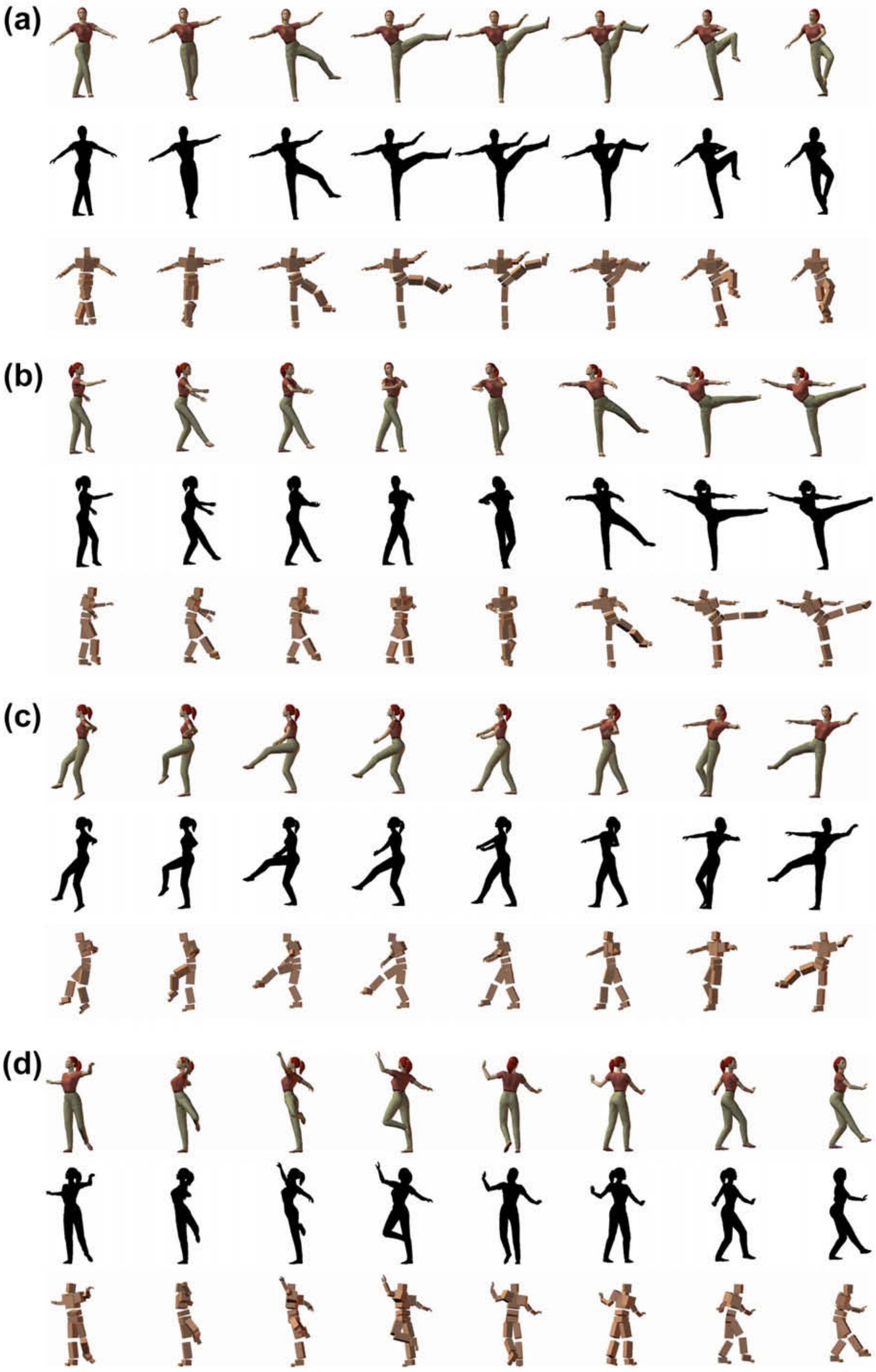

Fig. 9. Articulated human body tracking results.

angle vectors as training data and another different 511 frames with joint angle vectors for testing.

5.2.2. Result on comparisons and tracking

Fig. 7 demonstrates the comparison on the human body tracking error of TNPE–BME with KPCA–BME, ISOMAP–GRBF and GPDM–BME. In this experiment, GPDM has some problems because the subspace AR model is inadequate for the complex dynamics and large motion range hidden in ballet movements. Similar to Table 1, we compare the proposed method with other related methods on several aspects in Table 2. The quantitative comparison on the average error of joint angles over the whole testing data is shown in Fig. 8. In Fig. 9, we reconstruct the estimated configuration by Curious Lab Poser 6. There are four sequences of images. For each, the top row shows input images, the middle row shows the body silhouettes and the bottom row is the reconstruction of results estimated by TNPE–BME.

6. Conclusions and future work

We have presented a learning-based framework, integrated with non-linear motion manifold learning for articulated objects tracking. The framework is based on the mapping from the image feature space to the embedding space and the mapping from the embedding space to the configuration space. A novel dimensionality reduction method, TNPE, is proposed to learn the embedding non-linear manifold of time-series motion. BME and GMM are combined to model a non-linear, probabilistic and multi-valued mapping from the low-dimensional manifold space to the high-dimensional configuration space. The experiments on both articulated hand and human body tracking demonstrate the good performance on tracking accuracy. In the future, we intend to introduce dynamic process to smooth the motion and work at 3D human pose estimation in cluttered images.

Acknowledgements

This work was partially supported by the Natural Science Foundation of China (60675021), the Chinese High-Tech Program (2006AA01Z120), and Beijing key discipline program.

References

Agarwal, A., Triggs, B., 2006. Recovering 3D human pose from monocular images. IEEE Trans. Pattern Anal. Machine Intell. 28 (1), 44–58.

Campos, T.E., Murray, D.W., 2006. Regression-based hand pose estimation from multiple cameras. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, New York.

Elgammal, A., Lee, C.S., 2004. Inferring 3D body pose from silhouettes using activity manifold learning. In: Proc. IEEE Internat. Conf. on Computer Vision and Pattern Recognition, USA.

He, X.F., Cai, D.S., Yan, C., Zhang, H.J., 2005. Neighborhood preserving embedding. In: Proc. IEEE Internat. Conf. on Computer Vision, Beijing.

Lee, J., Kunii, T.L., 1995. Model-based analysis of hand posture. IEEE Comput. Graphics Appl., 77–86.

Rosales, R., Sclaroff, S., 2006. Combining generative and discriminative models in a framework for articulated pose estimation. Internat. J. Comput. Vision 67 (3), 251–276.

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

Sminchisescu, Cristian, Kanujia, Atul, Li, Zhiguo, Metaxas, Dimitris, 2005. Conditional visual tracking in kernel space. In: Advances in Neural Information Processing Systems.

Stenger, B., Mendonca, P.R.S., Cipolla R., 2001. Model-based 3D tracking of an articulated hand. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, Hawaii.

Tenenbaum, J.B., Silva, V., Langford, J.C., 2000. A global geometric framework for nonlinear dimensionality reduction. Science 290 (5500), 2319–2323.

Thayananthan, A., Navaratnam, R., Stenger, B., Torr, P.H.S., Cipolla, R., 2006. Multivariate relevance vector machines for tracking. In: European Conf. on Computer Vision, Austria.

Tian, T.P., Li, R., Sclaroff, S., 2005. Tracking Human Body Pose on a Learned Smooth Space. Computer Science Department, Boston University, Boston, MA.

Wang, Q., Xu, G., Ai, H., 2003. Learning object intrinsic structure for robust visual tracking. In: Proc. IEEE Internat. Conf. on Computer Vision and Pattern Recognition, Madison.

Wang, J.M., Fleet, D.J., Hertzmann, A., 2008. Gaussian process dynamical models for human motion. IEEE Trans. Pattern Anal. Machine Intell. 30 (2), 282–298.

Wu, Y., Huang, T.S., 2000. View-independent recognition of hand postures. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, South Carolina.

Wu, Y., Lin, J., Huang, T.S., 2001. Capturing natural hand articulation. In: Proc. IEEE Internat. Conf. on Computer Vision, Canada.

Xu, L., Jordan, M.I., Hinton, G.E., 1995. An alternative model for mixtures of experts. Advances in Neural Information Processing Systems, vol. 7. MIT Press, pp. 633–640.