Capturing 3D Human Motion from Monocular Images Using Orthogonal Locality Preserving Projection

Xu Zhao and Yuncai Liu

Shanghai Jiao Tong University, Shanghai 200240, China

Abstract. In this paper, we present an Orthogonal Locality Preserving Projection based (OLPP) approach to capture three-dimensional human motion from monocular images. From the motion capture data residing in high dimension space of human activities, we extract the motion base space in which human pose can be described essentially and concisely by more controllable way. This is actually a dimensionality reduction process completed in the framework of OLPP. And then, the structure of this space corresponding to special activity such as walking motion is explored with data clustering. Pose recovering is performed in the generative framework. For the single image, Gaussian mixture model is used to generate candidates of the 3D pose. The shape context is the common descriptor of image silhouette feature and synthetical feature of human model. We get the shortlist of 3D poses by measuring the shape contexts matching cost between image features and the synthetical features. In tracking situation, an AR model trained by the example sequence produces almost accurate pose predictions. Experiments demonstrate that the proposed approach works well.

Keywords: OLPP, Human Motion Analysis, Monocular Images.

1 Introduction

Capturing 3D human motion from 2D images is a significant problem in computer vision. For many image understanding applications, the 3D configurations of people in images provide usable semantic information about human activity. This is a challenging problem suffering from the obstacles conduced mainly by the complicated nature of 3D human motion and the information loss of 2D images. There are two main state-of-art approaches to deal with this problem [3]. *Discriminative methods* try to find the direct mapping from image feature space to pose state space by learning the mapping models from the training examples. This approach can supplies effective solution schemes for pose recovering problem if some additional issues can be well solved. However, the inherent one-more mapping from 2D image to 3D pose is difficult to learn accurately because the conditional state distributions are multimodal. The quantity and quality of training samples are also key factors, which can lead to some intractable problems to deal with. *Generative methods* follow the prediction-match-update philosophy. In the prediction step, the pose candidates are generated from the state prior distribution. The followed match step

V.G. Duffy (Ed.): Digital Human Modeling, HCII 2007, LNCS 4561, pp. 304–313, 2007.

evaluates the pose-image similarity with some measurement. Finally, the optimal solution is found by the state update operation. Such approach has sound probabilistic support framework but generally computationally expensive because of the complex search over the high dimension state space. Moreover, prediction model and initialization are the bottlenecks of generative method especially for the tracking situation.

In this paper, we present a novel generative approach, by which we try to widen the bottlenecks mentioned above with lower computing expense. We represent the human poses by a 3D body model explicitly, whose configurations are expressed by the joint degrees of freedom (DOFs) of body parts. In our body model, there are more than fifty full body DOFs. This is a very large state space to search for the correct poses matching with the given images. Hence, the state space should be cut in order to avoid absurd poses. In general, the reasonable pose datum pool in some compact subspace of the full state space. We extract the subspace with the OLPP of motion capture data. In this concise subspace, there are some advantageous characteristics for pose estimation, which will be introduced detailedly in the followed sections. Based on the consistency of human motion, the structure of this subspace is explored with data clustering and thus we can divide the whole motion into several typical phases represented by the cluster centers. States prediction is a common difficulty of complicated non-linear problems for the absence of effective prediction model. We choose the Gaussian mixture model as state prediction model because this model can well approximates the multimodal pose distribution with the outcomes of data clustering. By the efficient shape contexts [7] matching, we evaluate the pose predictions and finally recover the 3D human pose. In the tracking situation, an autoregressive process guides the state prediction.

2 Related Work

There has been considerable prior work on capturing human motion [1-3]. However, this problem still hangs in doubt because it's ill conditioned in nature. For knowing how the human 3D pose is configured, more information are required than images can provide. Therefore, much work focus on using prior knowledge and experiential data. Explicit body model embody the most important prior knowledge about human pose and thus are widely used in human motion analysis [1]. Another class of important prior knowledge comes from the motion capture data. The combination of the both prior information causes favorable techniques for solving this problem.

Agarwal and Triggs [6] distill prior information of human motion from the hand-labeled training sequences using PCA and clustering on the base of a simple 2D human body model. This method presents a good tracking scheme but has no description about pose initialization.

Urtasun et al. [8,9] construct a differentiable objective function based on the PCA of motion capture data and then find the poses of all frames simultaneous by optimizing the function. Sidenbladh et al. [10,11] present the similar method in the framework of stochastic optimization. For a specific activity, such methods need many example sequences for computing the PCA and all of these sequences must keep same length and same phase by interpolating and aligning. Huazhong Ning et al.

[14] learn a motion model from semi-automatically acquired training examples that are aligned with correlation function. Unlike these methods, we extract the motion base space from only one example sequence of a specific activity using the lengthways OLPP and thus have no use for interpolating or aligning.

The methods mentioned above utilize the prior information in generative fashion. By contrast, discriminative approach [15-17] makes use of prior information by directly learning pose from image measurements. In [15], Agarwal and Triggs present several regression based mapping operators using shape context descriptor. Sminchisescu et al. [16] learn a multimodal state distribution from the training pairs based on the conditional Bayesian mixture of experts models. These methods can bring the interest of fast state inference after finishing the training. However, they are prone to fail when the small training database are used.

The styles of using prior information are multiform. Mori et al. [12] contain the prior information in the stored 2D image exemplars, on which the locations of the body joints are marked manually. By the shape contexts matching with the stored exemplars, the joint positions of the input images are estimated. And then, the 3D poses are reconstructed by the Taylor method [18].

Comparing with these methods, extracting the common characteristic of a special motion type from prior information is of particular interest to us. At the same time, we ensure the motion individuality of the input sequences in the generative framework with a low computational expense based on the efficient analysis of prior information.

3 OLPP-Based State Space Analysis

In this study, we represent the 3D configurations of human body as the joint angles vectors of body model. These vectors reside somewhere in the state space. The potential special interests motivate us to analyze the characteristics and structure of this space. Such interests involve mainly modeling the human activities effectively in the extracted base space and eliminating the curse of dimension. The state space analysis is performed on the base of OLPP.

3.1 Orthogonal Locality Preserving Projection (OLPP)

Orthogonal Locality Preserving Projection (OLPP) is a novel subspace learning algorithm presented by Deng Cai et al. [4]. This algorithm is built on the base of the Locality Preserving Projection (LPP) method [5] and primitively devised to solve the problem of face recognition. Actually, OLPP is an effective dimensionality reduction method falling into the category of manifold learning.

Considering the problem of representing all of the vectors in a set of n D-dimensional samples $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n$ by n d-dimensional vectors $\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n$, respectively, $D > d$. The objective function of LPP [5] is as follows:

$$\min \sum_{ij} \| \mathbf{y}_i - \mathbf{y}_j \|^2 S_{ij} \tag{1}$$

where S is a similarity matrix. A possible way of defining S is as follows:

$$S_{ij} = \begin{cases} \exp(-\| \mathbf{x}_i - \mathbf{x}_j \|^2 / t), & \| \mathbf{x}_i - \mathbf{x}_j \|^2 < \varepsilon \\ 0, & otherwise \end{cases} \tag{2}$$

where ε is sufficiently small, and $\varepsilon > 0$. Here ε defines the radius of the local neighborhood. In other words, ε defines the locality. Therefore, minimizing the objective function is an attempt to ensure that, if $\mathbf{x}_i$ and $\mathbf{x}_j$ are close, then $\mathbf{y}_i$ and $\mathbf{y}_j$ are close as well. Finally, the basis functions of LPP are the eigenvectors associated with the smallest eigenvalues of the following generalized eigen-problem:

$$XLX^T \mathbf{w} = \lambda XDX^T \mathbf{w} \tag{3}$$

where $X = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]$ and D is a diagonal matrix ; $D_{ii} = \sum_j S_{ji}$. $L = D - S$ is the Laplacian matrix and $\mathbf{w}$ is the transformation vector. The algorithmic procedure of OLPP is stated below.

1. *PCA projection:* projecting the high dimensionality points $\mathbf{x}_i$ into the PCA subspace by throwing away the components corresponding to zero eigenvalue. The transformation matrix of PCA is W_{PCA}.
2. *Constructing the adjacency graph:* Let G denote a graph with n nodes. The i-th node corresponds to $\mathbf{x}_i$. Putting an edge between nodes i and j if $\mathbf{x}_i$ and $\mathbf{x}_j$ are close, i.e. $\mathbf{x}_i$ is among p nearest neighbors of $\mathbf{x}_j$ or $\mathbf{x}_j$ is among p nearest neighbors of $\mathbf{x}_i$.
3. *Choosing the weights:* according to equation (2).
4. *Computing the orthogonal basis functions:* Let $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_k$ be the orthogonal basis vectors, defining:

$$A^{(k-1)} = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{k-1}] \tag{4}$$

$$B^{(k-1)} = [A^{(k-1)}]^T (XDX^T)^{-1} A^{(k-1)} \tag{5}$$

The orthogonal basis vectors $\mathbf{w}_1, \mathbf{w}_2, \ldots \mathbf{w}_k$ can be computed as follow.

- Compute $\mathbf{w}_1$ as the eigenvector of $(XDX^T)^{-1} XLX^T$ associated with the smallest eigen-value.
- Compute $\mathbf{w}_k$ as the eigenvector of

$$M^{(k)} = \left\{ I - (XDX^T)^{-1} A^{(k-1)} \left[B^{(k-1)} \right]^{-1} \left[A^{(k-1)} \right]^T \right\} \cdot (XDX^T)^{-1} XLX^T \tag{6}$$

associated with the smallest eigenvalue of $M^{(k)}$.

5. *OLPP Embedding:* Let $W_{OLPP} = [\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_l]$, the embedding is as follows.

$$\mathbf{x} \rightarrow \mathbf{y} = W^T \mathbf{x}, \quad W = W_{PCA} W_{OLPP} \tag{7}$$

In this paper, utilizing the orthogonality of the base functions, we reduce the dimensionality of state space and then reconstruct the data.

3.2 Pose Representation

We represent the human pose using the explicit body model. Our fundamental 3D skeleton model (see Figure. 1a) is composed of 34 articulated rigid sticks. There are 58 pose parameters in our model, including 55 joint angles of body parts and 3 global rotation angles. Therefore, each body pose can be viewed as a point in the 58D state space.

Figure. 1b show the 3D convolution surface [19] human model, which actually is an isosurface in a scalar field, defined by convolving the 3D body skeleton with a kernel function. Similarly, the 2D convolution curves of human body as shown in Figure. 1c are the isocurves generated by convolving the 2D projection skeleton. As the synthetical model features, the curves will match with the image silhouettes.

Fig. 1. (a) The 3D human skeleton model. (b) The 3D human convolution surface model. (c) The 2D convolution curves.

3.3 Extracting the Base Space

All of the human poses distribute in the 58D state space. The poses belong to a special activity, such as walking, running, handshaking, etc., generally crowd in a subspace of the full state space. We extract this subspace from the motion capture data obtained from the CMU database [20].

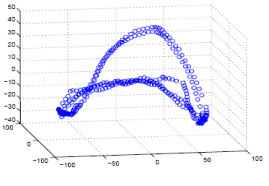

Fig. 2. The manifold of the walking sequences in 3D base space

According to the OLPP, any pose vector (t is the time tag) in a training sequence can be expressed as:

$$\mathbf{x}_t = \mathbf{y}_t W^T \tag{8}$$

where $\mathbf{x}_t$ and $\mathbf{y}_t \in \mathbb{R}^q$ are column vectors and W^T is the transformation matrix. $p(< q)$ is the dimension of the base space. Here, we take $p = 5$ \$p=5\$, which means that recovering the human pose in the 5D base space only lose negligible information. In this way, we extract the base space covering a special human activity from a single training sequence. Actually, the training sequences belonging to a same motion type but performed by different subjects can produce similar outcomes. For example, our experiments demonstrate that the walking training sequences generate the similar manifold in the 3D base space as shown in Figure. 2. Thus, by extracting the base space, we represent the pose as a 5D vector in base space.

The interests of extracting the base space include not only the dimension reduction but also the advantages for analyzing. We have known that the special human motion type shows the special manifold in the base space. Essentially, this manifold indicates the common identity of the motion type. Therefore, our focus is transferred from the base space to the more local part: special manifolds, which actually are the point set presenting special geometry shape in the base space. We analyze the manifolds with the k-means clustering. Based on the activity continuity, the set of $\mathbf{y}_t$ can be partitioned into different connected subsets and every subset represents a special motion phase. Here, we choose the number of clustering as 4. Every clustering center is the key-frame of the motion sequence. Figure. 3 shows the 3D human poses corresponding to the clustering centers. In the followed tracking process, the clustering outcomes are used in the pose prediction model.

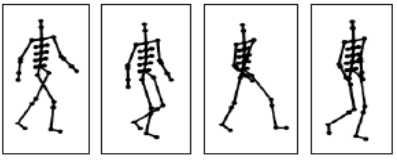

Fig. 3. The pose key frames in walking sequence

4 Image Matching Likelihood

In generative framework, pose recovering be formulated as a Bayesian posterior distribution inference:

$$p(\mathbf{y} \mid \mathbf{o}) \propto p(\mathbf{y}) p(\mathbf{o} \mid \mathbf{y}) \tag{9}$$

where $\mathbf{o}$ represents the image observations. The likelihood function $p(\mathbf{o} \mid \mathbf{y})$ is used for evaluating every pose candidate generated by the prediction models.

We choose the image silhouettes as the observed image feature as shown in Figure.4.

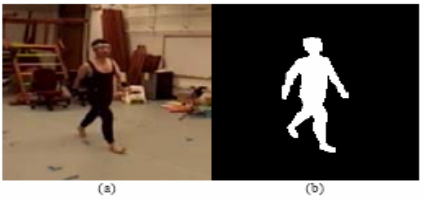

Fig. 4. (a) Original image. (b) Image silhouette.

We describe the image silhouettes and the convolution curves using shape context descriptor [7], a robust and discriminative shape descriptor. The likelihood function is constructed by the shape contexts matching [13]. In the matching process, we first sample the edge points of the image silhouettes as the query shape. Next, the point set sampled from the convolution curves are as the known shapes. Before matching, the image shape and the candidate shape are normalized to same scale. We denote the image shape as $\mathbf{S}_{query}$ and the candidate pose shape as $\mathbf{S}_i$. The matching cost can be formulated as:

$$C_v(\mathbf{S}_{query}, \mathbf{S}_i) = \sum_{j=1}^{r} \chi^2(SC^j_{query}, SC^*_i) \tag{10}$$

where SC is the shape context, r is the number of sample point in image shape, and $SC^*_i = \arg\min_u \chi^2(SC^j_{query}, SC^u_i)$. Here, we use the χ^2 distance as the similarity measurement.

5 Tracking

For image sequences, tracking means that the pose estimation in current time step depends on the outputs of previous estimation. Thus, the most important part of tracking is dynamical model that indicates how the state evolves with time. Another intractable problem in tracking is the state initialization. In this section, we deal with the problems of tracking based on the outcomes of state space analyzing in generative framework.

5.1 Initialization

Initialization is the first step of tracking, aiming for finding the correct pose of the first frame in a given image sequence. We present an auto-initialization scheme based on the Gaussian mixture model. In the base space depicted in section 3, a pose $\mathbf{y}$ can be viewed as a 5D random vector that is generated from a multimodal distribution. This distribution can be formulated as:

$$p(y) = \sum_{i=1}^{c} \omega_i \cdot N(\mathbf{y}; \mathbf{y}_{ci}, \Sigma_i) \tag{11}$$

where, $c = 4$ is the number of pose clustering, $\{\omega_i : i = 1,2,3,4. \sum_{i=1}^{4} \omega_i = 1\}$ are the weights of single Gaussian distributions and $\{\sum_i : i = 1,2,3,4\}$ are the variances of these distributions which can be computed from the training sequence.

The procedure of *auto-initialization* is performed as follows:

1. Estimating of the global rotations by:
- Partitioning the global angle scopes into 8 bins (Relying on the robustness of the matching method, 8 bins is enough.).
- Generating N samples from each single Gaussian distribution in every bin. (In our experiments, N=3.)
- Performing shape contexts matching between the query image and convolution curves produced by the sample poses.
- Evaluating the bins according to the matching score. The bin containing the minimum cost score wins. By the way, recording the matching scores of every sample pose.
2. Determining the pose in the query image.
- Generating pose samples from the Gaussian mixture distribution as formulated in the Equation (11). The weights are determined as follows: (1) Taking out the minimum matching score of each Gaussian distribution from the winning bin (see step 1). (2) Obtaining the weights by normalizing the matching scores to [0,1].
- Evaluating these pose samples and determining the pose shortlist in which there are n samples with minimum matching scores.
- The final pose is the weighted sum of poses in shortlist.

5.2 Dynamical Model

Because of the complicated nature of human motion, it's difficult to obtain an analytical physical model for it. We prefer to seek the statistical dynamical model of human motion from the training data. Similar to the model introduced in [6], we learn a second order Auto-Regressive Process (ARP) for the time domain prediction of pose in the base space. In tracking situation, the probability distribution of pose in time t can be formulated as:

$$p(\mathbf{y}_t \mid \mathbf{o}) \propto p(\mathbf{o} \mid \mathbf{y}_t) p(\mathbf{y}_t \mid \mathbf{Y}_{t-1}) \tag{12}$$

in our model, $\mathbf{Y}_{t-1} = \{\mathbf{y}_{t-1}, \mathbf{y}_{t-2}\}$. And the prediction distribution $p(\mathbf{y}_t \mid \mathbf{Y}_{t-1})$ is modeled by the second order ARP:

$$\mathbf{y}_t = \mathbf{M}_1 \mathbf{y}_{t-1} + \mathbf{M}_2 \mathbf{y}_{t-2} + v_t \tag{13}$$

where the fifth order matrices $\mathbf{M}_{1,2}$ and the variances of Gaussian white noise v_t are learnt from the training sequences. These parameters are corrected in the process of pose recovering according to the estimated outcomes. Guided by the dynamical model, we find the correct poses using particle filter. The computational expense of our method is low because the special manifold that represents the common

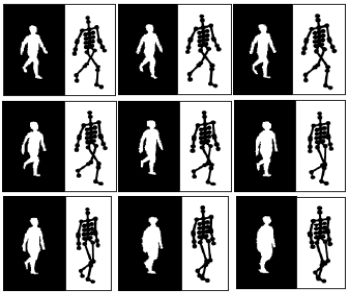

Fig. 5. The tracking results

characteristic of special motion type in the base space lead to the accurate dynamical model and therefore tracking can be proceeded with few particles. The results of tracking are shown in Figure.5. Experiments demonstrate that our method works well.

6 Conclusion

We have introduced a novel approach to tracking 3D human motion. This method extracts the compact base space from motion capture data that contain the prior information about human motion. Actually, in so doing, we extract the nature of a motion type and represent it by a compact way. Corresponding to a special motion type, a special manifold in base space indicates the common identity of this motion type. This can lead to the efficient estimation of human poses. We use the shape context matching to measure the similarity between the query image and the candidate poses. Experiments demonstrate that this is a robust and discriminative matching method. As the predict model, the Gaussian mixture model and the ARP model wok well in the process of tracking. In terms of future work, we will cover more types of human motions by including a wider range of training data. We plan to improve the matching method in order to reduce the computing expense further. And, the conception of base space will be extend to the recognition of human activity.

References

1. Aggarwal, J.K., Cai, Q.: Human Motion Analysis: A Review. Computer Vision and Image Understanding 73(3), 428–440 (1999)
2. Moeslund, T.B., Granum, E.: A Survey of Computer Vision-Based Human Motion Capture. Computer Vision and Image Understanding 81, 231–268 (2001)
3. Sminchisescu, C.: Optimization and Learning Algorithms for Visual Inference. In: ICCV 2005 tutorial (2005)
4. Cai, D., He, X., Han, J., Zhang, H.-J.: Orthogonal Laplacianfaces for Face Recognition. IEEE Transactions on Image Processing 15(11), 3608–3614 (2006)
5. He, X., Yan, S., Hu, Y., Niyogi, P., Zhang, H.-J.: Face recognition using laplacianfaces. IEEE Trans. on Pattern Analysis and Machine Intelligence, 27(3) (2005)

6. Agarwal, A., Triggs, B.: Tracking Articulated Motion Using a Mixture of Autoregressive Models. In: Proc. European Conf. Computer Vision (2004)
7. Belongie, S., Malik, J., Puzicha, J.: Shape Matching and Object Recognition Using Shape Contexts. IEEE Trans. Pattern Analysis and Machine Intelligence 24(4), 509–522 (2002)
8. Urtasun, R., Fleet, D.J., Fua, P.: Monocular 3D Tracking of The Golf Swing. In: Proc. IEEE CS Conf. Computer Vision and Pattern Recognition vol. 2, pp. 932–938 (2005)
9. Urtasun, R., Fua, P.: 3D Human Body Tracking Using Deterministic Temporal Motion Models. In: Proc. European Conf. Computer Vision, Prague, Czech Republic (May 2004)
10. Sidenbladh, H., Black, M., Sigal, L.: Implicit Probabilistic Models of Human Motion for Synthesis and Tracking. In: Proc. European Conf. Computer Vision, vol. 1 (2002)
11. Ormoneit, D., Sidenbladh, H., Black, M., Hastie, T.: Learning and Tracking Cyclic Human. In: Advances in Neural Information Processing Systems, vol. 13, pp. 894–900. The MIT Press, Cambridge (2001)
12. Mori, G., Malik, J.: Recovering 3D Human Body Configurations Using Shape Contexts. IEEE Trans. Pattern Analysis and Machine Intelligence 28(4), 1052–1062 (2006)
13. Mori, G., Belongie, S., Malik, J.: Efficient Shape Matching Using Shape Contexts. IEEE Trans. Pattern Analysis and Machine Intelligence 27(11), 1832–1837 (2005)
14. Ning, H., Tan, T., Wang, L., Hu, W.: People Tracking Based on Motion Model and Motion Constraints with Automatic Initialization. Pattern Recognition 37, 1423–1440 (2004)
15. Agarwal, A., Triggs, B.: Recovering 3D Human Pose from Monocular Images. IEEE Trans. Pattern Analysis and Machine Intelligence 28(1), 44–58 (2006)
16. Sminchisescu, C., Kanaujia, A., Li, Z., Metaxas, D.: Discriminative Density Propagation for 3D Human Motion Estimation. In: Proc. IEEE CS Conf. Computer Vision and Pattern Recognition. vol. 1(1), pp. 390–397 (2005)
17. Rosales, R., Sclaroff, S.: Learning Body Pose Via Specialized Maps. In: NIPS (2002)
18. Taylor, C.J.: Reconstruction of Articulated Objects from Point Correspondences in a Single Uncalibrated Image. Computer Vision and Image Understanding 80, 349–363 (2000)
19. Jin, X., Tai, C.-L.: Convolution Surfaces for Arcs and Quadratic Curves with a Varying Kernel. The Visual Computer 18, 530–546 (2002)
20. CMU database: http://mocap.cs.cmu.edu/