3D Ballet Motion Tracking from Shape Context by Using Differential Regression

Minglei Tong[1] , Hong Han[2] and Shudong Chen[3]

Abstract. An improved method, differential regression, is proposed. It is a more powerful discriminative approach for human pose estimation. The proposed methods investigate 3D body pose reconstruction by learning a simply regression between state vector differences and observation vector. A simulation model is established on a 57 dimensions human skeleton. The comparison experiments between proposed method and traditional regression are carried out by using a sequence of images on Ballet dancing. The calculated deviations are greatly reduced by 50%.

Keywords: 3D Human motion, Differential Regression, Shape Context

1 Introduction

The research of estimating 3D configurations of complex motion from monocular images is focused on, e.g. for applications requiring 3D human body pose or hand gesture analysis. This problem is made difficult by the considerable background clutter, camera movement, motion blur, poor contrast, body pose and shape variation, as well as illumination, clothing and appearance diversity. There has been a great mount of prior works on human pose recovering from a single view [1,2,3], but relatively little obtaining an outperforming result.

[1] Minglei Tong (✉)
School of Electric and Information Engineering, Shanghai University of Electric Power, Shanghai, PRC
Email:tongminglei@gmail.com
[2] Hong Han(✉)
Xi'dian University, Xi‘an, PRC
Email: hanh@mail.xidian.edu.cn
[3] Shudong Chen(✉)
Institute of Microelectronics of Chinese Academy of Sciences, Shanghai, PRC
Email: chenshudong@ciotc.org

Learning based approaches try to avoid the need for accurate 3D modeling and rendering, and to capitalize on the fact that the set of typical human poses is far smaller than the set of kinematic possible ones, by estimating (learning) a model that directly recovers pose estimates from observable image quantities[4]. In[5], a continue valued view manifold was learned via non-linear tensor decomposition that is able to interpolate visual observations of unknown views. This method was further extended in [6] to learn a gait manifold where a continuous-valued gait variable was proposed to characterize different walking styles. Unlike [5]where an explicit order is available for the view manifold, a key issue studied in [6] is how to determine an optimal manifold topology for gait interpolation.

Brand[7] models a dynamical manifold of human body configurations with a Hidden Markov Model and learns using entropy minimization. Mori [8] estimates the centers using shape context image matching against a set of training images with pre-labelled centres, then reconstruct 3D pose. Athitsos and Sclaroff [9] learn a perceptron mapping between the appearance and parameter spaces. Human pose is hard to ground truth, so most papers in this area use only heuristic visual inspection to judge their results. However, the interpolated KNN learning method of Shakhnarovich [10] used a human model rendering package to synthesize ground-truthed training. Aggarval[3] estimates full 54 DOF body pose and orientation with mean errors of only about 4 Degree.

In this paper, a learning based approach is considered, but instead of directly regressing between state vectors and observation vectors, we use state vector differential regression to distill a training sequence into a single compact model. We regress the difference between current pose (body joint angles) and prior pose against both image descriptors (silhouette shape) using a learned dynamical model. High dimensionality and the intrinsic ambiguity in recovering pose from monocular observations make the regression nontrivial. In this paper, pose is estimated indirectly, by regressing it against a dynamics based prediction and an observed shape descriptor vector. Regressing on shape descriptors allows appearance variations to be learned automatically; enabling us to work with a simple generic articulated skeleton model; while including an estimate of the pose in the regression allows the method to overcome the inherent many-to-one projection ambiguities present in monocular image observations. The proposed method is very similar with Aggarval's work. They, however, take advantage of regress the current pose (body joint angles) against both image descriptors (silhouette shape) and a pose estimate computed from previous poses using a learned dynamical model. Our strategy makes good use of the sparsity and simply linear regressor, which is called ridge regress. The human model is loaded by motion capture data (Ballet Dancing motion), furthermore, we represent 3D body pose by 57-D vectors x including 3 joint angles for each of the 19 major body joints. The input image descriptor is vector-quantized shape-context, which is the same feature in Aggarval's work

2 Tracking framework

2.1 Vector Quantized Shape Context

Histograms of edge information are a good way to encode local shape robustly. The shape context is intended to be a way of describing shapes that allows for measuring shape similarity and the recovering of point correspondences. The basic idea is to pick n points on the contours of a shape. For each point pi on the shape, consider the n − 1 vectors obtained by connecting pi to all other points. The set of all these vectors is a rich description of the shape localized at that point but is far too detailed. The key idea is that the distribution over relative positions is a robust, compact, and highly discriminative descriptor. So, for the point pi, the coarse histogram of the relative coordinates of the remaining n − 1 points, Here, shape contexts are used to encode silhouette shape over a range of scales, making use of their locality properties and capability to encode approximate spatial position on the silhouette. We resize the training images to the size of 128*128, which different style of motions, such as Ballet, Gongfu and Swin. Although the number of points along the edge in each images are different, we furthermore give a uniformly sampling along these points. Generally, in a 128*128 images, the number of points enclosing a human silhouette is about 1000, as shown in Fig 1. In the first row of the figure1, the sampling images of ballet dancing is given, whereas, the silhouette are shown in the second row . We subsample the points along the edge of the silhouette to about 200 uniformly. Then shape contexts of all points after re-sampling are clustered into 100 classes. The shape context distributions of all edge points on a silhouette are reduced to 100-D histograms by vector quantizing the 60-D shape context space using Gaussian weights to vote into the few histogram centers nearest to the contexts. Each image observation (silhouette) is thus finally reduced to a 100-D vector.

2.2 Tracking methods

The human pose in 3D world can only be observed indirectly via ambiguous and noisy image measurements, so it is appropriate to start by considering the Bayesian tracking framework in which our knowledge about the state (pose) x_0 given the 3D configuration in the first frame and the observations up to time t is represented by a probability distribution, the posterior state density $p(x_t \mid z_t, \cdots, z_0)$.

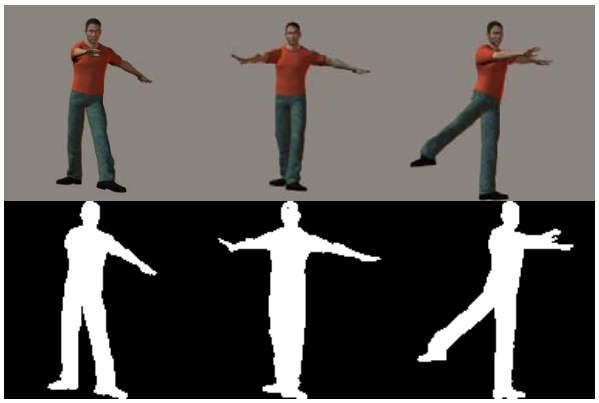


Figure 1. Ballet dancing images and silhouettes

Given an image observation z_t and a prior $p(x_t)$ on the corresponding pose x_t and the initial pose x_0, the posterior likelihood for x_t is usually evaluated using Bayes' rule, $p(x_t \mid z_t) \propto p(z_t \mid x_t) p(x_t)$. However, when tracking objects as complicated as the human body, the observations depend on a great many factors that are difficult to control, ranging from lighting and background to body shape and clothing style and texture, so any hand-built observation model is necessarily a gross oversimplification.

2.3 Dynamical Differential Regression

Aggraval [3] proposes a tracker estimating 3D body pose by using Relevance Vector Machine regression to combine a learned auto-regressive dynamical model with robust shape descriptors extracted automatically from image silhouettes in default of a reliable method for multi-valued regression. However, in our method, a dynamical differential regression is considered because the direct regression between state vector X and observation vector Z has given a relatively great derivations about 6 degrees as described in [3]

$$X = BZ$$

In which X are state vectors and Z are observation vectors, B is linear matrix. Nevertheless, we give the following regression

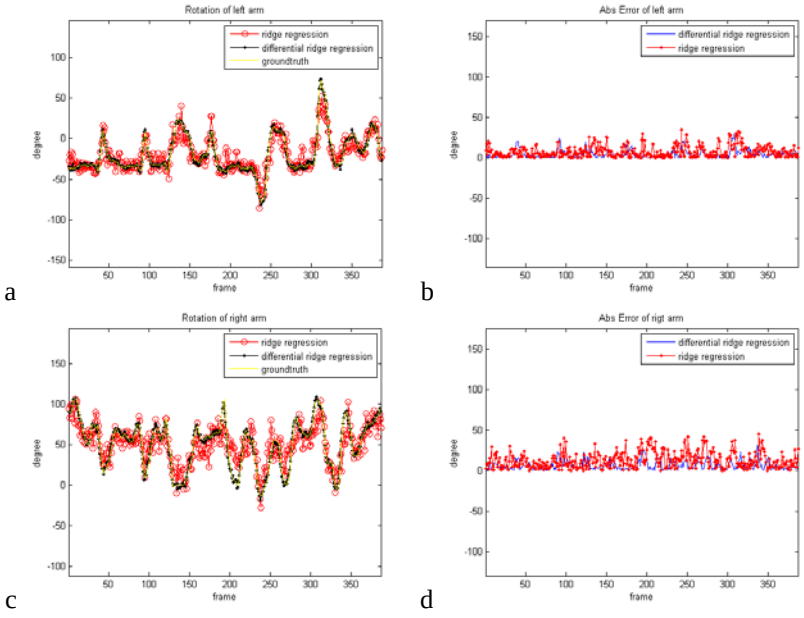
$$\Delta X = BZ ,$$

In which, $\Delta X = X_t - X_{t-1}$ .If the pose in the first frame is given, we can accurately calculate the current pose when the image observation z_t is inputted, which can greatly reduced the estimated errors.

3 Experimental results

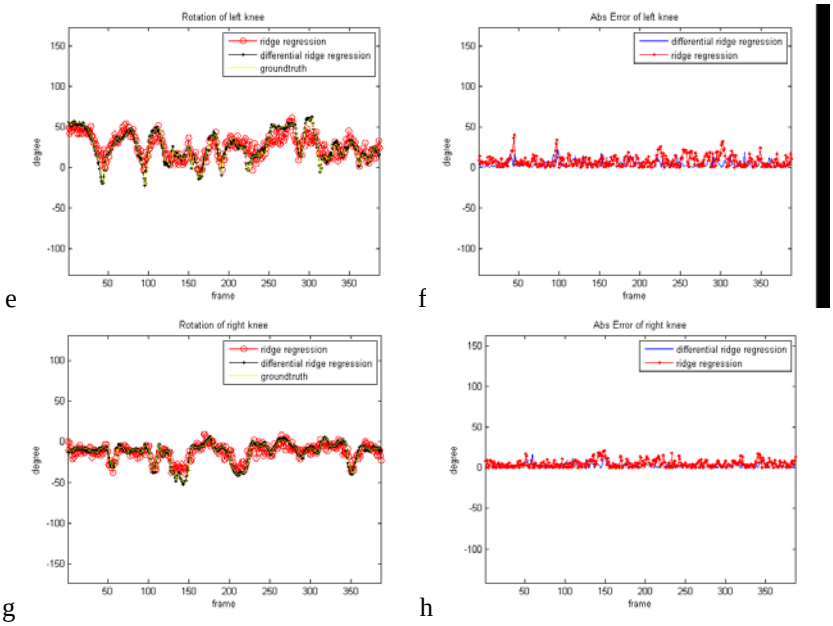
We set the similar experimental conditions as the details in Aggarval's work, but the number of joints is different . The DOF in our work is 57 and has 19 main joints. The mean (over all angles) RMS (over time) absolute difference errors between the true and estimated joint angle vectors is set in the same way.

A sequence of ballet dancing motion data has been inputted into the software Poser and the image sequence is recording by the Make-Movie function in Poser. We consequently calculate the edge of motion in every frame by Canny detector and refine the shape context of every point along the silhouette after re-sampling. The sequence of ballet dancing totally has 388 frames and the first 250 frames of images are considered as training data and the last 138 frames of images are given as test data. We compare the experimental errors by using direct regression and differential regression. In the fig 2, although we get an average RMS error reaching to 5 degrees at least, the RMS of joint of left knee is so high that in real 3D world the pose is totally deformed by the direct ridge regression. Table1 lists the absolute value of errors in different joints. From the table, the differential regression can bring a very small error in every joint.



a b

c d



e f g h

Fig 2 a,c, e,g are results using different methods. b.d.f.h are the Abs values using different methods.

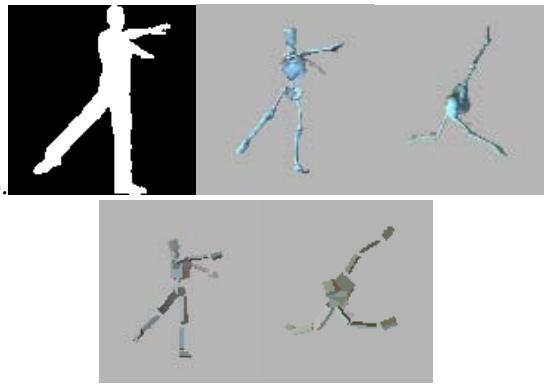
Table 1. Abs error of different joints

	Left arm	Right arm	Left knee	Right nee
Differential Ridge regression	2.2605	2.8900	2.3255	1.3451
Ridge regression	7.5597	12.7201	7.7388	5.1584

Furthermore, in our proposed way, we get a an average RMS error reaching to 2.1329 degrees as shown in fig2.b, the RMS of joint of left knee is very small. We also give some result of 3D reconstructing results by different methods. We also give some reconstructing result in Fig3. The Fig3.a gives the 51th frames and the first map is the silhouette. The second and third maps are ground truth in different views. The fourth and fifth maps are reconstruct -ing results of 51th frame.

a. 

b.



b.

Fig 3 a. gives the result of 51th frame,b. gives the result of 107-th frame. The Fig3.b gives the 107th frames and the first map is the silhouette. The second and third maps are ground truth in different views. The fourth and fifth maps are reconstructed results of 107th frame.

4 Conclusion and Future Work

We recovers 3D ballet dancing pose from sequences of monocular silhouettes by direct linear ridge regression of the difference between joint-angles against shape descriptors and dynamics based pose estimates. The method shows promising results on tracking video sequences, giving an average RMS error of 2.1 degree of angle on real motion capture data.

Future work: Although our work has great performance in the advance of prior knowledge to the information of the first frame, the performance of the proposed method will decrease a lot when the initial pose is not very accurate. The error propagation will influence on the algorithm greatly. We will find a new initial way in the further work to get a suitable starting pose and the multiple model will be included into tracking framework.

5 References

[1]C. Sminchisescu and B. Triggs. Estimating articulated human motion with covariance scaled sampling. IJRR, 2003.

[2] J. Wang, D. Fleet, and A. Hertzmann. Gaussian process dynamical models for human motion. PAMI, pages 283–298, 2007. 1282

[3]Agarwal, A., & Triggs, B. (2004a). 3D Human Pose from Silhouettes by Relevance Vector Regression. Int. Conf. Computer Vision & Pattern Recognition.

[4]Agarwal, A., & Triggs, B. (2004b). Tracking Articulated Motion with Piecewise Learned Dynamical Models. European Conf. Computer Vision.

[5] C.-S. Lee and A. Elgammal. Modeling view and posture manifolds for tracking. In Proc. of ICCV, 2007.

[6] X. Zhang and G. Fan. Dual gait generative models for human motion estimation from a single camera. IEEE Trans. On SMC-B, 40:1034–1049, 2010.

[7]Brand, M.. Shadow Puppetry. Int. Conf. Computer Vision. pp. 1237–1244,1999

[8]Mori, G., & Malik, J. (2002). Estimating Human Body Configurations Using Shape Context Matching. European Conf. Computer Vision (pp. 666–680).

[9]Athitsos, V., & Sclaroff, S. (2000). Inferring Body Pose without Tracking Body Parts. Int. Conf. Computer Vision & Pattern Recognition.

[10]Shakhnarovich, G., Viola, P., & Darrell, T. (2003). Fast Pose Estimation with Parameter Sensitive Hashing. Int. Conf. Computer Vision.