2010 International Conference on Pattern Recognition

3D Human Pose Reconstruction Using Millions of Exemplars

Hao Jiang
Computer Science Department, Boston College, Chestnut Hill, MA 02467, USA

Abstract

We propose a novel exemplar based method to estimate 3D human poses from single images by using only the joint correspondences. Due to the inherent depth ambiguity, estimating 3D poses from a monocular view is a challenging problem. We solve the problem by searching through millions of exemplars for optimal poses. Compared with traditional parametric schemes, our method is able to handle very large pose database, relieves parameter tweaking, is easier to train and is more effective for complex pose 3D reconstruction. The proposed method estimates upper body poses and lower body poses sequentially, which implicitly squares the size of the exemplar database and enables us to reconstruct unconstrained poses efficiently. Our implementation based on the kd-tree achieves real-time performance. The experiments on a variety of images show that the proposed method is efficient and effective.

1. Introduction

Our goal is to reconstruct 3D human poses from uncalibrated images by using only joint locations. Taylor [1] shows that to estimate a 3D pose of an articulated object, simply knowing the joint locations is not sufficient. There is an inherent ambiguity on the relative depth order between body part end points. If there are n body parts and their lengths are known, there are 2^n possible 3D poses. In a simple system, apart from the joint locations, user input also needs to specify the depth order of the end points of each body part before a 3D pose can be reconstructed. This procedure is usually tedious and prone to errors. We propose a method to reconstruct 3D human poses with only the joint correspondences: the required user input reduces to a few mouse clicks on the body joints, and the proposed method reconstructs 3D poses based on the user input and pose constraints.

To achieve this, we quantify the likelihood of an estimated pose and with the measurement we select the optimal estimation from all the possible 3D poses. Parametric models have been widely used to quantify the pose prior. Linear blending models are used in [3, 5]. Gaussian process has been used [4] together with nonlinear optimization to estimate 3D human poses. Recently, Wei and Chai [2] use a mixture of factor analyzers to model the pose prior in a pose editing human-computer interface. This parametric model is reported to be built on top of a very large pose dataset.

Parametric models can be quickly verified in the testing stage. However, as the pose database expands, parametric models become increasingly difficult to train. To model unconstrained poses, we need a huge number of pose samples. Currently, relatively small number of actions have resulted in millions of poses in the CMU motion capture dataset. Training parametric models on such a large dataset takes hours or days. The training time will increase rapidly when more data come in. Parametric methods therefore have to compromise the model accuracy and the training complexity. A non-parametric model is more suitable in dealing with very large pose dataset. Training such models usually needs much less effort. When properly constructed, a non-parametric model is also more faithful to the training data, performs more robustly for complex human poses, and can achieve fast inference. In this paper, we study how to use exemplar based method to reconstruct 3D human poses.

Exemplar based method has been used to estimate poses in specific action domains [6]. Another exemplar based method, locality-sensitive hashing, has been used to estimate 3D upper body poses [7] from images. In recent years, exemplar based methods become popular in many different areas such as object recognition [8] and scene understanding [9]. In this paper, we show how an exemplar based method can be used to estimate almost arbitrary 3D human poses in images by using a very large set of pose exemplars. From 2D joint locations on images, the proposed method first generates all the possible 3D poses. Each hypothesis pose is then normalized and compared with millions of pose exemplars to determine the optimum. Instead of directly processing the full body pose, we split it into upper body and lower body poses whose nearest neighbors in the pose exemplar database are found sequentially. This approach not only greatly increases the efficiency but also enables us to model a lot more poses with a relatively small exemplar database. We use the kd-tree to pre-organize the exemplars so that the nearest neighbor

1051-4651/10 $26.00 © 2010 IEEE
DOI 10.1109/ICPR.2010.414

search is efficient. Our approximate nearest neighbor implementation is able to reconstruct 3D poses in real time.

2. Our Approach

We would like to reconstruct 3D human poses from a single uncalibrated image. We assume that the 2D joint locations on the image are known. We use a 9-part body model which contains a torso, 4 half arms and 4 half legs. The user mouse clicks are only located on the limbs, while the torso end points are derived from the centers of the shoulder joints and waist joints. Therefore, the user input just contains 12 mouse clicks. Fig.1 illustrates the 2D joint locations on an image and the 3D pose reconstructed using the proposed method. There are totally 11 body structures, 8 half limbs, a torso, a shoulder and a waist, whose orientations need to be determined in the reconstructed 3D pose.

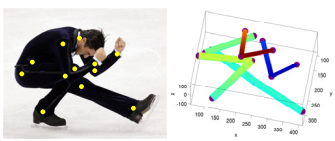

Figure 1. 3D human pose reconstruction.

2.1. Pose Hypothesis

Based on the 2D joint locations, we first estimate all the possible 3D poses. We assume a scaled orthographic projection. The hypothesis 3D poses can thus be obtained using Taylor's method [1]. Let the 2D joint locations be (x_i, y_i) and the scale of the projection be s. The depth difference dZ_k between two end points of body part k is

$$dZ_k^2 = (s \cdot l_k)^2 - (x_{k_1} - x_{k_2})^2 - (y_{k_1} - y_{k_2})^2$$

where l_k is the length of the kth body part; (x_{k_1}, y_{k_1}) and (x_{k_2}, y_{k_2}) are the two end points of the body part k. In this paper, the body part lengths are set to be the average person body part lengths. The sign of dZ_k is ambiguous in the 3D pose estimation.

The horizontal and vertical coordinates of the 3D joints are simply the corresponding xy coordinates on the image. Starting from a reference point, the neck point, we sequentially reconstruct the depth of all the joints by traversing the body tree. The scale s is a parameter that also needs to be determined, In this paper, we simply let $s = \max_k[\sqrt{(x_{k_1} - x_{k_2})^2 - (y_{k_1} - y_{k_2})^2}/l_k]$. The heuristic is found to work reasonably well since a 3D human pose usually has a body part that is nearly parallel to the image plane and achieves the largest ratio of the projected length to the body part length.

The depth difference dZ_k of the end points of body part k can be either positive or negative. A 2D pose therefore maps to many different 3D poses. However, there is usually only one correct pose. A human observer rarely makes mistakes in disambiguating poses from a monocular view. It is quite likely that a human observer hypothesizes different possible 3D poses and compares them with feasible ones in the memory to decide which pose is the most likely. The proposed method uses a similar strategy.

With a huge set of pose exemplars, we compute how similar the reconstructed 3D poses are to these exemplars. If we find a similar exemplar, the 3D pose hypothesis is likely to be true. For the comparison, we need to define a pose descriptor. In this paper, the pose descriptor contains the unit vector for each half limb. The half limb unit vectors are computed in a local coordinate system whose origin is at the neck point (the center of two shoulder joints). The right shoulder line and the torso line forms the xy plane, the torso line is the y axis, and the z and x axes are defined accordingly. Let A be the matrix whose columns are the x, y and z unit vectors for the local coordinate system. The pose descriptor is a concatenation of $\mathbf{v}_k$:

$$\mathbf{v}_k = A^{-1}(\mathbf{u}_{k_1} - \mathbf{u}_{k_2})/||A^{-1}(\mathbf{u}_{k_1} - \mathbf{u}_{k_2})||, k = 1..8$$

in which $\mathbf{u}_{k_1}$ and $\mathbf{u}_{k_2}$ are the two end points of body part k. With the pose descriptor, the pose similarity can be quantified by the angles between the feature vectors. As a simplification, we use the Euclidean distances to approximate the angle differences. Our experiments showed that these two measurements have similar performance.

Since we have 11 body structures, there are totally 2048 possible poses. A natural way to find the optimal pose is therefore to compare all the possible poses with the exemplars in the database and the optimal one has a descriptor with the smallest distance to those of the exemplars. Even though this is a working solution, it is quite slow due to the large number of hypothesises. To solve this problem, instead of directly matching full body poses, we split them into upper body and lower body poses. The upper body pose is determined by torso, shoulder and arms, and its descriptor contains 12 elements. The lower body contains waist and legs; its descriptor is also 12D. Note that the lower body descriptor is dependent on the torso and shoulder orientations. Therefore the lower body descriptor can be determined only when the upper body pose is estimated. This is why we use a sequential decision procedure. As follows we present details for exemplar pose database construction and 3D pose reconstruction.

2.2. Building the Exemplar Database

We construct an exemplar database that contains different poses in our everyday life. In this paper, the 3D pose exemplars are constructed using the CMU motion capture data. There are totally 4164772 pose exemplars. Even though the CMU dataset contains many different

poses, it does not cover all the possible ones. This does not pose a serious problem, since we match the upper body and lower body poses sequentially. Such a procedure implicitly squares the number of exemplars so that we can match almost arbitrary pose using a relatively small number of exemplars. We normalize the orientation of the pose exemplars using similar method to feature extraction in the previous section. The pose exemplars are quantified by 12D vectors for the upper body and the lower body poses. We determine how good a 3D pose reconstruction is by computing the shortest distance from the pose to the exemplars.

Directly matching hypothesis poses with exemplar poses is slow due to the very large database. We resort to the approximate nearest neighbor (ANN) method. In this paper, we use the kd-tree implementation to speed up the nearest neighbor search. Since we sequentially find the upper body poses and lower body poses, we build an upper body kd-tree and a lower body kd-tree separately. Compared with parametric models such as Gaussian Mixture Model (GMM), a kd-tree is much easier to train. Using a 2.8GHz machine, constructing a kd-tree with about 4 million pose descriptors takes a few minutes, while training a GMM with 50 Gaussian mixtures takes hours. In testing, the time of finding approximate nearest neighbor using a kd-tree is comparable to computing the probability density function of a GMM. Since we use all the data in the original pose dataset to evaluate pose hypothesises, our non-parametric method is more accurate than a simplified GMM model and we avoid the problem of determining how many Gaussian mixtures are really needed; for unconstrained poses, this number would be large.

2.3. 3D Pose Reconstruction

We are ready for 3D pose reconstruction. The procedure is summarized as follows:

Algorithm:

Input joint locations
N *= 64,* M*= 32*
for k *= 1 to* N
  n *= binary number of* $(k-1)$ *in 6 digits*
  generate upper body pose using n
  compute pose descriptor $\mathbf{v}_{up}$
  d *= ANN distance of* $\mathbf{v}_{up}$ *to the upper body exemplars*
  n^* *=* n*, if* d *is the smallest*
for k *=* 1 *to* M
  m *= binary number of* $(k-1)$ *in 5 digits*
  generate lower body pose using n and m*
  compute pose descriptor $\mathbf{v}_{low}$
  d *= ANN distance of* $\mathbf{v}_{low}$ *to the lower body exemplars*
  m^* *=* m*, if* d *is the smallest*
Merge the upper body pose corresponding to n^* *and lower body pose corresponding to* m^*

Here, *ANN* means approximate nearest neighbor. In this algorithm, the upper body pose contains the configurations of 6 bar structures and the lower body pose includes the configurations of 5 bar structures. The orientation of each upper body structure is determined by a single bit in n^* and the orientation of each lower body structure is determined by a bit in m^*. Using approximate nearest neighbor method and the kd-tree, the proposed method has an overall complexity $O(\log(K))$, where K is the number of exemplars. By separating the upper body and lower body pose estimation, we further speed up the computation by about 20 times. On a 2.8GHz machine, a 3D pose reconstruction takes a fractional second.

3. Evaluations and Discussions

We test the proposed method on images we took and images from the web. Fig.2 shows the 3D pose reconstructions using the images of a flexible toy. The yellow dots on the images are manually labeled joint locations. The red-blue stick figure next to each image shows the most similar pose in the exemplar database. The color coded stick figures show the estimated 3D poses using the proposed method. The color on the stick figures is cooler for small Z values and warmer for large Z values. Here the Z axis is perpendicular to the image plane and points to a viewer. The proposed method works quite well in reconstructing challenging poses. We further test the proposed method on randomly selected images from the web. These images cover many different human poses in sports and everyday activities. Our results are shown in Fig.3. The proposed method works very well in reconstructing different human poses.

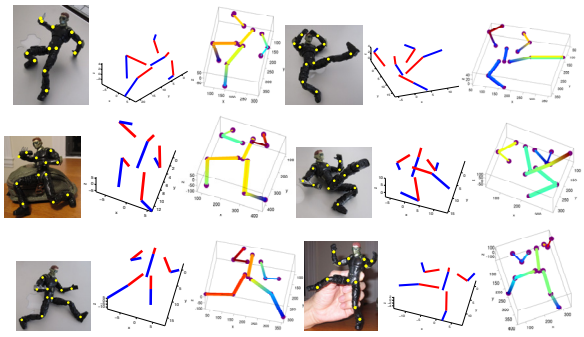

Figure 2. 3D pose reconstruction on the flexible toy images. Warmer color on the 3D stick figure indicates "closer" and cooler color indicates "farther away".

We proceed to compare the proposed method with a standard GMM method. We train two separate GMM models, one for upper body poses and the other for lower body poses. We use 50 Gaussian mixtures for each model. The training for each GMM takes more than 5 hours on a 2.8GHz machine, while the proposed method uses only about 3 minutes to generate the upper

Figure 3. 3D pose reconstruction sample results on the web images.

body and lower body search trees. Fig.4 compares the GMM result with the proposed method's result. The GMM method often makes mistakes in determining the orientations of the body parts, while the proposed method works much more robustly. The advantage of the proposed method over GMM is not a surprise. Due to limited Gaussian mixtures a GMM can handle when training with a very large dataset, it only models the ordinary poses, which are the large number of walking and running poses in the exemplar database. The proposed non-parametric method has no such problem and is therefore able to yield more reliable results.

4. Conclusion

We propose a novel exemplar based method to reconstruct 3D human poses from uncalibrated single view images using only a few body joint correspondences. Our implementation based on the kd-tree is both efficient and effective. The experiment on a variety of images shows that the proposed method is more robust than the widely used Gaussian Mixture Model in 3D pose estimation, especially when dealing with complex poses. The proposed method is useful for many applications including motion capture, animation and human computer interaction.

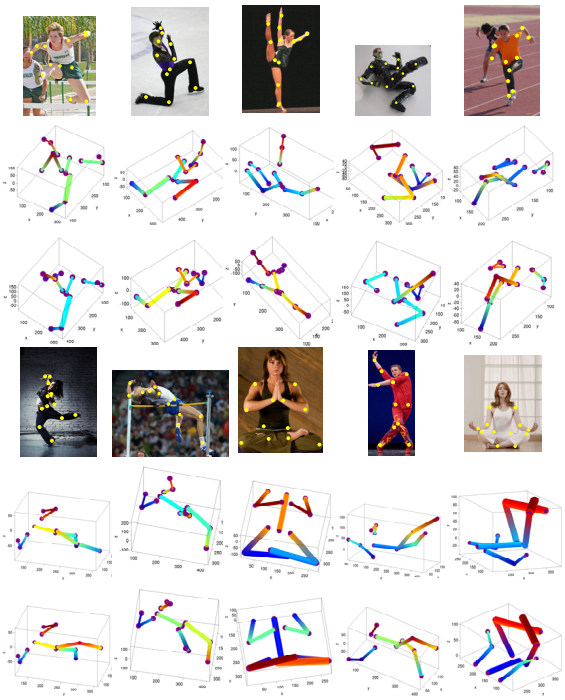

Figure 4. Comparison with GMM. Row1, 4: input images; Row 2, 5: the GMM pose estimation results for the corresponding images; Row 3 and row 6 show how the proposed method improves the results.

References

[1] C.J. Taylor, "Reconstruction of Articulated Objects from Point Correspondences in a Single Uncalibrated Image", Computer Vision and Image Understanding, vol.80, no.10, pp.349-363, 2000

[2] X. Wei and J. Chai, "Intuitive Interactive Human Character Posing with Millions of Example Poses", IEEE Computer Graphics and Applications, vol.30, no.1, 2010.

[3] L. Kovar and M. Gleicher, "Automated Extraction and Parameterization of Motions in Large Data Sets", ACM Transactions on Graphics (2004). 23(3):559-568.

[4] K. Grochow, S.L. Martin, A. Hertzmann and Z. Popovic, "Style-based Inverse Kinematics", ACM Transactions on Graphics (2004). 23(3):522-531.

[5] C.F. Rose,III, P.-P.J. Sloan and M.F. Cohen, "Artist-directed Inverse-Kinematics Using Radial Basis Function Interpolation", Computer Graphics Forum (2001). 20(3): 239-250.

[6] G. Mori and J. Malik, "Estimating Human Body Configurations Using Shape Context Matching", ECCV 2002.

[7] G. Shakhnarovich, P. Viola, and T. Darrell, "Fast Pose Estimation with Parameter Sensitive Hashing", ICCV 2003.

[8] D. Nister and H. Stewenius, "Scalable Recognition with a Vocabulary Tree", CVPR 2006.

[9] J. Sivic and A. Zisserman, "Video Google: A Text Retrieval Approach to Object Matching in Videos", ICCV 2003.