Discriminative Estimation of 3D Human Pose Using Gaussian Processes

Xu Zhao*, Huazhong Ning†, Yuncai Liu*, Thomas Huang†

* *Shanghai Jiao Tong University, Shanghai 200240, China. {zhaoxu,whomliu}@sjtu.edu.cn.*

†*ECE Department, U. of Illinois at Urbana-Champaign, Urbana, IL 61801. {hning2, huang}@ifp.uiuc.edu.*

Abstract

In this paper, we present an efficient discriminative method for human pose estimation. This method learns a direct mapping from visual observations to human body configurations. The framework requires that the visual features should be powerful enough to discriminate the subtle differences between similar human poses. We propose to describe the image features using salient interest points that are represented by SIFT-like descriptors. The descriptor encode the position, appearance, and local structural information simultaneously. Bag-of-words representation is used to model the distribution of feature space. The descriptor can tolerate a range of illumination and position variations because it is computed on overlapped patches. We use Gaussian process regression to model the mapping from visual observations to human poses. This probabilistic regression algorithm is effective and robust to the pose estimation problem. We test our approach on the HumanEva data set. Experimental results demonstrate that our approach achieves the state of the art performance.

1. Introduction

Recovering 3D Human pose from visual signal is an active research field in recent years. A wide spectrum of potential applications such as behavior understanding, content-based image retrieval, and visual surveillance motivate the endeavors to find robust solutions to this problem. However, this problem is extremely challenging due to the complicated nature of human motion and information limitation in 2D images.

The methods for human pose estimation can be summarized as two categories [19]: *generative and discriminative*. Generative methods [5, 17, 20, 1, 12, 11] follow the bottom-up Bayes' rule and model the state posterior density using observation likelihood or cost function. This class of methods are generally computationally expensive. A Discriminative method [19, 2, 15, 4, 7, 16] models the state posterior directly by learning an image-to-pose mapping. Once the training process is completed, pose estimation will be fast. In this work, we focus on recovering 3D human pose within the discriminative framework.

How to utilize the image information is critical to the problem of pose estimation. Many methods [2, 6, 19, 7] represent human images using body silhouettes. This representation has the advantage of containing strong cues for pose estimation while being invariant to appearance and lighting. Generally, it can be extracted by background subtraction. However, the information lose of interior appearance may introduce one-to-many ambiguities to the mapping from silhouette to pose. This multi-modal ambiguity of state posterior distribution is one of the main error source of pose estimation. Intuitively, it is beneficial to utilize the interior appearance information. At least it can alleviate the ill-condition of this problem. This was proved by experiments in some recent work [13, 3]. In this paper, we propose to describe the image feature using the salient interest points that are represented by SIFT descriptor [10]. Bag-of-words representation is used to model the distribution of feature space. This sparse and local image descriptor attempts to not only capture the spatial co-occurrence and context information of the local structure but also encode their relative spatial positions. These properties make the descriptor discriminative for the task of pose estimation. The descriptor also tolerates a range of illumination and position variations because it is computed on overlapped patches, instead of pixels.

How to model the mapping between pose space and feature space lies in the heart of pose estimation. The approaches to solving this problem varying from neural networks [15], fast nearest-neighbor retrieval [16, 11], regression methods [2, 1], to Bayesian mixture of experts [19, 13]. Our method is based on the non-parametric Gaussian Process (GP) regression. GP is flexible, fully probabilistic, and effective for small sample problem. Such properties are desirable to pose es-

978-1-4244-2175-6/08/$25.00 ©2008 IEEE

timation. We tested our approach on the real dataset HumanEva [18] and have achieved the state of the art performance. The sparse GP makes robust and accurate predictions very fast, once the model is trained.

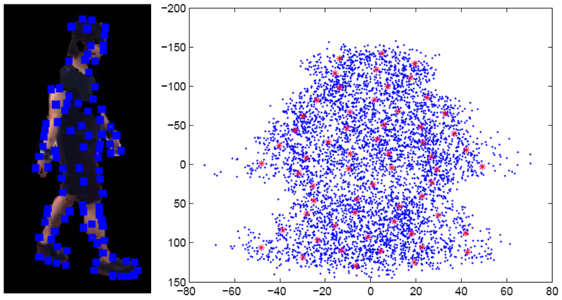
2. Image feature and descriptors

Pose estimation relies on elaborate design of feature representation. Image information should be utilized as much as possible to discriminate the subtle differences in human pose. The local structures and interior relative positions of body parts play an important role in pose estimation. Under such consideration, we design a specific descriptor to deal with the problem of pose estimation.

Our descriptor is extracted in the following steps. (1) For each image, the human window is detected and rescaled to a fixed size. (2) Detect interest points using Harris corner detector [9] within the image window (Extra background substraction is not necessary, but it may slighly improve the performance). Fig. 1.(a) depicts the interest points on a sample image frame. (3) Centered at each interest point, we compute a SIFT descriptor [10]. Denote the descriptor as a vector $\mathbf{c}$. (4) Find the relative coordinate (x, y) of each interest point. (5) The final descriptor therefore is $\mathbf{d} = (x, y, \mathbf{c})^T$.

In implementation, the local region of each interest point is partitioned into 9 cells, and a 9-orientation histogram is computed on each cell. Together with the relative coordinates of the interest point, the descriptor is a 83-dimensional vector. Our descriptor encodes the appearance, edge, and position information into a vector and alleviates the multi-modal ambiguities of posterior pose distribution to a large extent. It is meaningful to eliminate the left-right ambiguity of human body motion.

With the local descriptor ready at hand, we use the bag-of-word model [8] to represent the human image. Our bag-of-words model is obtained by an unsupervised method. First, the descriptors extracted from all training images are clustered by K-means, and the K cluster centers, called visual words, form a set $\mathcal{C} = \{\mathbf{c}_1, \mathbf{c}_2, \cdots, \mathbf{c}_K\}$ that is the so called code book. In our experiments, the number of visual words is 60. After the code book is available and given a testing image and its descriptor set $\mathcal{D} = \{\mathbf{d}_1, \mathbf{d}_2, \cdots, \mathbf{d}_m\}$, each descriptor votes softly with respect to the visual words. The bag-of-words representation, denoted as $\mathbf{x}$, is the accumulating scores of all descriptors. Fig. 1.(b) shows the relative coordinates of interest points and visual words.



(a)

(b)

Figure 1. (a) Interest points in a sample image frame. (b) The relative coordinates of interest points and visual words.

3. Gaussian process for pose estimation

3.1 Gaussian processes

Gaussian Processes (GP) [14] are generalizations of Gaussian distributions defined over infinite index sets. Thereby a GP can be used to specify distribution over functions. Given a training set $\mathcal{S} = \{(\mathbf{x}_i, y_i), i = 1, \cdots, N\}$ where $(\mathbf{x}_i, y_i)$ is an image-to-pose pair (y_i's are normalized so that they are zero-mean unit variance process), we suppose that the relationship between $\mathbf{x}_i$ and y_i is modeled by

$$y_i = f(\mathbf{x}_i) + \epsilon_i \tag{1}$$

where $\epsilon_i \sim (0, \beta^{-1})$ and β is a hyper-parameter representing the precision of the noise. Introducing a GP prior over functions f, we have

$$p(\mathbf{f}|\mathbf{X}) = \mathcal{N}(0, \mathbf{K}) \tag{2}$$

where $\mathbf{f} = [f_1, \cdots, f_N]^T$ is the function values, $f_i = f(\mathbf{x}_i), \mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_N]$, and $\mathbf{K}$ is a covariance matrix whose entries are given by a covariance function, $K_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j)$. In this paper, the kernel function we adopted is

$$k(\mathbf{x}_i, \mathbf{x}_j) = \theta_0 \exp\left\{-\frac{\theta_1}{2} \parallel \mathbf{x}_i - \mathbf{x}_j \parallel^2\right\} + \theta_2 + \theta_3 {\mathbf{x}_i}^T \mathbf{x}_j \tag{3}$$

For an unseen observation $\mathbf{x}_{N+1}$, the joint distribution therefore is

$$p(\mathbf{Y}_{N+1}) = \mathcal{N}(\mathbf{Y}_{N+1}|\mathbf{0}, \mathbf{C}_{N+1}) \tag{4}$$

where $\mathbf{Y}_{N+1} = [y_1, \cdots, y_N, y_{N+1}]^T$, and the covariance matrix $\mathbf{C}_{N+1}$ is given by

$$\mathbf{C}_{N+1} = \begin{pmatrix} \mathbf{C}_N & \mathbf{k} \\ \mathbf{k}^T & c \end{pmatrix} \quad (5)$$

$\mathbf{C}_N$ has elements

$$C(\mathbf{x}_i, \mathbf{x}_j) = k(\mathbf{x}_i, \mathbf{x}_j) + \beta^{-1}\delta_{ij} \quad (6)$$

where δ_{ij} is Kronecker delta function.

3.2 Model training

During training, the hyper-parameters $\Theta = \{\theta_0, \cdots, \theta_3, \beta\}$ of GP are learned by minimizing

$$-\ln p(\mathbf{X}, \Theta|\mathbf{Y}) = \frac{d}{2}\ln|\mathbf{C_N}| + \frac{1}{2}\mathbf{Y}^T\mathbf{C}_N^{-1}\mathbf{Y} + r \quad (7)$$

where $\mathbf{Y} = [y_1, \cdots, y_N]^T$, $r = N\ln(2\pi)/2$ is a constant, and d is the dimension of output y_i.

Once the GP model is learned, the conditional distribution $p(y_{N+1}|\mathbf{Y}, \mathbf{X})$ is a Gaussian distribution with mean and covariance given by

$$m(\mathbf{x}_{N+1}) = \mathbf{k}^T\mathbf{C}_N^{-1}\mathbf{Y} \quad (8)$$

$$\sigma^2(\mathbf{x}_{N+1}) = c - \mathbf{k}^T\mathbf{C}_N^{-1}\mathbf{k} \quad (9)$$

where $c = k(\mathbf{x}_{N+1}, \mathbf{x}_{N+1}) + \beta^{-1}$. With Eqn. 8 and 9, inference of new testing samples is easy and fast.

4. Experiments

We test our approach on the publicly available HumanEva dataset for the evaluation of human pose estimation, collected at Brown University [18]. The dataset was captured simultaneously using a calibrated marker-based motion capture system and multiple high-speed video capture systems. The video and motion capture streams were synchronized by software. It contains multiple subjects performing a set of predefined actions with repetitions.

4.1 Experiments setup

The original motion capture data provided by HumanEva were (x, y, z) locations of the body parts in the world coordinate system. There is a total of 10 parts: torso, head, upper and lower arms, and upper and lower legs. In this paper, we convert the (x, y, z) locations to global orientation of torso and relative orientation of adjacent body parts. Each orientation is represented by 3 Euler angles. We make the assumption that all the angles are independent and each has separate GP model. The joint angle trajectories are normalized so that it is a zero-mean unit variance process.

The HumanEva dataset was originally partitioned into training, validation, and testing sub-sets. We use sequences in the original training sub-set for training and those in the original validation sub-set for testing. The original testing sub-set is not used because motion data were not provided for it. A total of 2950 frames (first trial of subject S1, S2, and S3) for walking motion, 2345 frames for jog motion, and 2486 frames for box motion are used. All of the images are taken from a single camera (C1) because our approach recovers human pose from monocular images.

4.2 Experiments results

We report mean RMS absolute difference errors between the true and estimated joint angles, in degrees as in [2]:

$$D(\mathbf{y}, \mathbf{y}') = \frac{1}{m}\sum_{i=1}^{m} |(y_i - y_i')\mathrm{mod} \pm 180^\circ| \quad (10)$$

Table 1 reports the average RMS error over all joints angles. Our approach outperforms the state of the art algorithm [13] in estimation accuracy on the walking sequence, although it is slightly worse on jog and box sequences. Fig. 2 plots the estimation and ground truth of two joint angles in walking and boxing action respectively. The curves of estimation are close to the ground truth although they are less smooth. Fig. 3 shows some sample frames together with the estimated pose represented as the outline of a cylinder based human model superimposed onto the original images.

Table 1. Comparison of average RMS error over all joints, for sequences of walking, boxing, jogging.

Sequence	Walking	Jog	Box
Our work	6.38°	4.21°	6.15°
Ning [13]	6.68^o	4.12^o	5.50^o

5. Conclusions

In this work, we proposed a discriminative method based on Gaussian process regression. We designed an informative representation of visual observations. This image representation is suitable for human pose estimation because the position, appearance, and local structural information are encoded simultaneously. We use

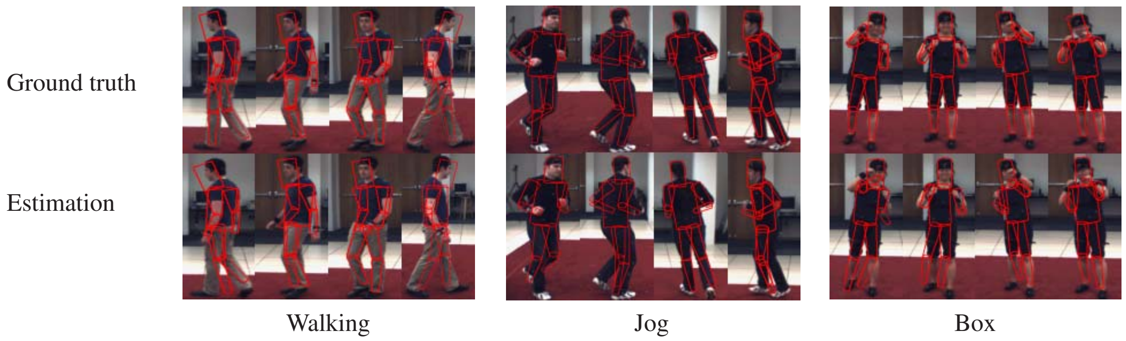


Figure 3. Sample estimation results. First row is the ground truth and second the estimation.

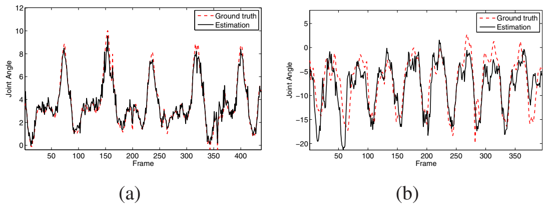


Figure 2. Joint angles: ground truth and estimation. (a) Right hip of subject S2 in walking; (b) Left hip of subject S3 in jog.

bag-of words representation to model the distribution of feature space. We use Gaussian process regression to model the mapping from visual observations to human poses. This probabilistic regression algorithm is effective and robust to the problem of pose estimation. We test our approach on the HumanEva data set. Experimental results demonstrate that our approach achieves the state of the art performance.

References

[1] A. Agarwal and B. Triggs. Tracking articulated motion using a mixture of autoregressive models. *ECCV*, 2004.

[2] A. Agarwal and B. Triggs. Recovering 3 D human pose from monocular images. *PAMI*, 2006.

[3] A. Bissacco, M.-H. Yang, and S. Soatto. Fast human pose estimation using appearance and motion via multi-dimensional boosting regression. *CVPR*, 2007.

[4] M. Brand. Shadow puppetry. *ICCV*, 2:1237, 1999.

[5] J. Deutscher, A. Blake, and I. Reid. Articulated body motion capture by annealed particle filtering. *CVPR*, 2000.

[6] C. H. Ek, P. H. Torr, and N. D. Lawrence. Gaussian process latent variable models for human pose estimation. *Lecture Notes in Computer Sciences (LNCS)*, 2007.

[7] A. Elgammal and C. Lee. Inferring 3D body pose from silhouettes using activity manifold learning. *CVPR*, 2004.

[8] L. Fei-Fei and P. Perona. A bayesian heirarchical model for learning natural scene categories. *CVPR*, 2005.

[9] C. Harris and M. Stephens. A combined corner and edge detector. *Alvey Vision Conference*, 15:50, 1988.

[10] D. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *IJCV*, 2004.

[11] G. Mori and J. Malik. Recovering 3 D Human Body Configurations Using Shape Contexts. *PAMI*, 2006.

[12] H. Ning, T. Tan, L. Wang, and W. Hu. People tracking based on motion model and motion constraints withautomatic initialization. *Pattern Recognition*, 2004.

[13] H. Ning, X. Wei, Y. Gong, and T. Huang. Discriminative Learning of Visual Words for 3D Human Pose Estimation. *CVPR*, 2008.

[14] C. Rasmussen and C. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[15] R. Rosales and S. Sclaroff. Learning Body Pose via Specialized Maps. *NIPS*, 2001.

[16] G. Shakhnarovich, P. Viola, and T. Darrell. Fast pose estimation with parameter-sensitive hashing. *ICCV*, 2003.

[17] H. Sidenbladh, M. Black, and D. Fleet. Stochastic tracking of 3D human figures using 2D image motion. *ECCV*, 2:702–718, 2000.

[18] L. Sigal and M. Black. Humaneva: Synchronized video and motion capture dataset for evaluation of articulated human motion. *Technical Report CS-06-08, Brown University*, 2006.

[19] C. Sminchisescu, A. Kanaujia, Z. Li, and D. Metaxas. Discriminative density propagation for 3d human motion estimation. *CVPR*, pages 217–323, 2005.

[20] C. Sminchisescu and B. Triggs. Covariance scaled sampling for monocular 3D body tracking. *CVPR*, 2001.