Human Upper Body Pose Estimation in Static Images

1. Research Team

Project Leader: Prof. Isaac Cohen, *Computer Science*

Graduate Students: Mun Wai Lee

2. Statement of Project Goals

This goal of this project is to develop techniques for estimating the pose of people in static images automatically.

3. Project Role in Support of IMSC Strategic Plan

Imagery data is an important component of multimedia content and appears commonly in the Internet domain, TV programs and movies. Analysis and interpretation of imagery data is therefore an important research area in IMSC. The project focuses on the human body, which is the most interesting object, and aims to develop techniques for estimating the body pose automatically. Potential applications include image understanding, categorization and retrieval. This research also addresses some of the problems in the modeling and learning of human body, including the pose, shape and clothing. The project is currently focusing on the upper body pose.

4. Discussion of Methodology Used

Overview

Estimating human pose in static images is challenging due to the high dimensional state space, presence of image clutter and ambiguities of image observations. We adopted an MCMC framework for estimating 3D human upper body pose which allows us to utilize a generative model that incorporates domain knowledge about the human articulated structure and to formulate appropriate likelihood measures for evaluating candidate samples in the solution space. In addition, we adopted a data-driven proposal mechanism for efficiently searching the solution space. We introduce the use of proposal maps, which is an efficient way of implementing inference proposals derived from multiple types of image cues.

Human Model

We propose to address this problem, by building an image generative model and using the MCMC framework to search the solution space. The human model (see Figure 1) consists of the articulated structure of the human body, (ii) a probabilistic shape model, and (iii) a cloth model. The articulated structure of the upper body consists of 7 joints, 10 body parts and 21 degree of freedom (6 for global orientation and 15 for joint angles). In the probabilistic shape model, each body part is represented by a truncated 3D cone and the shape of a 3D cone has 3 free parameters: the length of the cone and the widths of the top and base of the cone. The clothing model describes the type of clothing the person is wearing. For simplicity, we use sleeve length

to describe the portions of the arms the clothing covers. For computation efficiency, we quantized this parameter into five discrete levels.

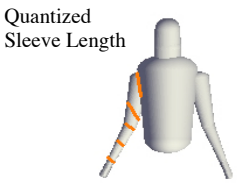

Figure 1: Clothing Model

Figure 2: Image observations, from left: (i) original image, (ii) detected face and head-shoulders contour, (iii) skin color ellipses extraction.

Observation

Image observations are used to compute data driven proposal distribution in the MCMC framework. The extraction of observations consists of 3 distinct stages: (*i*) face detection using the Adaboost technique proposed by [15], (*ii*) head-shoulders contour matching used a gradient-descent approach and active shape contour, and (*iii*) skin blobs detection using a histogram-based classifier. Examples of these observations are shown in Figure 2. These observations, weighted according to their saliency and their joint distributions, are used to generate "*proposal maps*", which represent the proposal distributions of the image positions of body joints. Figure 3 shows the pseudo-color representation of the proposal maps for various body joints. Notice that the proposal maps have multiple modes, especially for the arms, due to ambiguous observations and image clutters.

Original Image

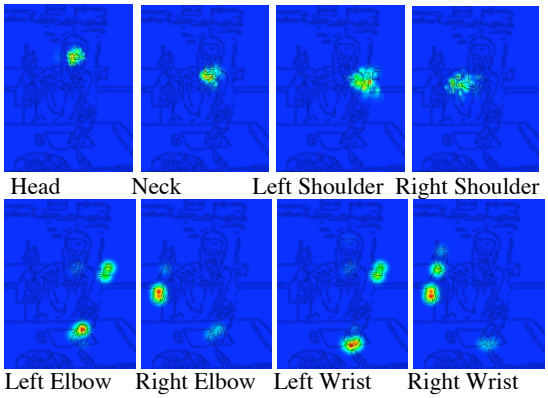

Figure 3: Proposal Maps for various Body Joints.

Pose Estimation

To search the solution space, we use the MCMC framework that allows us to sample a complex solution space that has multiple peaks. The set of sample solution hypotheses generated by the Markov chain has a stationary distribution equal to the posterior distribution. This framework is suitable not only for finding the optimal solution, but also for extracting a set of plausible alternative solutions to reflect the inherent ambiguities that exist in body pose observed from a single image.

Data-driven MCMC framework [16] allows us to design good proposal functions, derived from image observations, to explore the solution space The observations are first used to infer solutions on a set of 2D pose variables (can be viewed as hidden variables), and subsequently

generate proposals on the 3D pose using inverse kinematics. The use of proposal maps for the intermediate 2D pose variables improves considerably the search in the solution space by consolidating the evidences provided by the collective set of image observations.

The MCMC framework consists of 3 types of dynamics: *(i) diffusion dynamic*, which serves as a local optimizer by adding "noise" to the state, *(ii) proposal jump dynamic* that allows exploratory search across different regions of the solution space using proposal maps derived from observation. *(iii) flip dynamic* that involves flipping a body part (i.e. head, hand, lower arm or entire arm) along depth direction, around its pivotal joint [13]. The solution samples are evaluated using a cost function consisting of the prior distribution and the image likelihood function. The image likelihood function consists of two components: (*i*) a region likelihood, and (*ii*) a color likelihood. We have opted for an adaptation of the image likelihood measure introduced by [17].

Experimental Results

We used images of indoor meeting scenes as well as outdoors images collected from the Internet for testing. We had generated ground truth data by manually locating joint positions in the images and estimating the relative depths of the joints. Figure 4 shows the obtained results on various images. These images were not among the training data. The estimated human model and its pose (solutions with the highest posterior probability) are projected onto the original image and a 3D rendering from a sideward view is also shown.

The estimated joint positions were compared with the ground truth data, and a RMS error was computed. Since the depth had higher uncertainties, we computed two separate measurements, one for the 2D positions, and the other for the depth. Histograms of these errors (18 images) are shown in Figure 5. Figure 6 shows the RMS errors (averaged over test images) with respect to the MCMC iterations. As the figure shows, the error for the 2D image position decreases rapidly from the start of the MCMC process and this is largely due to observation-driven proposal dynamics. For the depth estimate, the kinematics flip dynamic was helpful in finding hypotheses with good depth estimation. It however required a longer time for exploration. The convergence time however varies considerably among different images, depending on the quality of the image observations. For example, if there were many false observations, the convergence took longer. On average, 1000 iterations took about 5 minutes.

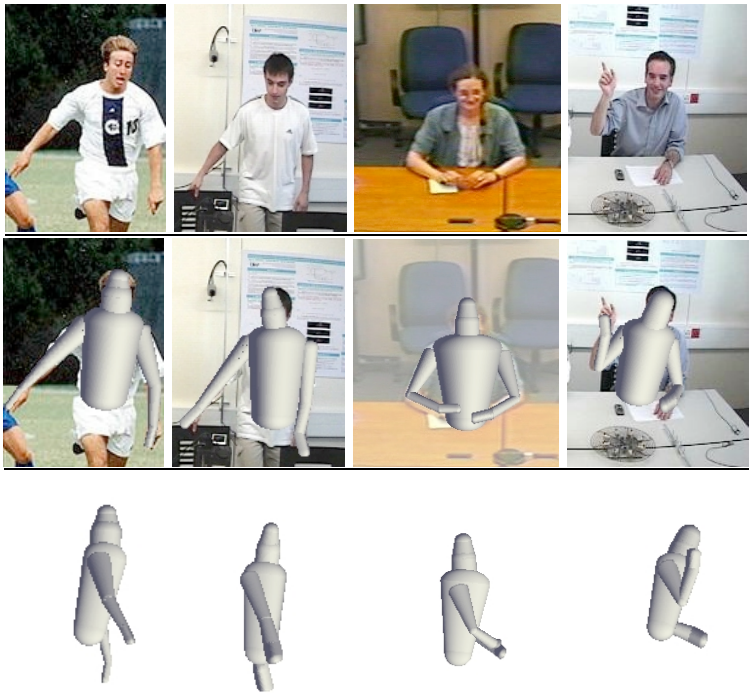

Figure 4: Pose Estimation. First Row: Original images, second row: estimated poses, third row: estimated poses (side view)

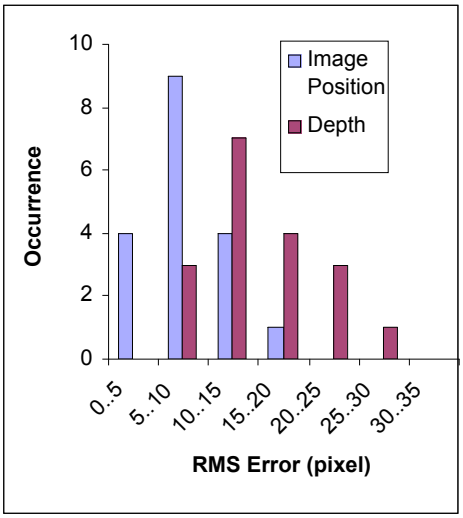

Figure 5: Histogram of RMS Error.

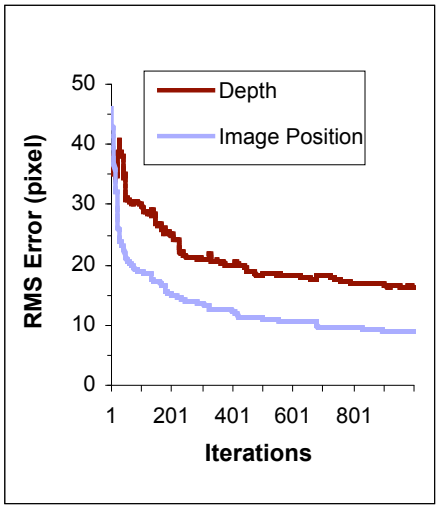

Figure 6: Convergence Analysis

5. Short Description of Achievements in Previous Years

In previous years, we have addressed the problem of body tracking with multiple cameras.

5a. Detail of Accomplishments During the Past Year

The proposed framework has been designed and implemented. A paper was prepared describing the details of this framework, and it is accepted for the upcoming European Conference for Computer Vision (ECCV 2004) [8].

6. Other Relevant Work Being Conducted and How this Project is Different

Pose estimation on video has been addressed in many previous works, either using multiple cameras [4] or a single camera [2][3][12]. Many of these works used the sequential Monte Carlo approach (particle filter) to estimate the body pose over time, by relying on a good initialization, temporal smoothness and sometimes a low dimensional dynamic model [3].

For static images, some works have been reported for recognizing prototype body poses using shape context descriptors and exemplars [9] and silhouette boundary [5]. Another related work involves the mapping of image features into body configurations [11]. These works however rely on either a clean background or that the human is segmented by a background subtraction and therefore not suitable for fully automatic pose estimation in static images.

Various reported efforts were dedicated to the detection and localization of body parts in images. In [10][5].[6], the authors model the appearance and the 2D geometric configuration of body parts. These methods focus on real-time detection of people and do not estimate 3D body pose. Recovering 3D pose was studied in [1][14], but the proposed methods assume that image positions of body joints are known and therefore tremendously simplify the problem.

7. Plan for the Next Year

We are currently extending our research to full body pose estimation and to video-based human body tracking.

8. Expected Milestones and Deliverables

For the next five-years, we will focus on expanding current method for recognizing a larger set of postures and body gestures for multimodal interactions.

2004-2005

- Improving the 3D body posture inference other image cues
- Improving the numerical complexity for achieving near real-time performances

2005-2006

- Gesture recognition from the fitted articulated body model
- Computer interaction using basic gestures.

2006-2008

- Multimodal interaction using body motion

9. Member Company Benefits

The purpose of this research is the ability to recognize the 3D posture and track its temporal variation for recognizing the person's gesture from a single camera. Various applications of this technique in human computer interaction are foreseeable.

10. References

[1] C. Barron, I. A. Kakadiaris. Estimating anthropometry and pose from a single image, CVPR *2000, vol.1*, pp. 669-676.

[2] C. Bregler, J. Malik. Tracking people with twists and exponential maps, *CVPR 1998,* pp. 8–15.

[3] K. Choo, and D.J. Fleet. People tracking with hybrid Monte Carlo, *ICCV 2001,* vol. 2, pp. 321-328.

[4] J. Deutscher, A. Davison, I. Reid. Automatic partitioning of high dimensional search spaces associated with articulated body motion capture, *CVPR 2001,* vol. 2, pp. 669 -676.

[5] I. Haritaoglu, D. Harwood, L.S. Davis. W^4: real-time surveillance of people and their activities, *PAMI, 22(8)*, pp.809–830, Aug. 2000.

[6] S. Ioffe and D.A. Forsyth. Probabilistic methods for finding people, *IJCV 43(1),* pp.45-68, June 2001.

[7] M. Lee, I. Cohen, "Human Body Tracking with Auxiliary Measurements ", *IEEE International Workshop on Analysis and Modeling of Faces and Gestures 2003*.

[8] M. Lee, I. Cohen, " Human Upper Body Pose Estimation in Static Images", accepted for *ECCV 2004*.

[9] G. Mori and J. Malik. Estimating Human Body Configurations using Shape Context Matching. *ECCV 2002*, pp 666-680.

[10] R. Ronfard, C. Schmid, and B. Triggs. Learning to parse pictures of people. *ECCV 2002*, vol. 4, pp. 700-714.

[11] Rosales, R.; Sclaroff, S. Inferring body pose without tracking body parts, *CVPR 2000*, vol.2, pp. 721-727.

[12] C. Sminchisescu, B. Triggs. Covariance Scaled Sampling for Monocular 3D Body Tracking, *CVPR 2001*, vol 1, pp. 447-454.

[13] C. Sminchisescu and B. Triggs. Kinematic Jump Processes for Monocular Human Tracking, *CVPR 2003*, vol.1, pp. 69-76.

[14] C.J. Taylor. Reconstruction of articulated objects from point correspondences in a single uncalibrated image. *CVIU 80(3)*: 349–363, December 2000.

[15] P. Viola, M. Jones. Rapid object detection using a boosted cascade of simple features, *CVPR 2001*, vol.1, pp.511-518.

[16] S. Zhu, R. Zhang, Z. Tu. Integrating bottom-up/top-down for object recognition by data driven Markov chain Monte Carlo, *CVPR 2000*, vol.1, pp.738 -745

[17] T. Zhao, R. Nevatia. Bayesian Human Segmentation in Crowded Situations, *CVPR 2003,* vol.2 pp.459-466.