High-Level Analysis of Human Pose

Gregory Shakhnarovich & Trevor Darrell

Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

http://www.ai.mit.edu

The Problem: Consider a photographic image of a person. It is usually easy for us to point to the arms, the legs, or the head, and to determine the person's rough overall pose. In most cases, when presented with a sequence of such images, we can easily recognize familiar activities (walking, waving hands, kicking). We would like to make a computer be able to perform such recongition and analysis tasks.

Motivation: For many applications of human body pose analysis, only high level reasoning is involved; that is, it is not necessary to infer the precise 3D location of the body parts (or of idealized parameters such as "joints"). Rather, the reasoning seems to refer to the relative position and motion of body segments. We hope to improve the performance of analysis applications by extracting the high-level information directly.

Previous Work: Analysis of a single image of an articulated object has been the subject of relatively few investigations. Most notable recent advances in this direction are the work of Rosales, Sclaroff *et al* [4], C.J.Taylor [5], and Malik and Mori [3]. Much more work has been done on the analysis of articulated motion; a comprehensive recent survey can be found in [2].

Approach: We believe that for many purposes, high-level reasoning about pose and activity does not necessarily require the precise location of joints or edges, but rather correct qualitative information about the location of the parts and their relative positions. We therefore forgo the simplified model representation (cylindrical, ellipsoidal, etc.) and directly label the pixels as belonging to specific parts of the articulated structure, or to the background. We are currently working in simplified settings, whereby we assume that the silhouette (albeit noisy) of the person is known (as in Figure 1(a)).

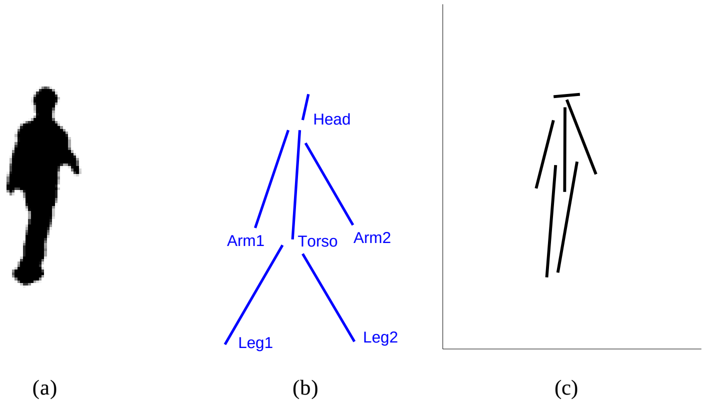

Figure 1: Input (a), model structure (b) and the resulting model (c) estimated from the input.

We propose a high-level representation of body segments, which is somewhat reminiscent of medial axis transform. In the simplest illustrative case, the model of a human consists of line segments associated with the main body parts (Figure 1(b)). Note that the segments are not necessarily attached – this is in fact a "loosely" articulated model.

The relationships between the parts are represented in terms of joint probability distributions of endpoint locations. Such a prior distribution can be obtained from the known anthropometric data, or learned from examples; we are pursuing the latter approach. Each pixel of the silhouette can then be labeled as belonging to a particular segment.

Figure 1(c) shows an example of a model inferred from the input silhouette in Figure 1(a). This result was obtained by iteratively finding the model aprameters Maximum A-Posteriori probability of the model given the estimated body part labeling, and then relabeling the body parts – a standard application of the Expectation Maximization algorithm (EM) [1]. Some examples of final labelings are shown in Figure 2.

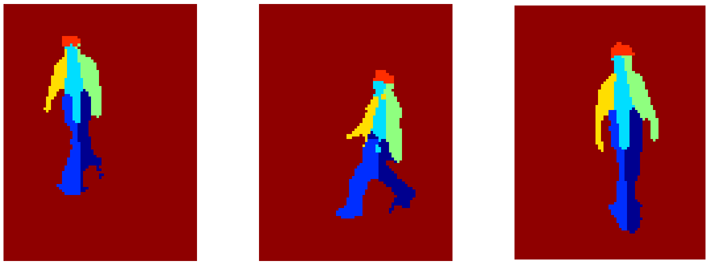

Figure 2: Body part labeling after 4 iterations of EM, using MAP model.

Impact: This work would make high-level, semantically meaningful information about the depicted object available for various reasoning applications: activity classification, recognition, gesture analysis etc. Analysis of a single image is a particularly challenging, largely unsolved problem, and progress in that direction would benefit many applications, for instance content-based retrieval from image databases, with queries phrased in terms of high-level descriptions of object pose.

Future Work: The assumption of a known silhouette of the subject is unrealistic; work remains active in the area of image segmentation, and we can not rely on the state-of-the-art segmentation algorithms to provide us with reliable silhouettes. Removal of this dependence on the silhouette would make our methodology much more powerful, and this is the main focus of our current work on this topic. In particular, we are looking into finding body segments in a cluttered image using an edge map representation.

Another important shortcoming of the simplistic approach presented here is the impossibility of representing occlusions, or in the more general case allowing for missing parts of the model. Development of a probabilistic framework that would naturally include such cases is an important future direction.

Finally, the linear segments of which our model currently consists unnecessarily limit its power; we plan to address this by introducing non-linear segments (e.g. curves instead of straight lines).

Research Support: This work is supported by the DARPA HumanID project, and by project Oxygen at MIT.

References:

[1] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern classification*. John Wiley & sons, New York, second edition, 2001.

[2] D. M. Gavrila. The visual analysis of human movement: A survey. *Computer Vision and Image Understanding*, 73(1):82–98, January 1999.

[3] Greg Mori and Jitendra Malik. Estimating Human Body Configurations using Shape Context Matching. In *Proceedings of European Conference on Computer Vision*, Copenhagen, Denmark, 2002.

[4] R. Rosales and S. Sclaroff. Specialized mappings and the estimation of body pose from a single image. In *IEEE Human Motion Workshop*, pages 19–24, Austin, TX, 2000.

[5] Camillo J. Taylor. Reconstruction of articulated objects from point correspondences in a single uncalibrated image. *Computer Vision and Image Understanding*, 80(3):349–363, December 2000.