Multi-view Body Part Recognition with Random Forests

Vahid Kazemi
vahidk@csc.kth.se
Magnus Burenius
burenius@csc.kth.se
Hossein Azizpour
azizpour@csc.kth.se
Josephine Sullivan
sullivan@csc.kth.se

CVAP / KTH
The Royal Institute of Technology
Stockholm, Sweden

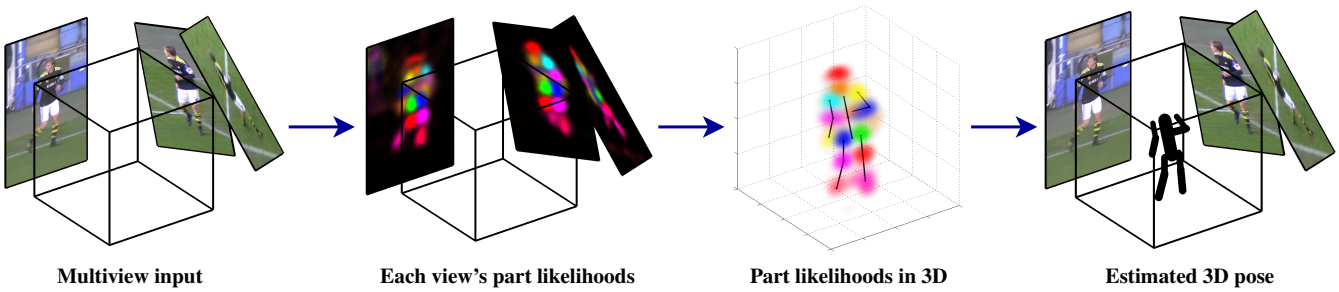

Figure 1: Overview of our 3D pose estimation framework. A random forest is first used to classify each pixel in each image as either belonging to a part or the background. The results are then back-projected to a 3D volume to get the 3D likelihood maps for each part. We find corresponding mirror symmetric parts across views by introducing a latent variable. Finally, a part-based model is used to estimate the 3D pose.

In this paper we address the problem of automatically estimating the 3D pose of a person seen from multiple calibrated cameras outside a studio environment [2]. Our particular focus is the estimation of the 3D pose of football players during a professional game. Football footage has several key characteristics some of which are shared between different sports. Most notably the images are commonly disturbed by motion blur because of the fast moving players and cameras. There is also a large variation in the players' 3D pose. On the other hand the variation in the players' clothing is limited and background clutter is not as severe as in less structured environments. Yet, low quality images and fast motion makes it hard to perform background subtraction reliably.

State of the art 2D part based models for human pose estimation rely on SVM classifiers applied to a HOG descriptor of an image patch[4]. However we opt to use a more efficient random forest approach for estimating the part likelihoods. We take our inspiration from the recent success of the *Kinect* system. Shotton *et al*. [3] use a random forest to estimate a person's 3D pose from a depth image. In this paper we consider ordinary visual images, as opposed to depth images, but similarly use a random forest to assign to every pixel a probability of being either a particular part or a background pixels. These probabilities form the basis for our part likelihood scores in 2D and 3D.

We create 3D part appearance likelihoods by aggregating the 2D likelihoods across all camera views. Care must then be taken regarding the correspondence of joints across the views. Because of the similar appearance of mirror symmetric parts, such as left and right arms and legs, and also the local nature of our part detectors, we can not directly distinguish the correct correspondences for each part. In this paper this issue is handled by introducing a latent variable into our model which represents the correspondence. At inference time we optimize for both the best pose and the best values of our hidden variable. We show that this approach is both feasible and effective and allows the 3D inference to perform well in the cases when the 2D likelihood maps are accurate.

To perform 3D pose estimation we follow the approach of [1] and discretize the search space. We assume that the person is within a bounding cube (fig. 1) and create a $64 \times 64 \times 64$ grid covering this cube. We evaluate our random forest based 3D appearance likelihoods (RF) for all grid points. We perform inference with and without the pose prior, imposing limb length and intersection constraints. We also perform inference with and without the latent variable, which handles the mirror ambiguity.

Figure 2 shows our estimated 3D poses (red) compared to the ground truth (blue), for four different frames. For this figure the inference was performed using the latent mirror variable but without any pose prior (uniform). The figure shows that our 3D appearance likelihoods accurately detect most of the body parts, even without imposing any pose prior. If we add the limb length and intersection constraints we are able to correct for some of the limited double counting that occurs for the lower legs.

[1] M. Burenius, J. Sullivan, and S. Carlsson. 3d pictorial structures for multiple view articulated pose estimation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2013.

[2] T.B. Moeslund, A. Hilton, V. Krüger, and L. Sigal. *Visual Analysis of Humans: Looking at People*. Springer, 2011. ISBN 9780857299963.

[3] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from a single depth image. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2011.

[4] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2011.

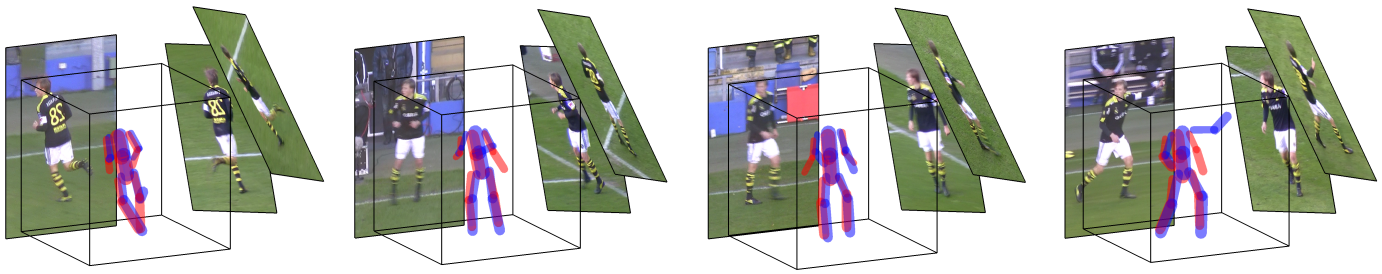

Figure 2: Example results. Final 3D poses obtained by taking, for each part independently, its most probable state over the grid. The mirror ambiguity is solved jointly.