Skeletal Graph Based Human Pose Estimation in Real-Time

Matthias Straka
straka@icg.tugraz.at
Stefan Hauswiesner
hauswiesner@icg.tugraz.at
Matthias Rüther
ruether@icg.tugraz.at
Horst Bischof
bischof@icg.tugraz.at

Institute for Computer Vision and Graphics
Graz University of Technology
Inffeldgasse 16/II, 8010 Graz
Austria

http://www.icg.tugraz.at/

Human pose estimation is a vivid topic in current literature due to its wide-spread applications such as motion-capture, telepresence or object manipulation in virtual environments. The process of human pose estimation is concerned with finding the pose parameters of a human body model that best fit to the observations in one or more input images. There exists a variety of algorithms that solve this task with high accuracy from multiple input images, depth images or even a single photograph. Unfortunately, these systems often require manual initialization and cannot process images at interactive frame rates. Often this can be overcome by learning poses from thousands of examples or fitting a rigid body part model to the data. However, model fitting algorithms are easily distracted by missing or spurious body-parts and depend on a good initialization. Therefore there is need for improvement in real-time body pose estimation methods to handle the full articulation space of the human body, support automatic single-frame initialization and tolerate outliers.

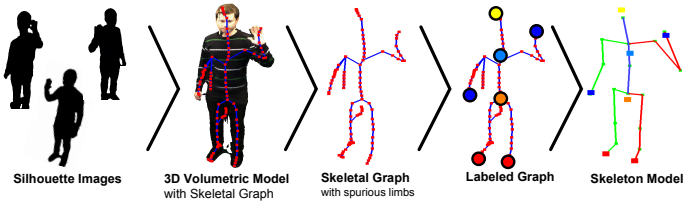
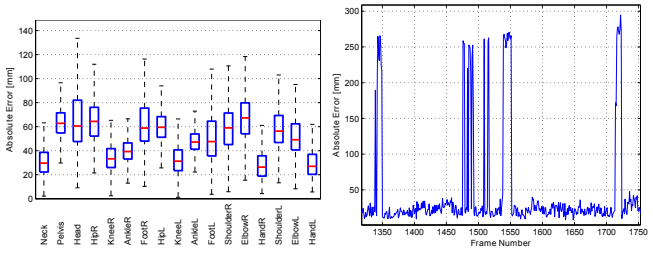


Figure 1: Extracting the human skeleton starting from silhouettes.

In this paper, we present a novel marker-less human pose estimation algorithm which uses a skeletal graph extracted from a volumetric representation of the human body. The skeletal graph is a tree that has the same topology as the human body (i.e. arms, legs and body). We generate this graph efficiently using a center-line tracing algorithm [3] applied on voxel data. As the center-line extraction produces spurious branches, we employ a novel pose-independent graph matching algorithm based on [1] which robustly labels graph end-nodes into head, hands and feet while ignoring such end-nodes that do not correspond to any valid limb. Our graph matching uses geodesic distances between each end-node of the graph and the head-node as features and matches them using dynamic programming. The labeled nodes allow for a good initialization of a human skeleton model. We optimize this model using a fast local optimization similar to [2] and ensure that all joints lie close to the skeletal graph and bones maintain the correct lengths. A graphical summary of our algorithm is shown in Figure 1.



(a) (b)

Figure 2: Statistics of the position estimation error for all joints (a). Distance between the estimated hand position and ground-truth over time (b).

In order to evaluate our approach, we render a polygon model that is animated by motion capture data from multiple views. This is how we obtain silhouettes and ground-truth joint positions in almost twenty-thousand frames. In Figure 2(a) we show that we are able to achieve an average joint position estimation error of around 50 mm in all of these frames. We do not perform any post-processing or tracking of joint positions over time but emphasize the importance of single-frame detection. Compared to systems that rely on tracking information, we can provide long-term skeleton tracking and do not get stuck in a wrong pose for a long time. Figure 2(b) shows how our algorithm automatically recovers from false estimations of the left hand position.

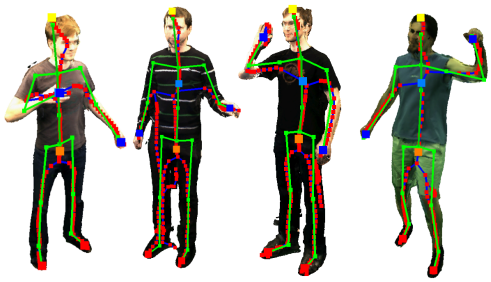


Figure 3: Examples of human pose estimation on real data.

We demonstrate the quality of our graph based tracking in Figure 3 where we show an overlay of the 3D human body, the skeletal graph (red) as well as the fitted skeleton (green) in various poses. Our pose estimation algorithm can be combined with a multi-view camera setup that is able to process ten video streams in real-time and allows for segmentation of the user in all camera views [4]. Using such silhouette images, the visual hull is generated efficiently through space carving. Image capturing and processing can be performed in real-time on a single computer using a state of the art GPU and CPU. On the same system, the visual hull can be generated within 10 ms, the skeletal graph is extracted in 6 ms and graph matching can be performed in less than one millisecond. This allows for human pose estimation interactively at up to 30 frames per second.

The key benefits of our method are the robustness of limb-labeling and its ability to perform frame-by-frame pose estimation at a low computational cost due to early reduction of the input data. More precisely, we reduce the amount of data from roughly 10^6 voxels to a skeletal graph which consists of merely 10^2 connected nodes. We do not require any learning phase nor a database with training images for human pose estimation, which makes our algorithm particularly easy to implement. Future work involves the integration of temporal tracking that aids the limb-labeling stage but avoids getting stuck in erroneous poses.

Acknowledgements: This work was supported by the Austrian Research Promotion Agency (FFG) under the BRIDGE program, project #822702 (NARKISSOS).

[1] X. Bai and L. J. Latecki. Path similarity skeleton graph matching. *Pattern Analysis and Machine Intelligence*, 30(7):1282–1292, 2008.

[2] I. Baran and J. Popović. Automatic rigging and animation of 3D characters. In *Proc. of the ACM SIGGRAPH*, 2007.

[3] A. Rodriguez, D. Ehlenberger, P. Hof, and S. L. Wearne. Three-dimensional neuron tracing by voxel scooping. *Journal of Neuroscience Methods*, 184(1):169–175, 2009.

[4] M. Straka, S. Hauswiesner, M. Rüther, and H. Bischof. A free-viewpoint virtual mirror with marker-less user interaction. In *Proc. of the 17th Scandinavian Conference on Image Analysis*, 2011.