Evaluating Lookup-Based Monocular Human Pose Tracking on the HumanEva Test Data

Nicholas R. Howe
Computer Science Department
Smith College
Northampton, MA 01063
nhowe@cs.smith.edu

Abstract

This work presents an evaluation of several lookup-based methods for recovering three-dimensional human pose from monocular video sequences. The methods themselves are largely described elsewhere [1, 2], although the work presented here incorporates a few minor enhancements. The primary contribution of this work is the evaluation of the results on a data set with ground truth available, which allows for quantitative comparisons with other techniques.

Methods relying upon silhouettes produced via background subtraction tend to act as a "straw man" in relation to the current state of the art; many recently proposed techniques work without reliance upon background subtraction and cite this feature as one of their advantages. Without disputing such reasonable claims, this work seeks to push the envelope for background-subtraction methods as far as possible. The goal of this effort is to provide a challenging baseline against which the performance of various alternatives may be assessed.

1 Background Subtraction

Given the commitment to background subtraction made here, the quality of the extracted silhouettes will strongly affect the result. Although the reconstruction methods may cope to some extent with noisy silhouettes, for the strongest comparison the silhouettes should be nearly error-free. Fortunately, recent work has demonstrated that graph-based techniques can extract high-quality silhouettes both reliably and quickly in most cases [5]. The foreground segmentation adopted here is based upon a different implementation with a similar philosophy, as detailed in the items below.

- The trim mean gives a robust Gaussian model of the background color at each pixel. For clips where the subject moves around sufficiently, the background model can be estimated directly from the action video.
- For a given frame, the number of standard deviations from the background color model at each pixel guides the foreground segmentation. For color images, separate models are developed for hue, saturation, and value at each pixel. A linear combination of the results of the three models weights each one according to its reliability. (The hue channel shows greater noise even after normalizing by the standard deviation, and is consequently weighted less than the other two.)
- To mitigate shadows, the model forgives luminosity decreases of up to τ_s from the computed background luminosity. This accounts for possible darkening due to shadows, and typically improves the segmentation where the subject's feet meet the floor. Occasionally, it may improperly label foreground regions as background if their color is slightly darker than the background region they occlude.

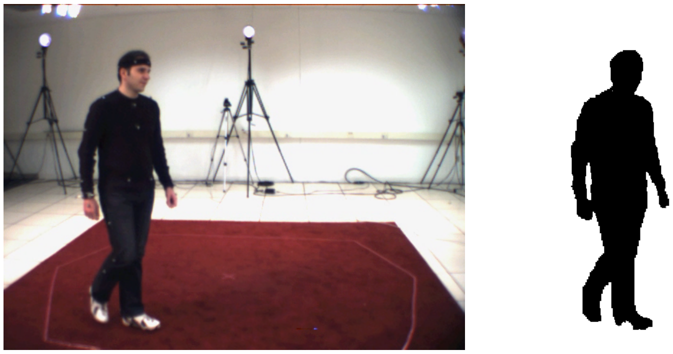

Figure 1: Example of typical foreground segmentation result. The precise boundaries and separation of body parts make further pose recovery steps easier. Nevertheless, errors can occur where there is poor contrast between the subject and the background, as in the darker portion of the left shoe.

- The minimum cut on a graph constructed from the image gives the foreground segmentation. Both four- and eight-connected neighbor edges are included in the graph, with weaker links to diagonal neighbors so that the solution favors neither straight nor diagonal boundaries. The graph omits neighbor edges where gradients appear in the frame that are not present in the background image. This strategy biases the foreground segmentation to follow object boundaries in the image. Figure 1 shows an example of a segmented frame.

2 Method

The method employed in this work is called *lookup-based* because it recovers 3D pose by retrieving perceptually similar examples from a previously computed collection. A small set of poses selected for each frame forms the basis of the next processing stage, which imposes a temporal Markov condition to select a sequence of poses with small momentum changes and high fidelity to optical flow observations. (This "stitching" process begins with 10 to 50 candidates per frame, depending on their similarity score.) Finally, smoothing the pose trajectories over time eliminates any remaining jerkiness.

The lookup stage may use any number of retrieval methods to generate the initial candidate pool, and this work combines several. It begins with three basic similarity measures, computed from the image silhouette and optical flow: moments of the image optical flow[2], turning angle[1], and half-chamfer distance. (Unlike full chamfer, half-chamfer can be computed in a small number of calculations.) The candidate pool includes poses that score the top rankings on all three measures, on the flow moments alone, and on the two shape measures together. Some are constrained to lie near poses in the candidate pool from the previous frame, ensuring continuity; a handful are chosen without this restraint to allow for recovery from errors. Although up to 65 candidates may be selected, overlap between the different retrieval modes means that the actual number will usually be lower. These candidates are registered to the image frame using a full chamfer match on the silhouette, and the most successful registrations compete for selection in the next stage.

The temporal chaining step employs a two-state stochastic chain model (sometimes called a second-order hidden Markov model) to find the sequence of states minimizing a global energy function. The terms in this global function represent three balanced concerns: the match of each candidate pose to its frame, the momentum change embodied in each three-frame section (including estimates of both linear and angular momentum), and the match of the implied optical flow to the flow observed in the image. Previous work on shorter video clips has suggested that including the latter two constraints in the energy function can prevent limb-swapping errors that sometimes appear when one leg crosses in front of another[2]. The results reported here suggest that they are not entirely successful, and further study of this issue may be warranted. If limb-swapping is suppressed via other means (for example, a state transition model for walking [3]), then a simpler formulation is possible that omits flow comparisons and employs only a first-order HMM. Nevertheless, the results presented here all use the more complex model described initially.

The smoothing stage of processing is mostly cosmetic, and can in some cases increase the error in comparison to ground truth. The smoothing treats individual pose parameters as one-dimensional functions over time, and removes any high-frequency variation. Previous work employed a method that modeled overlapping subsections via quadratics and blended the results together[1]. The results here instead perform a low-pass smoothing by zeroing high-frequency components of the DCT, but the result is quite similar.

3 Experiments

The HumanEva test set [4] forms the basis of these experiments. This abstract presents results only for a few selections from the walking sequences. We expect to process additional sequences as time permits.

Processing proceeds in several stages. The first stage computes a foreground segmentation and optical flow field for each frame. The second stage uses the foreground silhouette and the moments of the optical flow to retrieve matching silhouettes with known poses from a collection derived from the CMU motion capture library. For greater reliability, this stage performs several retrievals employing different combinations of similarity measures: flow moments and turning angle and half-chamfer matching all together, flow moments alone, and turning angle and half-chamfer matching together. In addition, some of the retrieved poses are constrained to remain close to the best matches from the previous frame, to ensure continuity in difficult regions. All the retrieved poses (at most 65) are pooled together, with any duplicates removed, and registered to the image frame. The third stage applies a second-order Markov chain to identify a temporal sequence of poses with low-energy transitions. Good sequences have small positional changes from frame to frame, and their calculated flow matches the flow observed in the frame. A final stage can smooth the temporal pose sequence to suppress jerkiness. However, this step can actually increase the error as compared to ground truth, so it is omitted here.

Reconstructed poses are registered to the video frame and therefore comparable to the known ground truth in the HumanEva suite. The analysis below computes the mean error over all the evaluation points, plus the mean error for each point over all the frames.[1]

The results reveal one glaring shortcoming of the method. Silhouette-based reconstructions are known to suffer from an ambiguity where the pose is inverted along the line-of-sight axis and the left and right sides of the body are exchanged. Such a transformation produces a physically reasonable pose with exactly the same silhouette as the original. In previous work the inclusion of flow data suppressed this problem [2], but it occurs often in the HumanEva sequences despite the flow constraints. One easy way to solve this problem would be to introduce a walking-specific motion model that requires alternating strides, as some have done[3], but this would compromise the generality of the lookup-based approach. Although this matter clearly invites further research, for the moment the results below present both the raw error (which may be quite high due to inversion of the recovered pose) and the minimum error chosen between the recovered pose and its inversion. Labels in the text will endeavor to indicate clearly which is which.

Indicative results appear here for the S3/Walking_1_(C2) and S3/Walking_1_(BW2) videos. The choice of subject and camera are not particularly significant, except that the background subtraction contains no extreme errors. Figures 3 and 2 show some statistics on the results, organized by frame and by evaluation point. The overall mean errors after allowing for pose inversion are 14 and 11 pixels, respectively. However, the figures reveal considerable variation. The individual frame errors show a standard deviation of nearly 9 pixels. The mean joint errors ranged from a low of 8 pixels for the pelvis to as much as 29 pixels for the extremities of the arms. The legs, despite their greater length, show lower mean error than the arms because they are more often visible in the silhouette.

The results presented in these figures have not been optimized in any way from the silhouettes selected by the optimal Markov chain. Naturally, the retrieved poses will not fit the observed motion exactly, and should allow further improvement via numeric optimization. A very preliminary experiment indicates the truth of this hypothesis: one pass of frame-by-frame optimization

[1] For the videos, visual inspection revealed what appeared to be an error in the ground truth data for the top of the head over many frames; consequently this point was omitted from the frame-by-frame analysis.

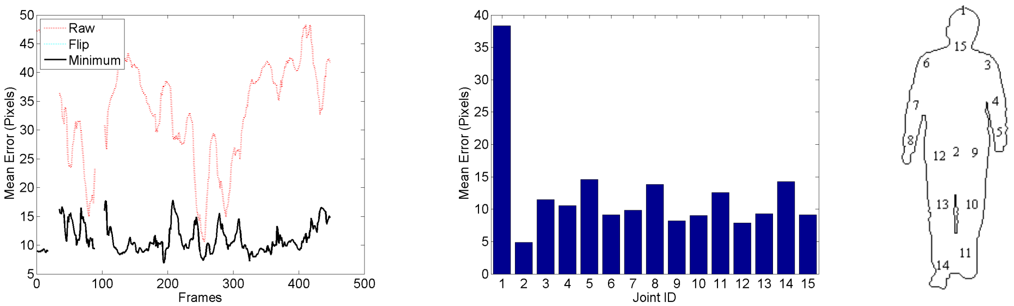

Figure 2: Summary of results for the S3/Walking_1_(BW2) sequence. At left, the red dotted line shows the error of the reconstructed pose. The solid line shows the minimum error of the reconstructed pose or its inversion (cyan dotted). Higher frame error in this sequence appears correlated with sideways walking. At right, the bars show the mean error for each evaluation point over all the frames. They show that error increases towards the extremities. The large error in the first point incorporates apparently erroneous data.

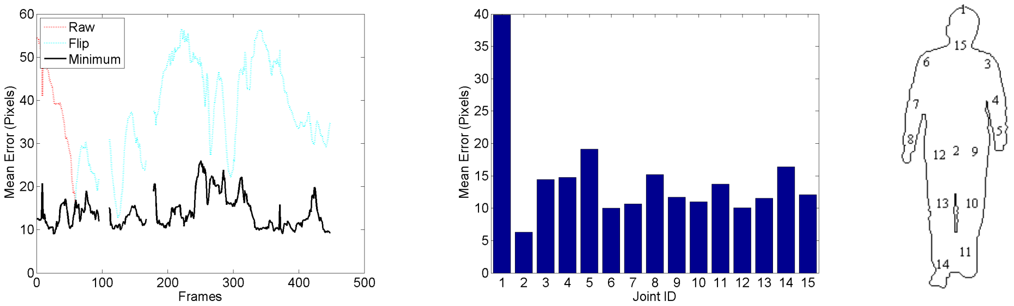

Figure 3: Summary of results for the S3/Walking_1_(C2) sequence. At left, the red dotted line shows the error of the reconstructed pose. The solid line shows the minimum error of the reconstructed pose or its inversion (cyan dotted). The larger error around frames 250-310 may be correlated with instability in the background subtraction. At right, the bars show the mean error for each evaluation point over all the frames. They show that error increases towards the extremities. The large error in the first point incorporates the apparently erroneous data.

using chamfer-distance and temporal smoothness criteria reduces the mean ground truth error in the S3/Walking_1_(BW2) video from 11 to 10 pixels.

4 Conclusion

The numeric results show considerable room for improvement, either through refinements of the silhouette lookup method or through completely different approaches. These results are presented in the hopes that they will add to understanding of the state of the art, and spur further research in the field. It is our intention to add results on additional HumanEva sequences as soon as possible.

Acknowledgment

This manuscript is based upon work supported by the National Science Foundation under Grant No. IIS-0328741. The data used in this project was obtained from mocap.cs.cmu.edu. The database was created with funding from NSF EIA-0196217.

References

[1] N. Howe. Silhouette lookup for automatic pose tracking. In *IEEE Workshop on Articulated and Nonrigid Motion*, 2004.

[2] N. Howe. Flow lookup and biological motion perception. In *International Conference on Image Processing*, 2005.

[3] X. Lan and D. Huttenlocher. A unified spatio-temporal articulated model for tracking. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume I, pages 722–729, 2004.

[4] 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, September 2006.

[5] Y. Sun, B. Yuan, Z. Miao, and C. Wan. Better foreground segmentation for static cameras via new energy form and dynamic graph-cut. In *ICPR (4)*, pages 49–52, 2006.