On the structure of ambiguity in reconstructing human motion over multiple frames

Akash Kushal and Mithun Das Gupta

Final Project Report
CS598daf - Human Tracking

Abstract. This report presents an empirical study on the structure of the kinematic ambiguities that arise when reconstructing human motion from monocular video sequences. In particular, we look at the advantages of working with multiple frames of video in reducing the kinematic ambiguity. We artificially generate the possible ambiguous configurations for an arbitrary viewing direction and use anatomical constraints such as joint angle limits and body non self-intersection constraints to prune them. A simple 2nd order linear dynamical model is used to connect up the ambiguous configurations in neighboring frames. Possible ambiguous motion sequences are then generated by searching for paths in these configurations through the motion frames. We observe that the number of ambiguous motion sequences is considerably reduced by imposing continuity constraints via the dynamical model. However, removing the ambiguity entirely is not possible for most motions. Also, the number of ambiguous paths that remain is highly correlated with the orientation of the observed person with respect to the camera. We validate these results by performing multiple experiments on a motion database.

1 Introduction

Reconstructing 3D human motion from a monocular image is ambiguous for several reasons. In many cases it becomes hard to correctly label the limbs of the human figure. For example, in a profile view of a walking person it is hard to determine which hand and leg (left or right) is in front and which is at the back. Similarly, in a frontal view of a person it may be hard to determine whether the person is facing toward the camera or away from it without using a face detector.

Even if we assume that we can resolve this limb labeling ambiguity in some way and can generate accurate joint locations in the image, the 3D body configuration still remains ambiguous. This is because, each limb may be pointing toward the camera or away from it and in both cases the projection of the joints into the image will remain the same. Since each limb could be flipped toward or away from the camera independent of the others the total number of ambiguous configurations becomes extremely large. Many of the resulting configurations get ruled out because they are anatomically impossible. However, many still remain if one is dealing with a single image.
If however one is dealing with tracked image sequences, many configurations that were possible in a single frame may now become impossible. This happens because there may be no path through the possibility space from the first frame to the last frame passing through that configuration. Indeed, Howe et al. [1] claim to overcome this ambiguity when using motion snippets of 11 frames each to reconstruct the corresponding motion instead of just single frames. However, they only claim to achieve correct tracking when motions similar to the one being tracked are present in the training set. Hence, it might be the case that the training set of motions was not large enough to contain many ambiguous motions and since the test motions were similar to some motion in the training set no wrong tracks were generated. Using motion snippets will clearly reduce the number of ambiguous configurations but it is not clear if using motion snippets can completely disambiguate many motions.

The goal of this project is to determine how much reduction in the number of possible configurations can be achieved by looking at multiple frames of motions together instead of single frames. We start by artificially generating all the ambiguous configurations for a motion capture snippet that arise when the motion is viewed from an arbitrarily selected viewing direction. This generates a set of possible body configurations for each frame of the motion snippet which are then pruned using the joint angle limits and body non self-intersection constraints. The above constraints are imposed as hard constraints and refine the set of possible ambiguous configurations generated for each frame. A dynamical model is used later to link up the ambiguous configurations in three consecutive frames that could have realistically followed each other. Possible ambiguous motions correspond to paths in the obtained graph.

2 The Dynamical Model

We use a human motion capture database to learn a 2$^{nd}$ order linear dynamical model (similar to [2]). The regression model learnt is of the form

$$\hat{x}_n = Ax_{n-1} + Bx_{n-2}$$  \hfill (1)

Here $A$ and $B$ are the parameters of the model and $\hat{x}_n$ denotes the predicted state at time $n$. Following [2], we do not estimate this model by directly minimizing $||\hat{x}_n - x_n||^2$ since the number of parameters to estimate (entries of $A$ and $B$) is quite large. Instead, we rewrite the model in a different form and add a regularizing term to the objective function as shown in equation 2. This regularization penalizes deviations from a base model $\hat{x}_n = 2x_{n-1} - x_{n-2}$.

minimize

$$||\hat{x}_n - x_n||^2 + \lambda(||P||^2_{frob} + ||Q||^2_{frob})$$

where

$$\hat{x}_n = (I + P)(2x_{n-1} - x_{n-2}) + Qx_{n-1}$$  \hfill (2)
Initially, we tried to fit the above model over the state space of joint angles (again following [2]). The joint angles are represented as triplets of Roll, Pitch and Yaw (RPY) angles. The model learnt in the joint angle space did not work well and the prediction errors for some angles were of the order of 60° for some frames of motion in the database. A closer examination of the offending frames showed that there were unreasonably large variations in the joint angle values between those frames and their predecessors. This difference however was not being reflected in the displayed motion as any sudden motion of the bone connected to the joint. Further investigation revealed that the reason for the failure of the joint angle based model in these cases is due an inherent extreme non-linearity of the the RPY angle representation near certain rotations.

2.1 Ambiguity in the Roll-Pitch-Yaw (RPY) representation

The RPY angle representation suffers from various forms of ambiguities/degeneracies that create problems for the dynamical model.

1. Dual Representation Ambiguity: This ambiguity exists due to the fact that the rotation corresponding to the triple \((R, P, Y)\) is the same as the rotation corresponding to \((\pi + R, \pi - P, \pi + Y)\) for all \(R, P, Y\).

2. \(\pi/2\) Degeneracy: The \((R, P, Y)\) representation becomes degenerate around \(P = \pi/2\) and any small change in rotation around this value will result in large deviations in the \((R, P, Y)\) representation. The direction of maximum deviation lies along the line \(R + Y = constant, P = constant\). In other words, changing \((R, P, Y)\) along this line will result in negligible change in the rotation.

Apart from these ambiguities there is also the standard \(2\pi\) ambiguity which is there for all angle representations.

Figure 1(a) shows pictorially the different ambiguities in the RPY representation. All points in the same color are randomly sampled angles that lie within...
5° rotation of a base angle. This true rotation was computed by converting the relative rotation to a quaternion representation. The $R$ and $Y$ angles for the base rotations are kept fixed and their $P$ values are varied from 60° to 120° in intervals of 10°. As the $P$ get closer to 90° the scatter around the base angle increases and covers the entire $2\pi$ range along the $R + Y = constant$ line at $P = 90°$. One can also notice the Dual-Representation ambiguity mentioned above from the plot. The scatter plot for each base angle is composed of two separate clusters, separated by a $\pi$ in both $R$ and $Y$ and flipped about $P = \pi/2$. Finally, the $2\pi$ ambiguity in the representation of angles also shows up in the plot and the scatter cloud for $P = 90°$ is split into 2 clouds because of this.

It turns out that one can preprocess most motions to correct for the Dual-Representation ambiguity and the $2\pi$ angle ambiguity (described later in section 3). However, the $\pi/2$ Degeneracy results in the dynamical model not working well. Figure 1(b) shows the degeneracy of RPY in a single motion around $P = \pi/2$ (the dual-representation ambiguity and the $2\pi$ circularity has been corrected). There is a sudden change in the RPY value along the ambiguous direction $R + Y = constant, P = constant$ when $P$ gets close to $\pi/2$.

Hence, we decided to learn the dynamical model in the space of the relative 3D location of the bone endpoints for all the bones in the model instead of RPY-angles. More precisely, we combine the difference of the 3D coordinates of the endpoints of all the 14 bone segments in our model to form a 42 dimensional vector. The dynamical model predicts the vector for the next frames configuration based on the current and previous frame vectors. Since this model uses relative joint positions it cannot ensure that the bone lengths will remain constant. However, this model suffices for our application since we only use it for choosing among different configurations.

3 Imposing Constraints on Human Pose

3.1 A simple human body model

We use the database of motions used by [3] for our experiments but simplify the human body model used before proceeding. This is done by removing some joints, such as toes and palms, which have very small projections in the motion frames. We do not consider this a serious issue, since most human motion tracking systems are not very reliable in localizing these joints in the image, and hence, the ambiguities that occur while lifting these joints to 3D are of little interest. The joint angles of the reduced body model are calculated so that the structure of the new motion remains as close to the original motion as possible. Figure 2 shows a frame from the original and the simplified motion.

3.2 Motion Constraints

Valid human configurations must satisfy a number of constraints. The two most important constraints considered by the human motion tracking community are
the joint limit constraints and the body non self-intersection constraints. We use an extensive human motion database to generate the joint angle limits, necessary to prune the possible 3D configurations. The joint angle constraints are obtained by computing the maximum and minimum values of all the joints over all the frames of all the motions in the motion database.

However, the Dual-Representation ambiguity mentioned in the previous section will also create a problem while computing the joint angle limits for joints with 3 degrees of freedom. We must first transform the angles in each motion frame so as to remove these ambiguities by minimizing the joint angle spread in RPY space. In other words, we want to flip a joint angle value between its dual representations in every frame so as to bring them all to the same representation. We employ a simple 2 stage algorithm for correcting the values of a single joint.

− In the first step we move along the frames of each single motion and correct the angle rotation in the current motion frame (either flip or don’t flip its representation) so that its RPY value is closest to the previous frame of the motion. Note that the distance between the angle values for the representations must be measured modulo $2\pi$. This process takes each single motion into a common representation. However, the motions themselves might still be in different representations.

− To correct this we compute the mean joint angles for every motion and build a complete graph with the motions being the nodes of the graph and the edge lengths being the difference in rotations corresponding to the mean angle values. Next, we compute an MST (minimum spanning tree) on this graph. Then, we pick a random node as the root and do a traversal of the MST flipping visited motions (if required) so as to bring the mean joint RPY values for the motions corresponding to the edges as close as possible. This process brings all the motions into a single RPY representation.

Finally to correct for the $2\pi$ circularity, we transform the joint angle values for all the frames in the database (by adding or subtracting $2\pi$) so that the spread of the angles is minimized. Figure 3 shows the a scatter plot of the RPY
angles for a specific joint before and after the correction procedure. The left plot shows the 2 clouds corresponding to the Dual-Representation ambiguity (the 4 point clouds on the 4 corners are actually a single cloud and are split into 4 due to the circularity around $2\pi$) and the right plot shows the single cloud obtained after correction. This cloud is used to determine the minimum and maximum joint limits.

![Fig. 3. RPY ambiguities in original motions. RPY corrected motions.](image)

The second constraint used is the body-non self intersection constraint. Each bone is assigned a thickness value and every configuration for every motion frame is checked for intersections among the bones. Finally, all the intersecting configurations are removed from the possible set for each frame.

### 4 Ambiguity Graph

This section describes the process of computing the Ambiguity Graph which encodes all the ambiguous motions possible for a given motion and viewing direction. The vertices of the graph are tuples $(C^j_f, C^k_{f+1})$ where $C^j_f$ denotes the $j^{th}$ possible configuration in frame $f$ and $C^k_{f+1}$ denotes the $k^{th}$ possible configuration in frame $f + 1$. We must use configuration tuples consisting of (previous state, current state) instead of simply the current state since we are using a second order dynamical model (described in section 2) which uses both the current as well as the previous state to predict the next state. This approach helps us separate the path which moves through the configurations $C^j_{f-1}, C^j_f, C^k_{f+1}$ from the path that moves through the sequence $C^a_{f-1}, C^b_f, C^c_{f+1}$. If we represent vertices using singleton configurations, having both these paths would imply that the crossing paths $C^j_{f-1}, C^j_f, C^k_{f+1}$ and $C^a_{f-1}, C^b_f, C^c_{f+1}$ would also be allowed in the ambiguity graph even though they might be infeasible w.r.t. the dynamical model. An edge is placed between 2 vertices $(C^j_{f-1}, C^j_f)$ and $(C^j_f, C^k_{f+1})$ if the error between the configuration $C^k_{f+1}$ and the dynamical prediction based on
$C'_{j-1}$ and $C'_j$ is small. As was mentioned earlier the dynamical model works on a vector of relative joint locations and hence the error is also measured in this space. More precisely, an edge is added between $(C'_{j-1}, C'_j)$ and $(C'_j, C'_{j+1})$ if

$$||\text{RelJointPos}(C'_{k+1}) - D(\text{RelJointPos}(C'_{j-1}), \text{RelJointPos}(C'_j))|| \leq E$$

Here, $\text{RelJointPos}(C)$ denotes the vector consisting of the difference in 3D locations for the endpoints of each of the 14 bones for the configuration $C$ and $D$ represents the dynamical model. The $<$ operator returns true iff all the entries of the vector on the LHS are less than all the entries of the vector on the RHS. Each entry of the $E$ vector is computed as the maximum value of the LHS of that entry over all the motions in the database. In other words the constraint is kept as tight as possible while allowing all the motions in the database to pass the dynamical model check. This ensures that at least one path (the correct motion) is always present in the graph and removes as many other edges as possible.

5 Experiments and Conclusions

For the graph construction experiments, we have worked on a subset of 20 motions from the database of 118 motions originally used for training the dynamical model. The ambiguity graph was constructed for 5 arbitrarily chosen viewing directions for each of the 20 motions. Figure 4 shows 2 of the ambiguity graphs obtained. The left graph was obtained for a case when the human is facing toward the camera for most of the motion. In this motion the person runs toward the camera while holding an object. The graph obtained for this viewing direction is sparse and contains 504 possible paths over the entire 80 frame sequence. The second figure shows the ambiguity graph obtained when the same motion is viewed from the side. This graph is much denser and the number of paths in this graph is about 2.5 million. This is because there are a lot more possible flips available for the legs and hands when viewed from the side than there are when viewed from the front. However, many of the ambiguities in the second case occur when the bones in the legs and the arms become nearly perpendicular to the viewing direction and hence, the difference in 3D positions of the joints between the different flipped positions is not too large. The short dense regions of the graph correspond to such motion frames.

To study the variation of the ambiguity with respect to the viewing direction we break up the motions into overlapping batches of 15 frames each and compute the multiplicative increase in the number of paths for each batch for the different viewing directions measured w.r.t. the orientation of person during the 15 frames of the batch. We choose to work with 15 frame batches because the batches formed are long enough to effectively compute the multiplicative increase statistics and are short enough so that the motion direction of the human remains constant throughout the sequence. Figure 5 shows a smoothed plot of the natural log of this multiplicative factor with respect to the the azimuth and
elevation angles for the viewing directions w.r.t. the human azimuth angle. The human orientation is estimated by averaging the normal of the chest of the human over all the frames in the batch. The plot is smoothed to be able to extract some meaningful information from the data. The blue circular points on the plot are the viewing directions (relative to the human’s azimuth) that were available in the subset of motions that we tested. The value of the points were smoothed by using a gaussian kernel with $\sigma = 10$ on each point and performing a weighted average using weights from the kernel. We observe that for most human motions the frontal viewing direction creates the minimum ambiguous paths whereas the the side views create many more ambiguous motions.

Fig. 4. Graphs for 2 viewing directions for the same motion.

Fig. 5. Analysis of paths obtained for a motion for multiple viewing directions.
References