Improving Human Pose Estimation Using Optical Flow and Morphological Constraints

Team 4
Georgia Institute of Technology

1. Introduction

Reliable human pose estimation has recently received increased attention by the research community. Already having many applications in fields such as gaming and healthcare, accurate pose estimation remains a challenge for immersive human-computer interaction. This type of research is important because activity recognition is becoming a booming computer vision topic. Activity recognition relies on accurate pose estimation, which is still far from optimal and needs to be improved. We aim to improve the current state-of-the-art Kinect skeletal tracking results from [8]. We intend to use optical flow and morphological constraints in order to remove noise from the pose estimate and fix incorrect estimations.

2. Related Work

Many different approaches have been tried to estimate human pose. Shotton et al. [8] use a single depth image to quickly predict the 3D positions of body joints. They use randomized decision forests trained on several thousands depth images. Other approaches based on RGB images usually rely on regression or body parts modeling. Several methods using recognition from models of distinguished body parts have been used in the past. One of the most popular techniques developed by Felzenszwalb and Huttenlocher [3, 4] uses histograms of oriented gradients of deformable body parts that are arranged using springs with kinematic constraints and a classifier. These tend to often fail in unusual poses. Regression has been widely used in 2D human pose estimation but tends to lack accuracy. Mori and Malik use shape context matching with a kinematic chain-based deformation model to retrieve exemplars [6]. Shakhnarovich et al. use locally sensitive hashing to interpolate indexed poses efficiently [7]. Great improvements have been made using body part classification and offset joint regression [8]. However, this approach still suffers from occlusion and noise.

3. Approach

Pose estimation using only depth and without temporal information becomes unstable when the user moves too quickly or stands too close to Kinect, although the majority of the body is still captured by the RGB camera. We combined optical flow from color data and the previous depth image to correct the 3D positions of body joints that are predicted by the offset joint regression algorithm from [8]. Our approach thus outputs corrected skeletons, i.e. it reduces the motion of skeletons that are unlikely due to occlusion or fast movement. We then use morphological constraints to further enforce that certain motions are not possible. This can detect and improve improper estimations from self occlusion or leaving the Kinect’s field of view.

3.0.1 RGBD scene flow

Dense scene flow can be computed from optical flow [9] by using past and present RGBD point clouds. Our approach involves a consecutive pair of RGBD images. First, the depth values are converted to XYZ coordinates in 3D
space by using the internal and external calibration parameters of the visible and IR cameras on the sensor to register the RGB and depth images. Then we compute dense optical flow between the two images. Previous extensions of optical flow to 3D [1] assume a continuous spatiotemporal volume of intensity values on which differentiation is then performed. In our case, however, the depth values are already known, thus enabling a much simpler calculation. We compute 2D optical flow \((\Delta i, \Delta j)\) using a consecutive pair of RGB images using Farneback’s method [2]. Therefore, for the value \(D_{n-1}(i, j)\) in the depth map at location \((i, j)\) at time \(n\), there exists a corresponding depth value \(D_n(i + \Delta i, j + \Delta j)\) in the next frame. The optical flow along the \(z\) axis, which we call \(w\), is then simply the difference between the two depths to find \(w\):

\[
w = D_n(i + \Delta i, j + \Delta j) - D_{n-1}(i, j) = z_n - z_{n-1}
\]

(1)

We also project the motion of the \(i, j\) pixels into 3D space using the camera calibration parameters. The optical flow in the \(i\) direction is defined as:

\[
\Delta i = i_n - i_{n-1}
\]

(2)

These can each be projected into the world coordinate frame using the following equation:

\[
x_n = (i_n - C_x) \frac{z_n}{F_x}
\]

(3)

where \(C_x\) is the center of projection of the depth camera in the \(x\) direction and \(F_x\) is the focal point of the depth camera in the \(x\) direction. Substituting (3) into (2) yields:

\[
u = (i_n - C_x) \frac{z_n}{F_x} - (i_{n-1} - C_x) \frac{z_{n-1}}{F_x}
\]

(4)

Since this is a real-time system, we can assume that \(w\), the difference between \(z_n\) and \(z_{n-1}\), is a small number and (4) can be reduced to:

\[
u = (i_n - i_{n-1}) \frac{z_n}{F_x}
\]

(5)

The same operation can be done for \(v\). Hence, the RGBD scene flow can be solved by combining (1) and (5) and is defined as:

\[
(u, v, w) = (i_n - i_{n-1}) \frac{z_n}{F_x},
\]

\[
(j_n - j_{n-1}) \frac{z_n}{F_y},
\]

\[
D_n(i + \Delta i, j + \Delta j) - D_{n-1}(i, j).
\]

(6)

We then smooth along the \(w\) region in order to enforce motion smoothness. Scene flow allows us to rectify the predicted position of a body joint. For each body joint, if the value of the flow at that joint is low (less than 7.5 cm per frame) and the skeletal movement is high, then the position of the body joint is predicted from its last position using the value of the flow. If the flow is low and the skeletal movement is high, we assume it is an error in the calculation of the flow due to noise and we keep the prediction made using [8]. 7.5 cm was found empirically to be the most the joint could reasonably move in real-time without causing errors.

3.0.2 Morphological Constraints

In addition to making correction with the scene flow, we impose morphological constraints on the skeleton. [8] often fails to draw an anatomically correct skeleton when body parts are occluded or in certain extreme situations such as crouching or picking up objects from the floor. To detect and eliminate such incorrect skeletons, constraints on the ratios of the lengths of body parts are imposed. In particular, the mean waist to hip ratio has been found to be between 0.76 and 0.99 [5]. This ratio can detect when hips are too close together. We also detect when the head is below the hips as this is often a case in which the skeleton is in an improper state. We also detect skeletal clustering, a phenomena that occurs when a person walks out of the Kinect boundaries or towards the Kinect. The skeletal estimation is a clustered state which we can detect using the standard deviation of the mean joint movement. When this is detected, the skeleton ceases to exist. This can be seen in Figure 3.

4. Evaluation

We took annotations of each joint location in scenes to find the skeleton ground truth. In each figure, the cyan skeleton is the one derived from [8] and the yellow skeleton is our method. In the annotated frames, the magenta skeleton is the annotated skeleton. A series of results are shown in Figure 3. Figure 2 shows a graph of the error over time between our method and the annotations and [8] and the annotations. The error per frame is computed as the sum of the euclidean distance between each joint and the annotated skeleton. Each point on the graph is found using that equation. The mean total joint error of [8] was 6.204 meters per frame counting erroneous skeletons going out of frame and 3.95 meters without counting the bad skeletons. Our method yielded an error of 3.89 meters per frame.

5. Discussion

In the annotated sequence shown in Figure 3, our method reduced the mean error per frame of [8] by 2.7 meters when counting erroneous skeletons going out of frame. Without counting clustered skeleton errors, our method reduced the mean error per frame of [8] by 67 centimeters. Our method also maintains real-time precision with a higher than 30 fps
rate. As shown by Figure 2, on average we are as good if not slightly better than the method described by [8]. We yield much better results when the person has walked out of frame and detect that the skeleton is not in the world plane anymore faster than [8]. This can be easily seen in the last two frames of Figure 3. We believe this is mainly due to our clustered constraint and morphological constraints. We believe that the scene flow gives slightly better results per normal frame by minimizing jumps in the joints that are caused by occlusion. These occlusion corrections can be seen in the left arm of the 3rd frame and the left leg of the 5th frame in Figure 3. We discovered that simple morphological and motion constraints can drastically improve [8] when a person self-occludes or moves out of the field of view. Additional work could be done using sparse optical flow on the joint locations rather than dense optical flow. The work could be further improved using a full deformable model of the human body with modeling constraints.

References