Will Whitney

Real-Time Human Pose Recognition in Parts from Single Depth Images

1. **One of the main contributions of this paper is their reformulation of a seemingly difficult problem – real-time body pose estimation – into the already well-studied problem of object recognition. The goal of object recognition in computer vision is to assign to each pixel in an image one of a predetermined set of labels (e.g., car, building, sky, etc). What are the predetermined labels in the proposed approach? In what sense is the method in this paper based on object recognition?**

   The predetermined labels are body parts, as in hand, elbow, left torso, etc. They turned the problem into object recognition in that each of the pixels in the frame is weakly classified as belonging to a particular part of the body. This drastically improved the performance and the generalizablitiy of their system, since these labels didn't depend on their surroundings.

2. **The proposed approach uses low-level image features based solely on depth information. Eq.1 computes the depth disparity between two points of the body in "world space" coordinates. (Jargon: world space is the one you're used to, a coordinate system with its origin at some fixed placed in the real world; it is distinguished from an object space coordinate system, which is simply one that has its origin at the center of an object.) What would happen if we ignored the normalization term dI(x) in Eq.1? That is, how would the value of Eq. 1 change if it were instead dI(x+u) - dI(x+v)?**

   This normalization accounts for the offset of the entire body in space. Without it, the value of the equation would be dependent on where the person is standing.

Explain why the normalization term makes the feature computed by Eq. 1 depth-invariant (i.e., the same no matter how far the person is from the Kinect sensor).

This essentially just normalizes the distance away that it finds another point by the depth of the first point. By dividing this out, the system looks for points that are about the same distance away on the body each time for defining characteristic vectors.

3. **The approach uses a random decision forest technique to train the body part classifier. The core to understanding this technique is in Eqs. 5-6. Shannon entropy, H(Q), measures the heterogeneity of a set of examples Q. It has its maximum value of 1 for a set that has an equal number of examples of each class (e.g., {+ - - ++ -+-}) and a minimum of 0 if all examples in the set Q have the same label (e.g., {+ + + +}). (If you took 6.034 you saw this when building decision trees.) What choice of ϕ will make G(ϕ) large or small? Explain how Eq.5-6 gives us a good splitting candidate.**

   This equation is saying that a split (ϕ) is selected from the set of possible splits, Q, such that the entropy of each side, left and right, of the split will be minimized (normalized by the number of elements in each side). That is, it maximizes the 'percent uniform' of the sides of the split. As such, a ϕ which most cleanly divides + into one side and - into the other will maximize G.