Automatic Photo Pop-Up – Hoiem, Efros, and Hebert, SIGGRAPH 2005

I think the most significant contribution in this paper is the way that a large set of different image features are used to incrementally build up knowledge about the scene. It seems like most works on problems like this tend to pick some particular type of image feature that they are going to use, and try to come up with a complicated way of using that feature alone to compute the desired result (in this case, a segmentation). Like how graph cuts are used in many of the segmentation-based papers we have seen. The algorithms are very 'clever' but the underlying assumptions are so simplistic that it's not hard to imagine how they are going to fail.

In this work they take the opposite approach, using a large set of dead-simple features and throwing them all into a standard (and perhaps somewhat 'boring') classifier, namely boosted decision trees. This classifier figures out which are the best features to use in each situation, filtering out any features that don't provide useful discriminative information. This reduction simplifies the problem computationally, and it's also kind of interesting because it tells us which of the features aren't actually useful, which might help when considering additional features to include in the classifier. The iterative pixel to superpixels to constellations build-up relying on this classifier is also very simplistic, but it works. By using so many different types of information culled from the image, one does get a sense that it's likely going to be more robust than, say, a standard segmentation technique would be. The results do seem to demonstrate this reasonably well. Even though they are trying to do something quite basic, the number of successes is honestly kind of surprising, and the failure cases aren't really unexpected - it's easy to see that they clearly violate the assumptions of the model.

One of the biggest limitations of this work is the simplistic 'pop-up' technique for converting the 2D labeling into a 3D scene. Basically the scene is segmented into a set of vertical planar segments, and the image texture is mapped onto the planes. The pop-up only considers straight edges fit to the boundary of the final non-ground region. But looking back at the superpixels and constellations, it seems clear that there is a lot more information available than just these linear boundaries. Of course, the problem is that there is no 'right' constellation, but even between the four constellations shown for the train image, there are consistent segmentation boundaries in some places, like around the left side of the train. It seems likely that one could use a RANSAC-style technique here to extract segmentation consistencies between a number of possible constellation groupings. This would produce more accurate boundaries in the image segmentation, which could be mapped onto the linear pop-up boundaries the paper is using to get more accurate edges. It could also help with the problem they have where overlapping vertical objects are assumed to be on the same vertical plane – if they were segmented, they could possibly be disambiguated and mapped to individual planes. These more accurate segmentation boundaries would also be useful if we added a bit of interaction to the system. The user could click on boundaries or vertices in the current pop-up to add edges to the current pop-up, and then the system would re-compute the 3D geometry. Combined with some image-based metrology tools, this could make it quite efficient to extract a better 3D model from the image.