Andrew Pillsbury
10/9/14
Computer Science Honors Thesis Proposal

Variable Sized Image Segmentation From Motion Using Temporal Superpixels

Andrew Pillsbury

The Problem:

In order to train computer vision algorithms to identify and track objects in videos, one must have training data with ground truth of where the objects actually are. The Superpixel Selection System (S3) gathers this data from Amazon Mechanical Turk (MTurk) by splitting a video clip into frames and segmenting those frames into superpixels, and then having workers select the superpixels that make up a given object in each frame. This gets accurate data when the given object takes up a large portion of the screen, as can be seen with the van in Figures 1 and 2.

Figure 1: Segmented image and original image.

Figure 2: Segmented image with the object (van) selected.

When the object is small, however, this method is inaccurate because often the object in question is comprised of just one or two superpixels that include more than just the object, as is seen in Figure 3.

Figure 3: The outline of the pedestrian in the box is not accurately captured by the superpixels.

One solution to this problem is to use very small superpixels (see Figure 4), but that requires many superpixels for each frame, which causes the online MTurk task to run very slowly and deters potential workers.

Figure 4: A segmentation of the frame shown above using many more superpixels. The online MTurk task starts to run slowly at roughly 200 superpixels, and this segmentation has 1,238.

The Solution:

Our proposed solution is to make the superpixels of varying size, with smaller superpixels near the edges of objects and larger superpixels in the middle of objects and in the background. This allows for a moderate number of superpixels to be used, while still giving a user enough flexibility to accurately select the superpixels that exactly make up the object.

We will base our algorithm off of the one described in the paper "A Video Representation Using Temporal Superpixels".[1] This algorithm uses Temporal Superpixels (TSPs) which track objects moving through frames, and draws edges between the TSPs at the edges of moving objects.

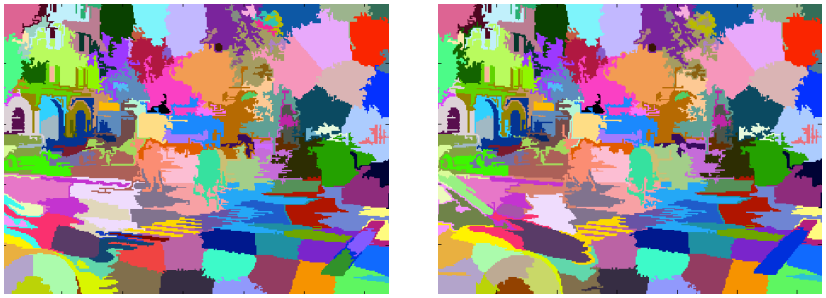

Figure 5: Results from the TSP algorithm. Note that for the most part the TSPs track the motion (or lack thereof) of the objects that they correspond to. For example, you can see that the purple TSP on the hood of the van (lower left) moves with the van from frame to frame.

The TSP algorithm provides a coarse edging, but it is not always accurate. We propose to use the TSP optical flow algorithm to roughly detect the edges of objects from motion, and then use that information to create smaller superpixels at those edges and larger superpixels elsewhere (see Figure 6).

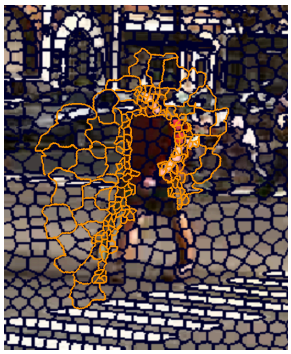

Figure 6: The yellow lines in this frame show a (partial) example of what our segmentation might look like, and the black lines are the segmentation from the TSP algorithm.

Using our new algorithm in conjunction with S3, we hope to make a robust system that will be able to quickly and accurately acquire ground truth for where precisely objects are at each frame of a video.

---

[1] Chang, J., D. Wei, and J. W. Fisher III. "A Video Representation Using Temporal Superpixels." *Proceedings of the IEEE Computer Vision and Pattern Recognition (CVPR)*. 2013. http://people.csail.mit.edu/donglai/paper/chang13_CVPR.pdf.