CSC494: Final Report

Rohan Chandra
996274142

Overview:

The algorithm discussed implements a simplified version of the Matthieu Bray, Pushmeet Kohli, and Philip H.S. Torr Posecut method discussed in the Simultaneous Segmentation and 3D Pose Estimation of Humans using Dynamic Graph-Cuts paper. The algorithm uses an automatically generated belief for the location and major axis of pose of the human as a prior for the segmentation process. The segmentation is performed as a dynamic graph cut over a graph generated from the image. In particular, a graph is constructed such that each node in the graph represents a pixel in the image. The overall minimal energy labelling can be found as the minimal graph cut of the graph generated from the image. Given this separation, it is then possible to infer a pose model that optimally describes the segmented person. The pose can be computed by finding the joint configuration that leads to the subsequent lowest energy segmentation.

Generating initial priors and subdividing the image

At higher resolutions, the generation of the graph and solving for the min flow can take an prohibitively long amount of time. Additionally, in a larger image, it is possible that multiple people appear in the scene, for which multiple segmentation and pose models are preferable. It is also likely that there are large areas of the image for which no people appear and do not necessarily need to be processed.

Thus, an initial stage of the algorithm is to locate probable regions in which a single person is located and perform the segmentation and pose estimation on these regions individually. Given the time complexity of the segmentation it is also more efficient to process smaller regions at the possible cost of having slightly overlapping regions.

Regions of interest are first discovered by computing zero crossing of the Laplacian of the image. Pixels that occur at a zero crossing of the Laplacian are expanded slightly and can then be grouped together using a connected components method. A bounding box is then generated for each connected component.

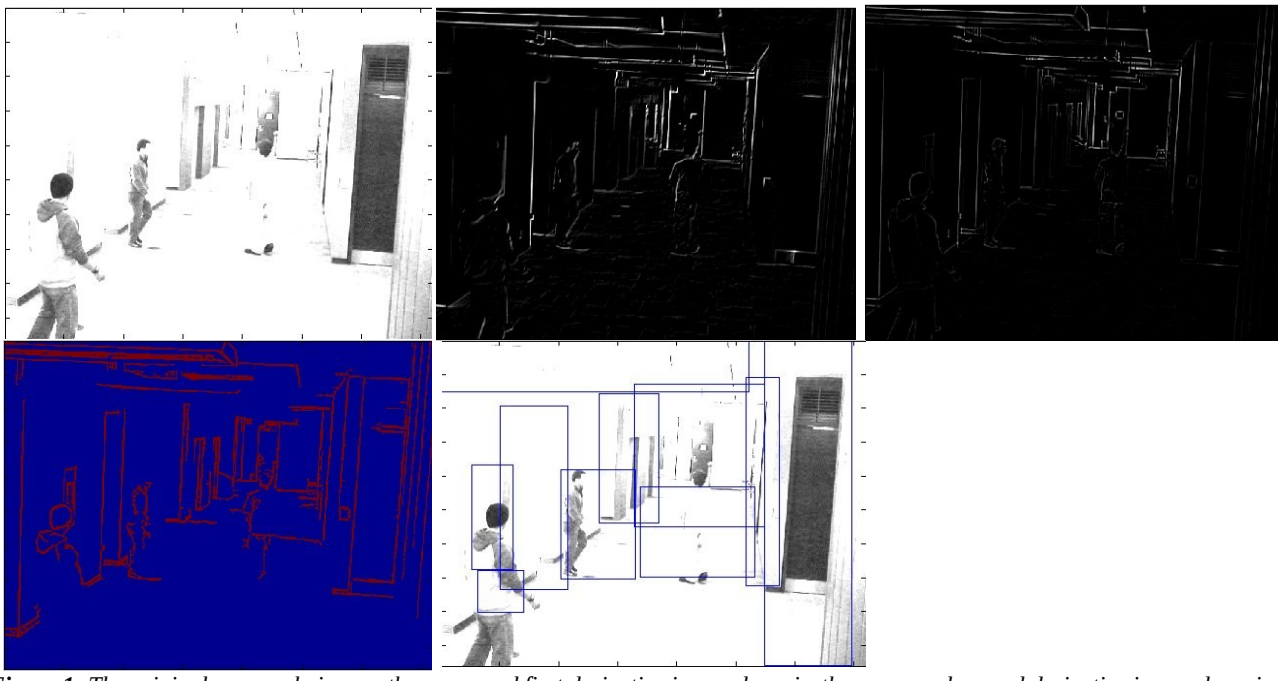

Figure 1: *The original grey scale image, the composed first derivative in x and y axis, the composed second derivative in x and y axis, the edgels produced from the zero crossings of the laplacian, and finally the regions of interest overlaying the original image*

Then, within each of these regions of interest a learned head detector is run to locate the portion within the region that has the highest probability of being a human head, should a head exist within the image. From there, the major axis of the pose of the person, i.e the spine of the individual, and its origin can be estimated.

Issues and possible solutions:

- When generating the edgels, a highly noisy image results in to many zero crossing of the Laplacian that exceed the noise threshold. This excessive amount of edgels creates a set of edgels that are close enough in proximity and sufficiently distributed across the image so as to create a region of interest in the image that approaches the original dimensions of the image. Put simply, the image may contain enough noise that the algorithm cannot create smaller regions of interest by the above mentioned process. While this is not catastrophic, it removes any benefits of the initial cropping process while still incurring the computation cost of determining the region.
  - This issue could be mitigated by determining the noise threshold of the zero crossings of the Laplacian based on the variance of pixel intensities in the image, perhaps varying in local patches.
- There are many situations in which the head detector creates false positives and can be unreliable. Particularly, the head detector gives significant response in almost any roughly circular region.
  - However, this could be remedied by using multiple part detectors and using the consensus between them to help in the case that any one part detector mislabels within the target region. A simple deformation model can be built based on a Gaussian deformation cost between where parts were located in comparison to the ideal location in which they would occur relationally to each other.

Graph creation:

Given the image, the graph is generated by creating a node for each pixel in the graph. Each nodes has an edges to its 8 immediate neighbouring pixels and to a given source and sink node. The edges between neighbouring pixels then represents the cost of setting the two neighbouring pixels to differing labels, while the edges to the sink and source represent the cost associated with setting the pixel to be in the foreground or the background respectively.

To find the segmentation, it is necessary to minimize the following energy equation

$$\Psi 3\,(x, \Theta) = \sum_{i}(\varphi(D|xi\,) + \varphi(xi\,|\Theta) + \sum_{j}(\varphi(D|xi\,, xj\,) + \psi(xi\,, xj\,)))$$

where the value of x denotes a specific labelling of a pixel and Θ represents the joint configuration for the pose.

Further, we define the unary costs as:

$$p(xi = \text{figure}|\Theta) = 1 - p(xi = \text{ground}|\Theta) = 1/(1 + \exp(\mu * (d(i, \Theta) - dr\,)))$$

This value represents the cost of setting a pixel to a specific label, given its relation to the pose prior, where d(i, Θ) defines the distance of the pixel to the pose, dr defines the thickness of limbs in the pose, and μ determines the ratio of the penalty.

Additionally, we add another unary cost as

$$\varphi(D|xi) = -\log \Pr(i \in Vk\,|Hk\,) \text{ if } xi = Xk$$

This value represents the cost of setting a pixel to a specific label in Vk, based on a histogram Hk of the foreground. Specifically, the closer the colour relative within the histogram of the given pixel is to the foreground, the more likely it is to be a part of the foreground.

Then, two binary costs are defined as follows:
$\varphi(D|x_i , x_j) = \gamma(i, j)$ if $x_i = x_j$
0 if $x_i = x_j$.

and $\gamma(i, j) = (\lambda \exp(-g(i, j)/2\sigma^2))/dist(i,j)$
where g(i,j) represents the absolute pixel difference between pixels i and j and dist(i,j) represents the physical difference between the two pixels. The intent of this value is to increase the cost associated with setting pixels in close proximity to each other higher as the difference in pixel intensity value between them increases.

Finally the last binary cost
$\psi(x_i , x_j) = K_{ij}$ if $x_i = x_j$,
0 if $x_i = x_j$.
Controls how flexible the graph cut can be in terms of separating neighbouring pixels, where the higher k is, the less likely dissimilar intensity valued pixels are to be given differing labels.
Given this we can construct the graph as shown in the following two node example and the dramatic increase in complexity for a 4 node graph.

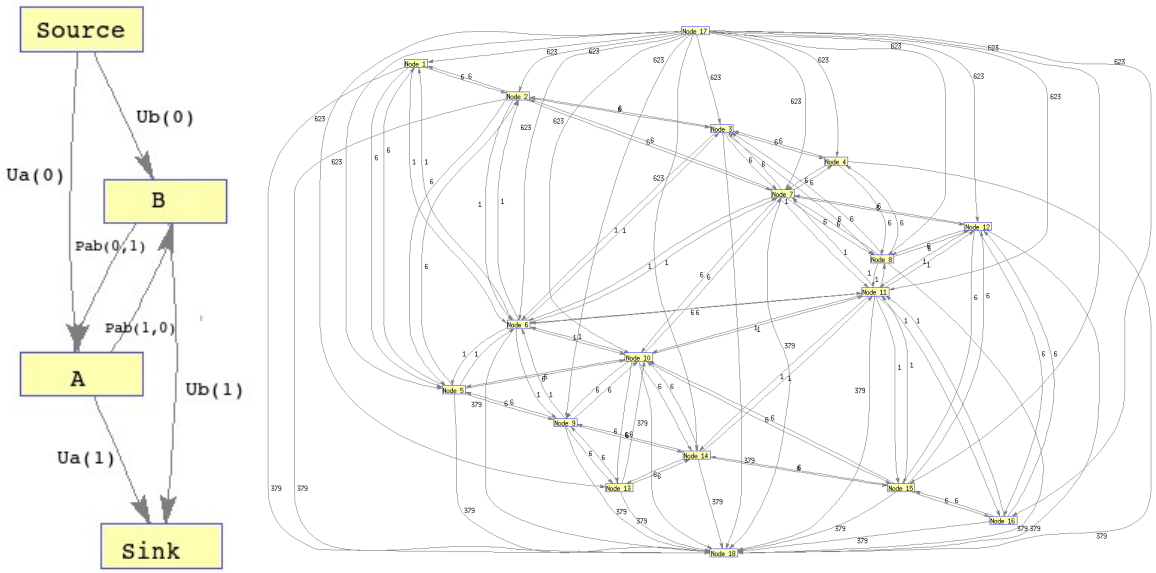

Figure 2: *A two node example of the graph construction and a four node graph*

By the choice of the energy function, it is possible to infer that the following constraint must hold
$E(0, 0) + E(1, 1) \leq E(0, 1) + E(1, 0)$
Specifically, the cost of setting pixels to hold the same label is lower than the cost of setting them to be different and thus the graph constructed is sub-modular. Then, the energy minima can be found in a single graph cut. However, there is a possibility that the unary terms assigned to their respective edge

may be negative by the construction given above. In such a case, it is necessary to reparamatize the edge by adding a value alpha such that Ua +alpha > 0. By adding this value to both unary terms (i.e to Ua(0) and Ua(1)) the min flow of every valid solution increases by exactly alpha, but the cut itself is not changed. Thus, the segmentation of foreground and background pixels can be found in polytime using a single graph cut.

Issues and possible solutions:

- The unary term relies on having a colour model of the foreground in advance, so as to compare it to given the intensity of the given pixel. However it is difficult to have a very accurate colour model. Differing choices of this histogram create very different segmentation results.
    - A functional work around was to use a small number of bins and form the histogram over the pixel in the area of the pose prior generated previously, as they are the most likely pixels to be part of the foreground at the time of initially performing separation.

Effect of varying parameters

The parameters that required particular tuning due to their effect on the results when varied are discussed as follows.

Changing the number of histogram bins has a large impact on the unary terms. The more bins within the histogram, the more sensitive the unary terms are to differences in colour. However having colours within a certain range of each other being placed in the same bin of the histogram, helps the algorithm be more robust towards noise. This can be beneficial when pixels that should be apart of the foreground differ slightly from each other in their intensity values. However, it is also detrimental as there are many occasions in which the intensity of pixels in the foreground is very similar to that of the background. In practice, it is difficult to find a suitable general choice of the number of histogram bins for all possible images.

Changing the value of k, a binary cost associated with setting two different pixels controls how conservative the graph cut is when assigning pixels differing labels. As the pairwise costs become lower, it is more possible to set differing labels on neighbouring pixels, and gaps within the segmentation appear. Otherwise, as K increases the costs associated with setting differing labels is higher than having them the same, and the regions within the segmentation become more uniform and tend not to contain gaps.

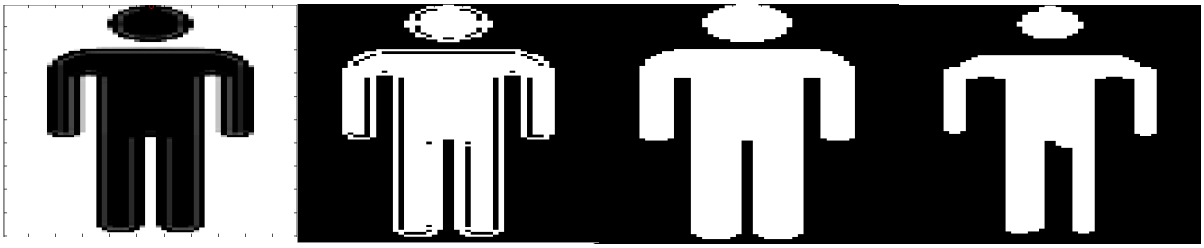

Figure 3: *A simple test case; the original image shown leftmost and, from left to right, the segmentation result for progressively increasing values of K. Note, a 2 pixel line from the centre of the head to midway of the chest was used as the int ital pose parameter.*

Segmentation results:

In ideal situations, such as the silhouette present in figure 3, the segmentation functions quite well when just using a simple pose and foreground colour prior. However, several issues make it difficult in more general real world situations. In grey scale images, the colour of the foreground tends to be very similar to that of the background, as can be seen in the following figure.

Figure 4: *From left to right, the original image, its grey scale, and the segmentation result using the previously discussed optimal values of k and the number of histogram bins.*

As can be seen in the above, the colour of the person's sweat shirt, particularly on his back and sleeve, in the grey scale image closely match to the background surrounding the person. In this case, the segmentation favours keeping them in the foreground, as there isn't much evidence to suggest that they should have differing labels. However, if colour information were utilized, it have been immediately apparent that the red sleeve has a distinct edge from the background and a more coherent segmentation would be possible.
Attempts to remedy this issue by using more histogram bins, to increase the sensitivity in the unary terms to the differences in grey scale intensity, produces the following two segmentation as seen in the below figure. The two images differ depending on whether the location of the initial pose prior is more to the left or to the right of the centre of the head.

Figure 5: *From left to right, using the same greyscale image as in figure 4, the effects of using a larger number of histogram bins and starting the prior in different locations.*

In this case, using a larger number of histogram bins increases the sensitivity of the algorithm to the starting location of the pose prior, which did not previously have a large effect on the results.

Additional segmentation results are pictured below.

Figure 6: *From left to right, the original image, its grey scale and the resulting segmentation*

As can be seen in figure 6, the algorithm tends to do better when the foreground has a fairly distinguishable pixel intensities in comparison to the background. However, as seen in the mislabeling of the face and floor siding, the method will tend to mislabel foreground pixels that predominately appear in the background and background pixels with similar intensities to pixels those that predominately appear in the foreground.

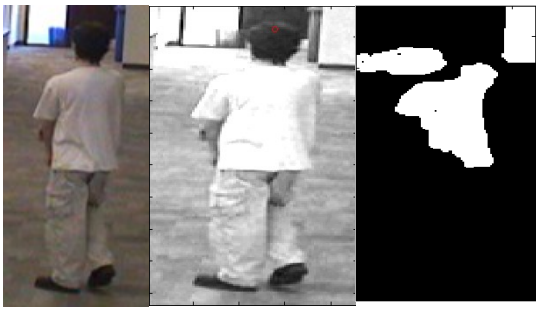

Figure 7: *From left to right, the original image, its grey scale and the resulting segmentation*

In the case of figure 7, the intensity of the pixels in the individual's head are closely matched to those in the door as are those in the pants to the floor. Additionally the individuals torso is closely matched to the wall and bright patches on the floor, causing the algorithm to perform very poorly. Again, these areas are more distinguishable in the original RGB image than the greyscale, but much of the information is lost when the channels are averaged.

Pose estimation:
The optimal pose parameters can be infered as follows:
$\Theta opt = \arg \min(\min(\Psi 3\ (x, \Theta))$.

Unfortunately, depending on the degrees of freedom of joints permissible, it is possible that the associated graph would be multi-modular, and that a single graph cut may lead to a local minima rather than a global one.

Conclusion

Ultimately the method provides a powerful means for segmentation, but suffers from several key issues, particularly in its initial assumptions. The method proposed relies on having a basic pose and colour model estimation of the foreground at the time of initial segmentation. However, it is difficult to generate an initial estimate that can be used before the foreground is identified. Possible solutions are to use a neutral stance pose and measure the response at every possible origin location and at every possible rotation, but such a solution is computation prohibitively. Additionally, it may not be suitable to use a neutral pose to discover a person who is in a vastly different pose, such as sitting. Experiments with iteratively trying to improve the pose and segmentation model, by starting with a guessed pose and foreground model and then using the results to perform multiple passes over the image also seemed unsuccessful. Particularly, if the algorithm is mislead into creating a bad separation initially, the pose it estimates is also likely to be poor. Then, if the image is processed again using these poor estimates, the algorithm is further mislead into creating increasingly erroneous segmentation and pose estimates. Thus, it may be more suitable to use a separate detection mechanism to try and identify humans in the image first and then use this method to segment the specific region containing the person.

External source packages used:

- To determine the min cut I used David Gleich's Math_BGL library
  http://www.mathworks.com/matlabcentral/fileexchange/10922