1 Software

The software design of the project consists of three central components: segmentation, registration, and coring location positioning. Segmentation involves grouping the scanned biopsy images into separate regions as well as recognizing the cancerous regions of the images. The registration step maps the scanned biopsy image to an image of the current paraffin block ready for coring in the TMA machine. The coring location positioning step involves choosing the best coring locations near the highest density of cancerous tissue that pack into the tissues shape most efficiently. With all three of these components in operation, successful TMA's can be created.

2 Image I/O

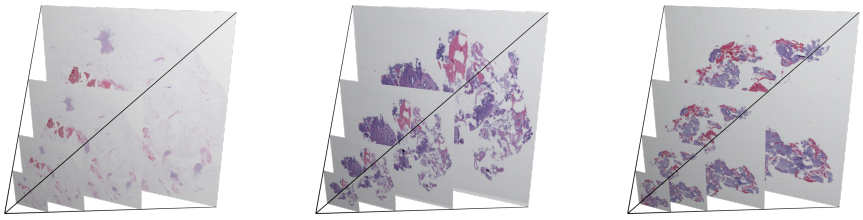
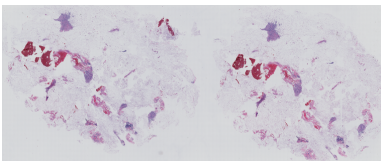


Figure 1: Pyramid image visualizations.

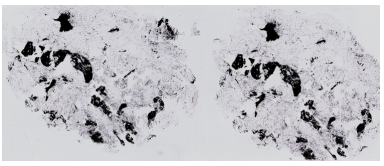
The images acquired from the medical scanner had very high pixel counts and were compressed into two special formats, JPEG2000 and Aperio's Medical Imaging format. Each of the images was around 90,000x90,000 pixels. When compressed, the images were around 300MB each in size, but uncompressed they were around 15GB. The images could not be viewed in their entirety without filling the RAM of a modern computer. This presented a challenge for processing images at such a large scale. Also, in many cases the full resolution was not needed or could not be used due to computational cost of processing large quantities of pixels.

It was decided that using the original image formats was inconvenient. Both Aperio and JPEG2000 are not standard formats accepted by most image processing libraries, making it difficult to read in data. It was important to convert the images to a more standard format that could be used by both OpenCV and Matlab. First an attempt was made to directly convert the images from the Aperio format using the OpenSlide library directly into the TIFF format. This involved having a detailed understanding of each image format-standard and was not a focus of the project. Learning all three image formats(Aperio's, JPEG2000, OpenTIFF) would have taken too long to understand in great detail. The Kakadu JPEG2000 image library and the VIPS image library were eventually used to convert from the non-standard image formats into the TIFF image format.

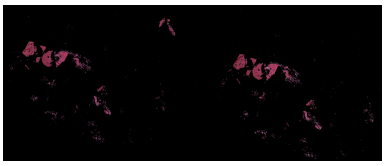
The OpenTIFF library was used to process the images and store them in a tiled and pyramid image format. OpenTIFF is an open source library readily available on the Linux operating system making it convenient for this application. A pyramid image stores several image resolutions at once in a single image file. The lowest level is the highest resolution and each subsequent level is half the original image size. Fig **??** illustrates the concept of an image pyramid. Image pyramids allowed for different image resolutions to be loaded into memory without further processing and provided memory management by using tiled images. Tiled images allow for random access to section of an image without the need to load the entire image into RAM. VIPS and Kakadu were the libraries used to convert the original image formats from the Aperio and JPEG2000 to OpenTIFF respectively. Once in the TIFF format it was easy to load the images into RAM for processing. Additionally the TIFF format was directly supported by both Matlab and OpenCV.



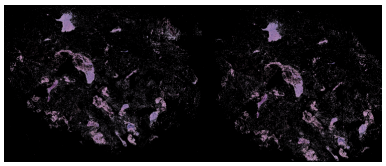
(a) Original image.



(b) Segmented Background.



(c) Segmented color cluster.



(d) Segmented color cluster.

3 Image Segmentation

The scanned slide images prior to acquisition were stained providing color differences between cancerous and non-cancerous cells. At full resolution individual cell color contribution could be seen, but at lower resolutions the image appeared to have different color region descriptions. Large concentrations of cancerous cells represented one color region and non cancerous cells represented another color region. This made color an obvious descriptor for segmentation. Figures 4 5 6 show the results of the three methods developed on several scanned paraffin block images.

The first segmentation attempt used histogram analysis to segment the pixels into regions. Each image was converted into the HSV(Hue Saturation and Value) color space. The Hue value represents the color component of an image and was used as a descriptor for segmentation. A histogram of the Hue channel of the image was created and the highest probable local maxima of the data set were accepted as the main color regions of the image. The HSV image was segmented by searching for pixels in the range of each of the probable maxima. This naive approach provided poor results and did not adequately segment the images. Large sections of the images were left unclustered. Fid 4 shows results from the histogram method. The histogram method did provide useful segmentation information about the number of regions in the image which could be used to properly initialize a K-Means algorithm.

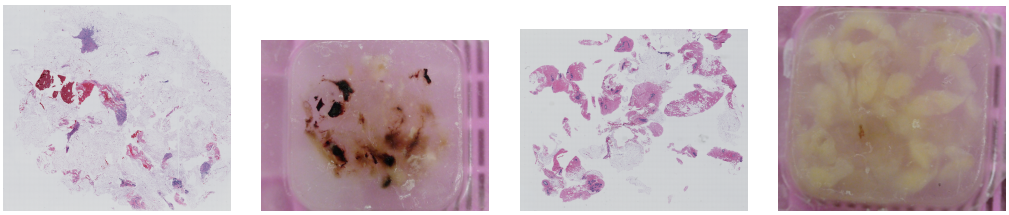
The next approach used a K-means clustering segmentation to break up the image into cancerous and non-cancerous regions. Here the image was converted into the LAB(Lightness and a-b color-opponent) color space. The LAB colorspace was used because pixels in the space are perceptually uniform across color and luminance. This makes it useful to compare the changes in the two color channel values against each other. Euclidean distance between pixels in the LAB space represent the difference in color of the pixels. The K-Means algorithm will use Euclidean distance to determine the difference between two pixels. LAB image data was used for a 3-ways K-means segmentation. This limited the regions that K-means sorted the pixels to just 3 groups. 3 groups were chosen initially because typically the images contain only a cancerous region, non-cancerous region and background region. The results from the K-means squared segmentation were superior to the previous histogram method. The K-means was able to successfully segment the cancerous and non-cancerous regions of the image with strong edges defined between the different regions. Additionally each region did not contain pixels from other regions of the image. The resolution however was an important factor in both computation time as well as successful segmentation of every pixel in an image. As the image size increased the K-Means algorithm began to fail on every iteration resulting in unsuccessful segmentation. It was found that lower resolutions produced better results for each of the tested methods. Figures 5 6 4 show a few resulting images segmented by each method at different resolutions.

The K-means segmentation was extended further by the work of the histogram method. The histogram provided useful information on the number of regions in the image. To improve results from the histogram data filtering was used to smooth the histogram to provide better local maxima detection and remove spurious maxima. The local maxima were thresholded below a probability and then counted to determine the number of regions present in the image. This value was then sent to the K-means algorithm to determine the proper regions for segmentation.

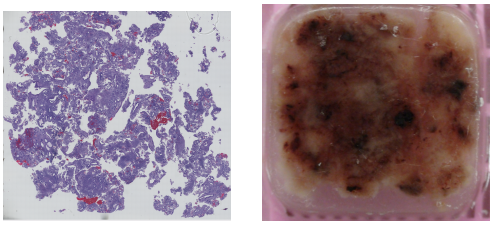
In addition, the segmentation problem was reduced by removing all background pixels from the K-means

and histogram step. The background of each image not containing any tissue was always a roughly constant white. This allowed for simple thresholding to eliminate large portions of the image. This also reduced the number of K-means groups by one.

4 Registration



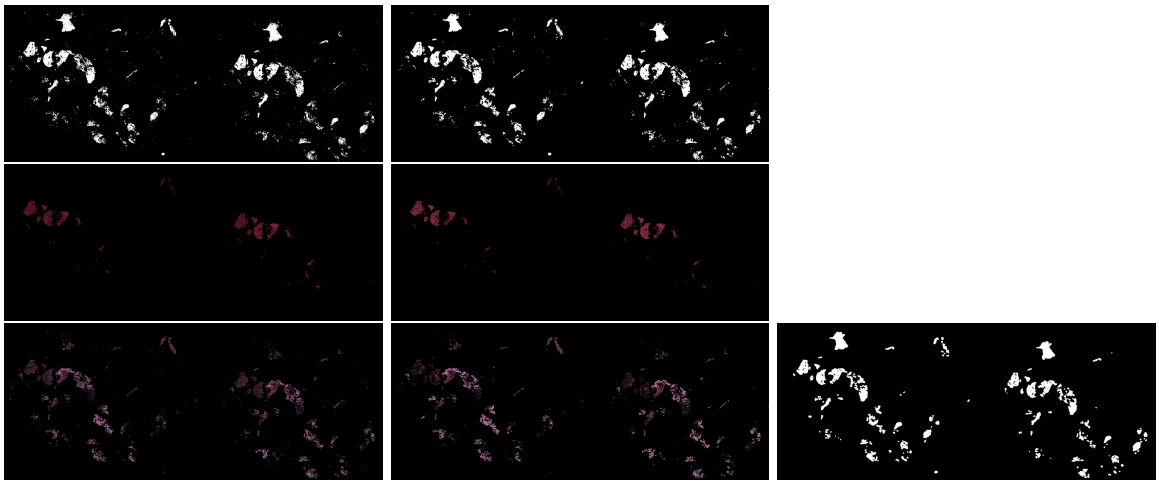
(a) Scan (b) Block (c) Scan (d) Block



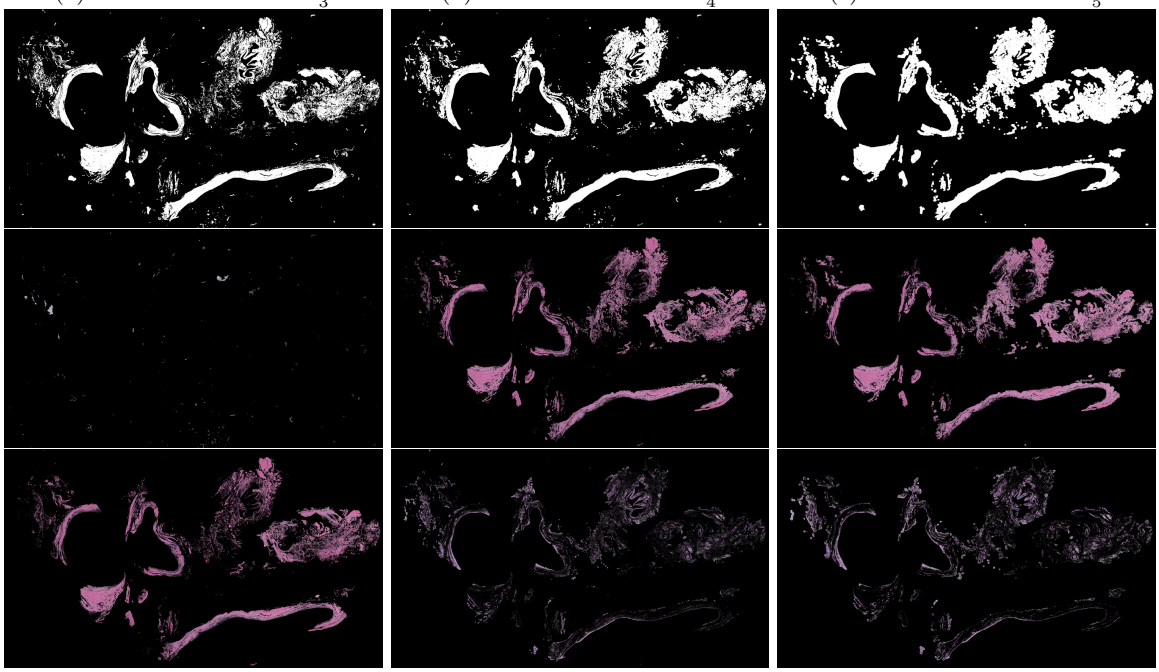
(e) Scan (f) Block

Figure 3: (a) and (b) group Block and Scan, (c) and (d) group Block and Scan, and (e) and (f) group Block and Scan representing the issues of subsurface tissue.

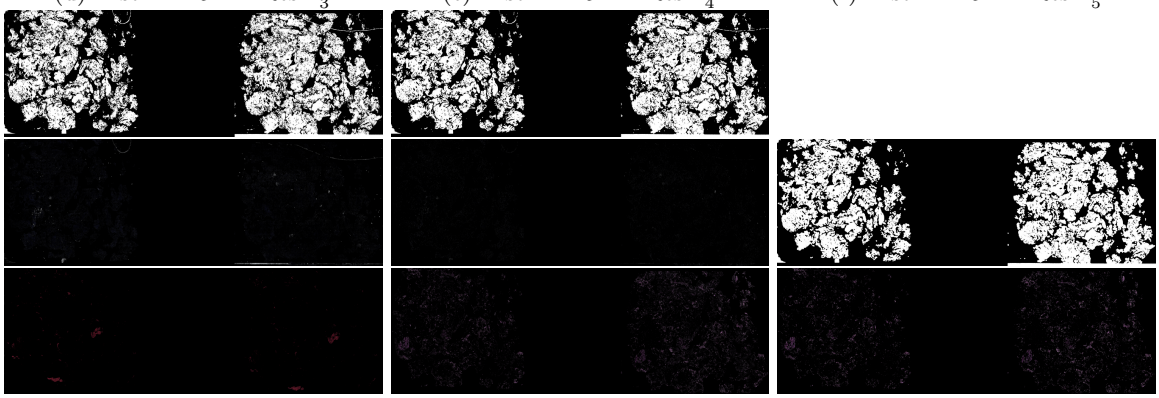
Image registration is the step that maps between the scanned image and the physical space of the paraffin block inside the mechanical TMA machine. Registration is central for coring the paraffin blocks, without accurate registration, the coring positions defined on the scanned image will not lineup with the paraffin block and the TMA will contain bad and unwanted tissue. Image registration proved to be a very difficult component. The two images that need to be registered are typically very different making it difficult for reliable registration. As seen in Fig 3 the relation between the scanned image and the paraffin block image is limited. The lack of similarity is due to the fact that the scanned image is from a top layer slice of the paraffin block and does not contain subsurface tissue information that is present in the paraffin block image.



(a) Hist - ID:516 - Res: $\frac{1}{3}$ (b) Hist - ID:516 - Res: $\frac{1}{4}$ (c) Hist - ID:516 - Res: $\frac{1}{5}$

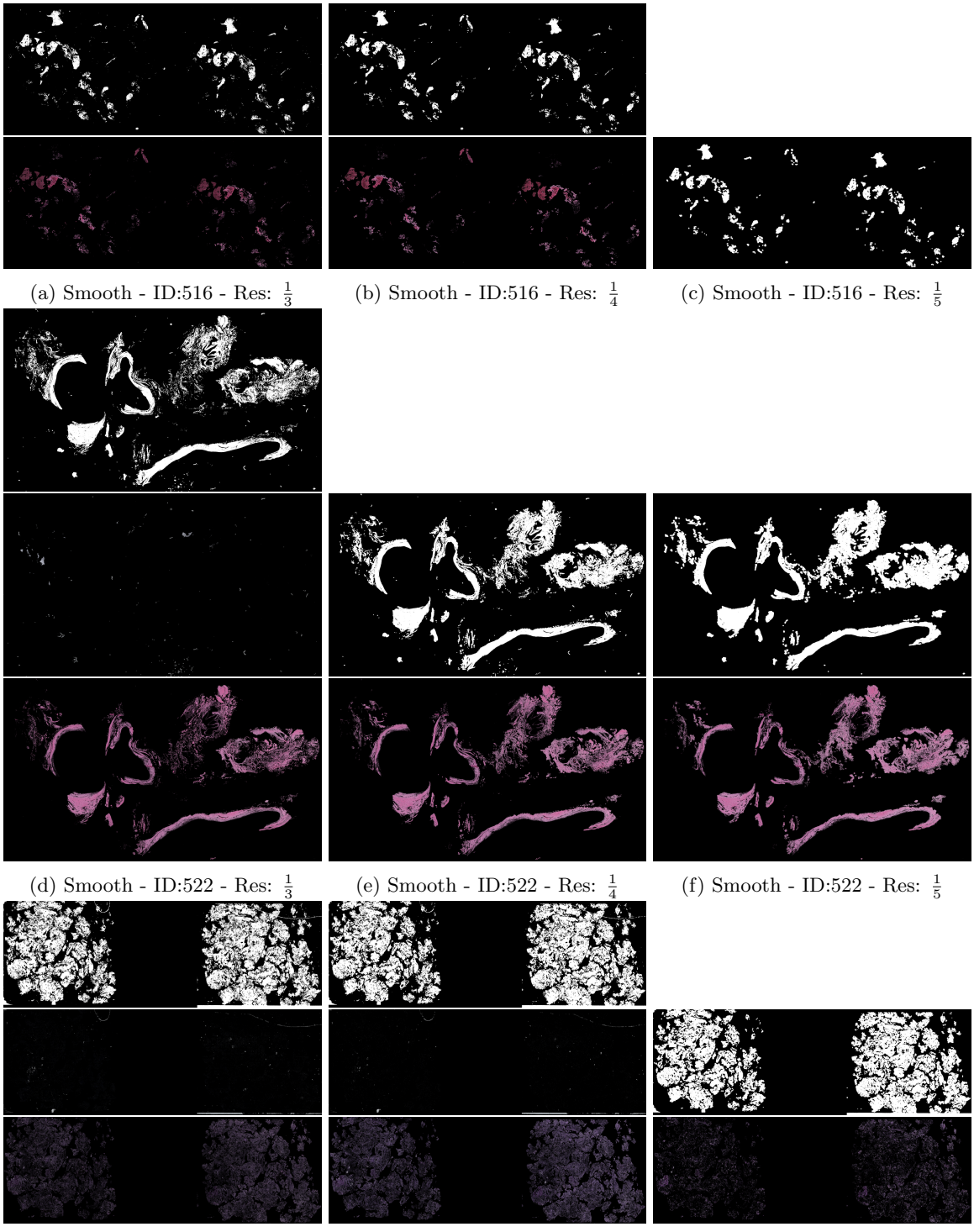


(d) Hist - ID:522 - Res: $\frac{1}{3}$ (e) Hist - ID:522 - Res: $\frac{1}{4}$ (f) Hist - ID:522 - Res: $\frac{1}{5}$



(g) Hist - ID:482 - Res: $\frac{1}{3}$ (h) Hist - ID:482 - Res: $\frac{1}{4}$ (i) Hist - ID:482 - Res: $\frac{1}{5}$

Figure 4: Histogram Method

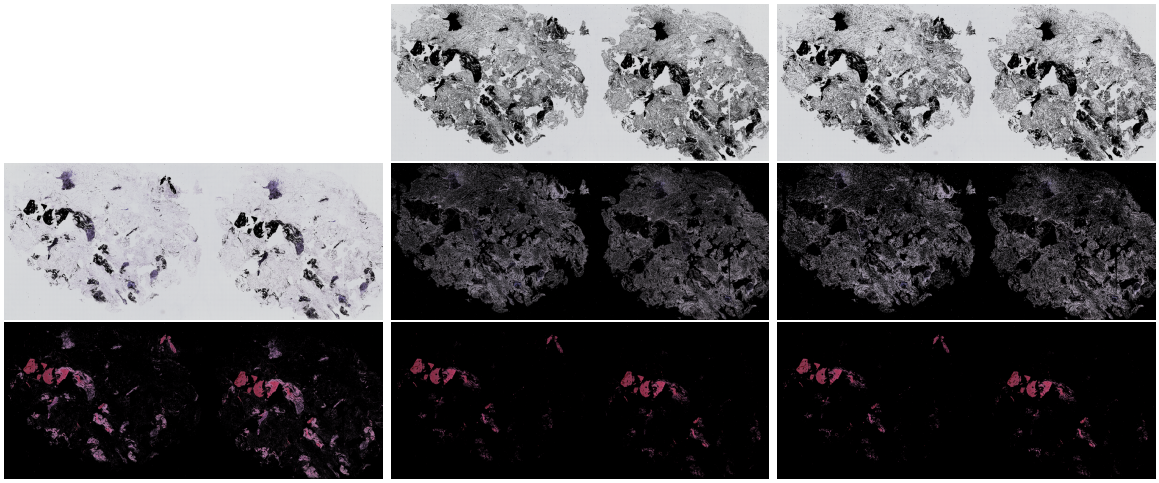


(a) Smooth - ID:516 - Res: $\frac{1}{3}$ (b) Smooth - ID:516 - Res: $\frac{1}{4}$ (c) Smooth - ID:516 - Res: $\frac{1}{5}$

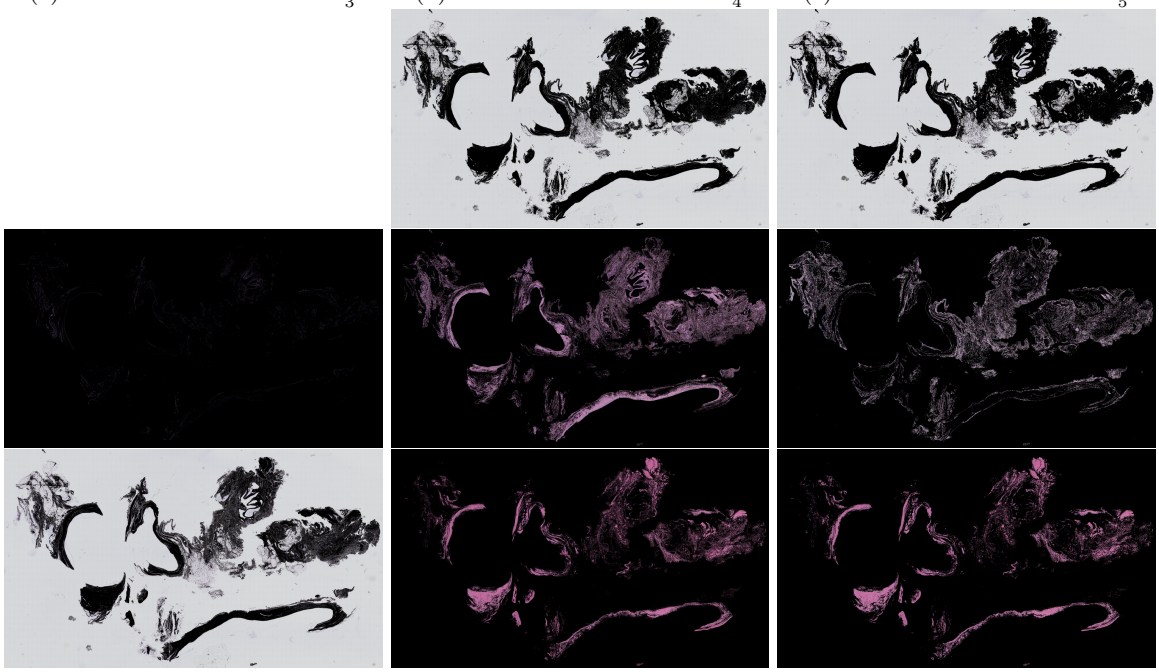
(d) Smooth - ID:522 - Res: $\frac{1}{3}$ (e) Smooth - ID:522 - Res: $\frac{1}{4}$ (f) Smooth - ID:522 - Res: $\frac{1}{5}$

(g) Smooth - ID:482 - Res: $\frac{1}{3}$ (h) Smooth - ID:482 - Res: $\frac{1}{4}$ (i) Smooth - ID:482 - Res: $\frac{1}{5}$

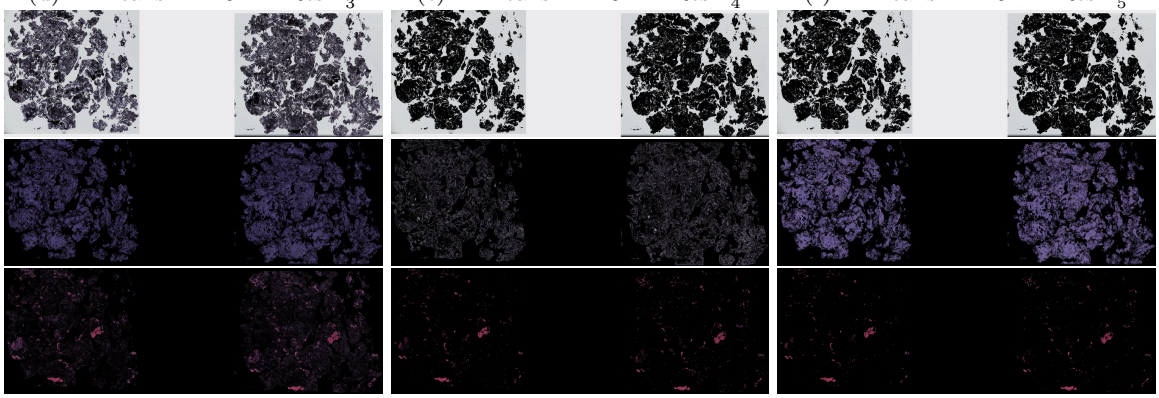
Figure 5: Smoothed Histogram Method



(a) K-Means - ID:516 - Res: $\frac{1}{3}$ (b) K-Means - ID:516 - Res: $\frac{1}{4}$ (c) K-Means - ID:516 - Res: $\frac{1}{5}$



(d) K-Means - ID:522 - Res: $\frac{1}{3}$ (e) K-Means - ID:522 - Res: $\frac{1}{4}$ (f) K-Means - ID:522 - Res: $\frac{1}{5}$



(g) K-Means - ID:482 - Res: $\frac{1}{3}$ (h) K-Means - ID:482 - Res: $\frac{1}{4}$ (i) K-Means - ID:482 - Res: $\frac{1}{5}$

Figure 6: K-Means Method