Image Processing 101

Donald J. Peck and Hamid Soltanian-Zadeh*

Department of Radiology, Henry Ford Hospital, Detroit, MI 48202, and
*ECE Department of the University of Tehran, Tehran 14399, Iran.

Introduction

Digital image acquisition and display are becoming commonplace for many imaging modalities. With the utilization of digital images, the interpretation of the data can incorporate post-acquisition processing methods that enhance and quantify the information available. The application of post-processing algorithms to medical data is not new. From its onset nuclear medicine, computed tomography (CT) and magnetic resonance images (MRI) have utilized some form of processing to reconstruct and interpret the information acquired. Yet with the increased emphasis on acquiring both physiological and anatomical information more advanced processing algorithms are being developed for all digital modalities. For example, information pertaining to the flow, uptake and permeability of a contrast material has been utilized to differentiate normal from pathological tissue. This information can be analyzed qualitatively by visualizing the passage of the contrast material through the different tissues or by visualizing where enhancement occurs on post-contrast images. In addition, the size of the contrast enhancement or the time verse signal change of the passage of the contrast material through the tissue can be used to quantify or characterize tissues. However, to accurately characterize tissue or define a change or difference between tissues the calculated values must be compared. This implies that the different tissues are isolated or segmented in some manner that enables the quantitative values from them to be evaluated. In addition to segmentation, the visualization of the tissue of interest and the calculation of different parameters to characterize them may be more robust if pre-processing algorithms are used to enhance the data in some manner. Although segmentation and pre-processing algorithms are used to improve the information available, the effects they have on the acquired data must be well established in order to understand if further analysis or interpretation will be biased by the data manipulation. If the image processing method is not reliable, then any qualitative or quantitative result will not be reliable. Although the goal of any image-processing method is to extract more information, all image processing destroys some of the information present in the acquired data. That is, image processing can be said to remove unwanted information (e.g. noise, artifacts, normal tissue) allowing the visualization and interpretation of the tissue of interest to be accomplished more accurately.

Many different processing methodologies have been applied to medical images and new methods are being developed every day. In this article some of these methods will be reviewed. Although a comprehensive review is not possible, the basic structure to many of the schemes used in these methods can be summarized. In addition, a rigorous mathematical approach will not be given since there are many good texts that cover these methods in detail . Instead, the generalized concepts behind the methodologies are presented and their application in medical imaging is demonstrated.

1.1. Data Representation

Most medical images are a two-dimensional (2D) representation of an anatomical location or projection where the signals acquired from different positions within the field of view are displayed as picture elements or pixels. Pixels contain information pertaining to the signal acquired from the tissue enclosed in the corresponding volume elements (voxel) or projection ray. That is, information relative to the signal from the tissue within a volume and the spatial location where the signal was acquired are both contained in the image. An example MRI of a patient with a glioblastoma is shown in Figure 1 with a region or neighborhood of the image expanded to demonstrate the individual pixels in the image.

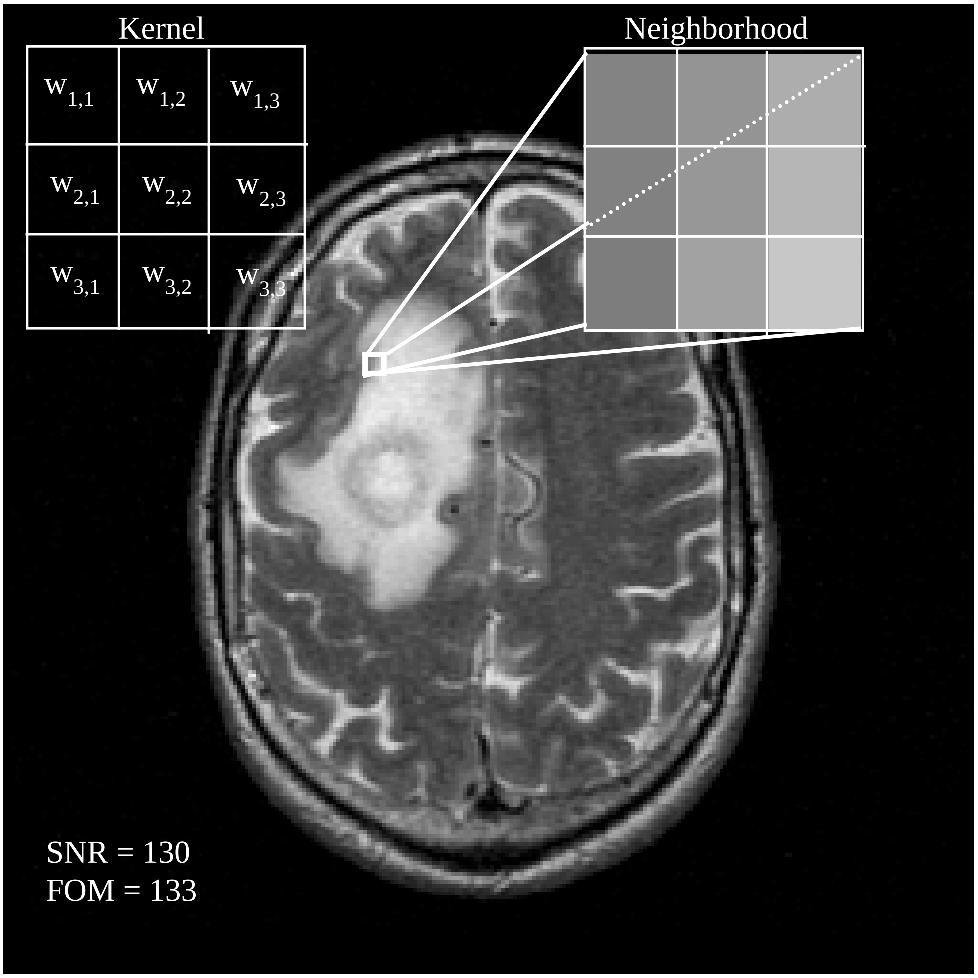

Figure 1

If a sequence of images from the same anatomical location is acquired, the spatial location and the signal change between images for each pixel can be utilized individually or in combination in image processing methods. When multiple images are utilized the signals from each pixel can be considered as a vector. In the vector representation each pixel would have a corresponding pixel vector associated with it, where the elements of the pixel vector are the signal from each pixel in each image in the sequence. Examples of image sequences are different T2- or T1-weighted MRI or images acquired at different times during the passage of contrast. In Figure 2 a sequence of MRI from the representative study is shown. In this example five 256x256 images were acquired and therefore there would be 65536 pixel vectors with five elements in each. In order to represent the data as pixel vectors it must be assumed the spatial coordinates, i.e. pixel locations, for all images are the same or the data must be registered to assure this assumption. In this review we will not discuss registration algorithms , although registration is considered a pre-processing method for many analyses. Initially we will review image analysis methods utilizing a single image and then expand the review to image processing methods that use the multi-image (i.e. pixel vector) representation for the analyses.

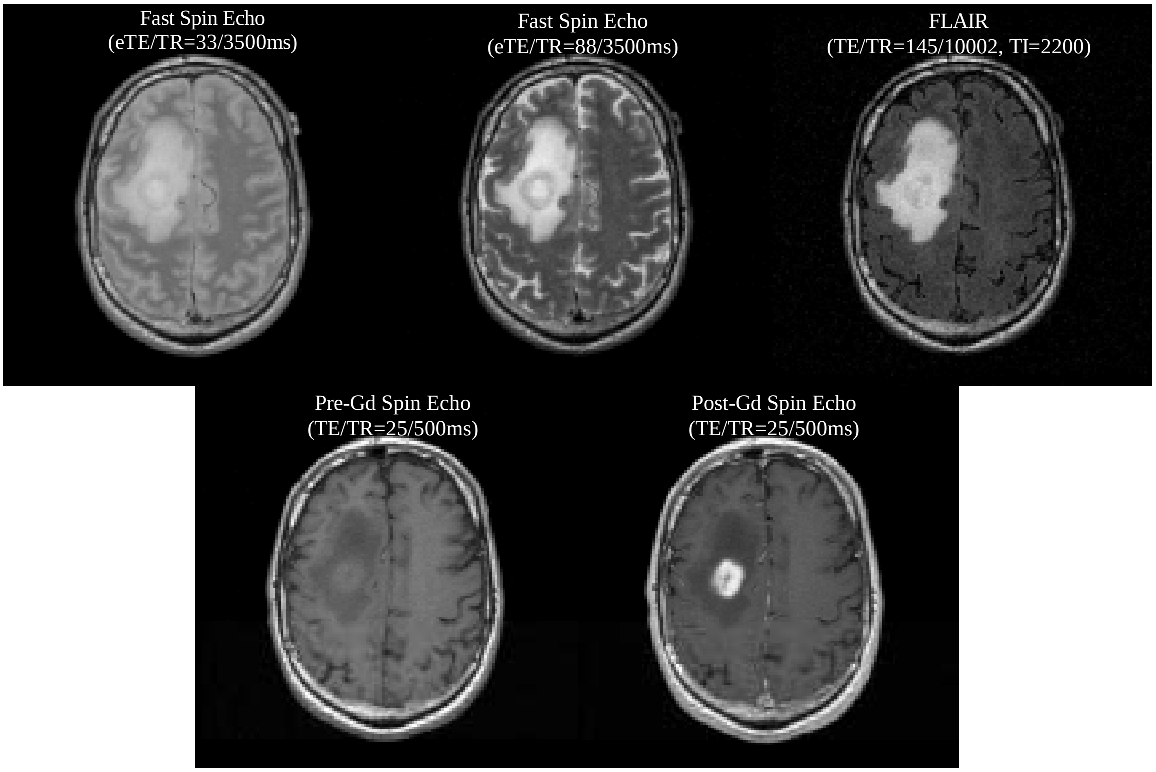

Figure 2

Image data can be represented in either a spatial or a frequency domain. Conventional medical images are a spatial domain representation of the acquired data since the displayed pixel signal corresponds to the spatial location where the signal originated. However, an image can also be represented by a combination of sine/cosine functions for different frequencies that can be used to reconstruct the spatial image. In this representation the number of each sine-cosine pair required to accurately reconstruct the image is displayed. The conversion from different data representations is termed transformation and specifically to transform data between the spatial and frequency domain is done utilizing a Fourier transformation. The mathematical theorems associated with Fourier transformation are not discussed here, but the concept of representing the data in either domain is important and will be reviewed next.

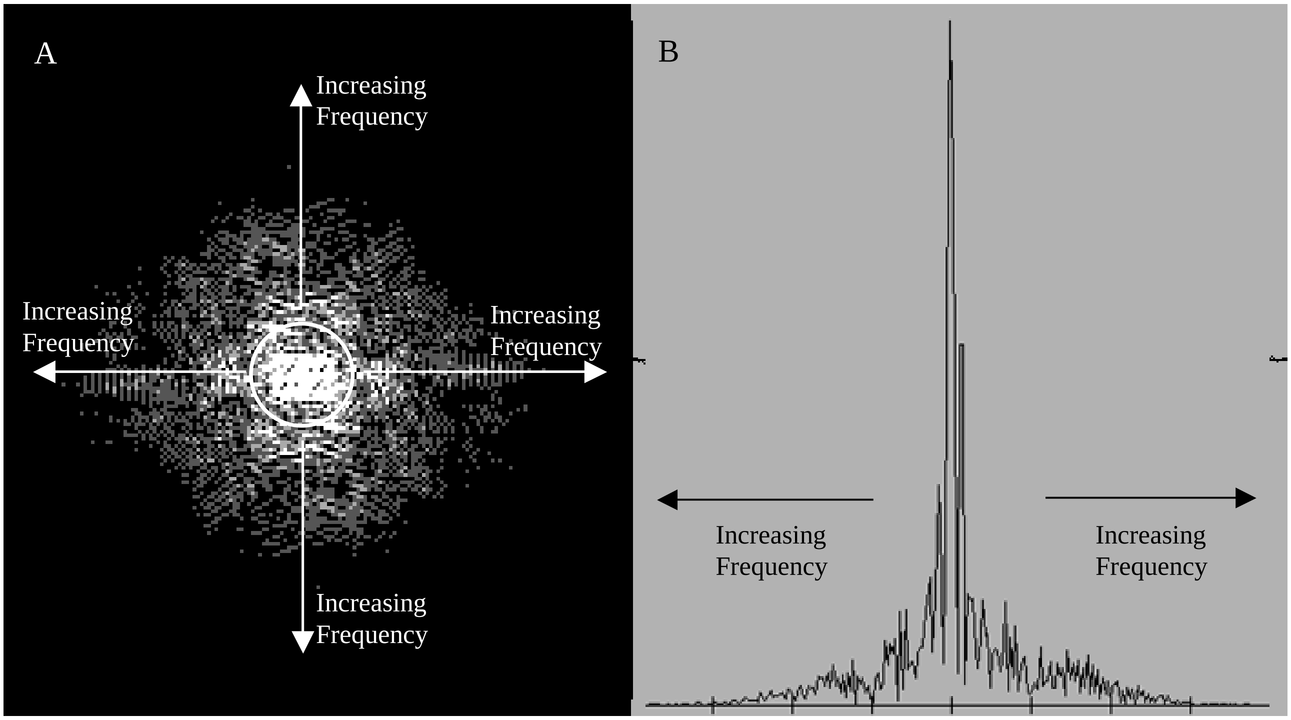

Figure 3

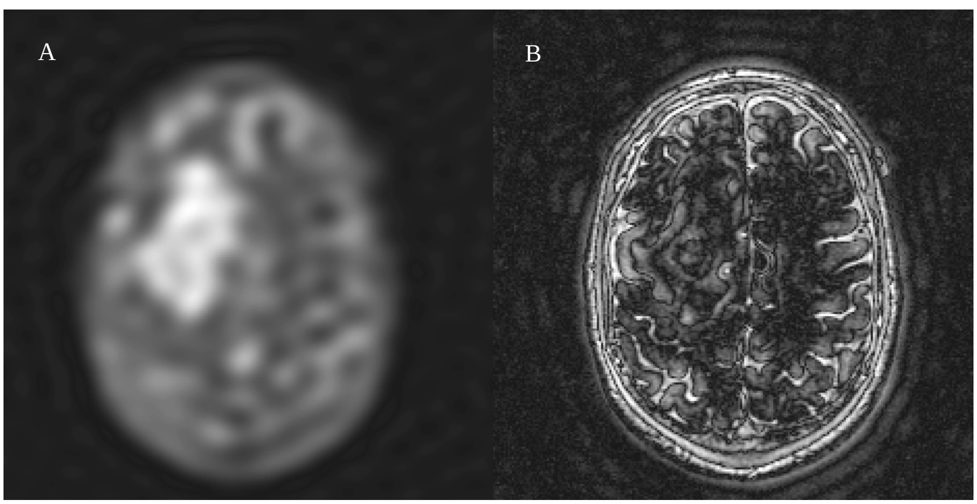

Figure 4

In Figure 3 the frequency representation of the image in Figure 1 is shown. Figure 3A shows the frequency distribution as a 2D image where the number of each sine-cosine function required to reconstruct the spatial image is represented as a gray level in the image. The center of this image would correspond to sine-cosine functions at low frequencies and as you move away from the center, in any direction, the frequency of the sin-cosine function increases. Figure 3B shows a one-dimension (1D) or profile through the frequency distribution in Figure 3A. In the 1D representation the magnitude of the curve is related to the number of sine-cosine functions at each frequency required for the reconstruction. It can be seen in Figure 3 that the majority of frequencies required to reconstruct the image in Figure 1 are contained in the center of the frequency representation, i.e. at low frequencies. In Figure 4 the original image is reconstructed by using only sine/cosine functions in the center of the frequency representation, i.e. at frequencies contained within the circle drawn on Figure 3A, and then again using only the functions outside the circle at higher frequencies. Note that in the resulting image the functions corresponding to the low frequencies provide the information necessary to reconstruct the general shape of the image and the relative contrast between tissues, but lack any sharp boundary information. The high frequency functions are required to reconstruct the edges between objects and tissue. In addition to the edge information contained in the high frequency functions, Figure 4B demonstrates that these functions also include more information relative to the image noise. Note that both the random and structured components (e.g. "ringing or ghost artifacts" outside the head) of the noise are more apparent in the image reconstructed using only the high frequency functions. This ability to differentiate contrast, resolution (i.e. boundary definition) and noise is one of the primary reasons for using the frequency representation of data in image processing methods.

Another form of data representation is related to the range of gray level values used. In many image analysis methods images are processed as binary data. The term binary means the pixel values can have only two values, often 1 or 0. Although most images are made of hundreds or thousands of gray level values, e.g. an 8-bit image has 256 possible gray levels for each pixel, the use of binary data can simplify the processing algorithms and the interpretation of the results. The most direct method to process gray level images as binary is to apply the processing method to only specific pixels that are defined as a region of interest (ROI). A ROI can include any number of pixels and can be defined as all pixels within a specified gray level range, the ROI can be defined by an operator arbitrarily (i.e. pixels selected or drawn on the image by an operator), or the ROI can be defined by a prior image processing method. Once the pixels are defined as either belonging to or not belonging to the ROI image analysis methods can be applied to the ROI as binary operations. Following image processing the ROI can be overlaid onto the original images to visualize or extract the gray level information from the pixels contained within the ROI. As stated above many image processing methods can be used to define ROI. These processing methods include many of the Feature Extraction and Segmentation algorithms discussed below. Therefore, in several of the examples the image processing methods may be applied to original gray level images or to ROI and the results may be displayed as a ROI or a gray level image.

2. Pre-Processing

All acquired data has associated with it additional information that may not be directly related to the tissue or physiological process being investigated. The most common type of "unwanted" information is noise. All acquisition methodologies have some form of noise. Noise can be random or deterministic. The random distribution of noise can be modeled based on known characteristics of the acquisition methodology and the object being imaged. This allows the noise to be utilized for establishing the limits or confidence levels used in many analyses. This implies that noise contains information pertaining to the acquisition method, but it may also contain information related to the signal characteristics of the tissue being imaged. In addition to noise other artifacts are often associated with imaging. The term artifact in this manner is generalized to mean any information that is included in the data but was not planned on being acquired. Artifacts are often structured or deterministic in nature and, as with noise, may contain information about the acquisition technique, tissue structure, and/or signal characteristics of the tissue. For example, inhomogeneity seen on MRI has information regarding the magnetic field and coils employed during the acquisition and the size, chemical content and shape of the object being imaged.

Although artifacts and other noise signals contain information they usually overshadow the tissue of interest and therefore pre-processing algorithms are used to remove them. The term pre-processing implies that these algorithms are the initial step in the analysis. Although this is often true these algorithms are utilized alone or in combination in many image-processing procedures. Therefore, when these algorithms are termed pre-processing it is assumed they are done to simplify or improve the results obtained in subsequent processing steps.

2.1. Neighborhood Analysis

Image-processing algorithms often utilize the relationship between individual pixels to make decisions on the amount or type of processing to be done. The manipulation of the data based on the values of the surrounding pixels from a neighborhood is termed neighborhood analysis. In neighborhood analysis processing is accomplished by applying different weighting factors to the values for each element of a neighborhood. The weighting factors are organized in a kernel where each weighting factor corresponds with a neighborhood element. An example of a kernel is shown in Figure 1. The values and location of weighting factors within a kernel are used to make up different image processing filters. Many standard filters are applied by multiplying the weighting factor in each element of the kernel to the corresponding element of the image neighborhood and then the central value of the neighborhood is replaced with the sum or average of all multiplications. This procedure is performed on the entire image by utilizing every pixel as the central element in multiple applications of the kernel until all pixels have been used. Following the application of the filter the resultant value for each pixel is stored in a new image called a filtered image. This filtered image can then be used as the result of the image processing method or the original images can be modified by the values in the filtered image. Therefore, different filters are created by varying the weighting factors in the kernel, by modifying the mathematical manipulation of the weighting factors within the neighborhood, or in the manner in which the filtered image is applied to the original data.

2.1.1. Noise Reduction and Smoothing

Noise reduction is clearly desirable to enhance image interpretation for many situations. Yet, the frequencies that include a large component of noise also include information pertaining to boundaries or resolution. Therefore, noise reduction methods are often termed smoothing algorithms since they reduce noise at the cost of blurring or smoothing boundary definition. A common smoothing filter is the low pass filter. The opposite of a low pass filter is a high pass filter. Although the high pass filter is not considered a smoothing or noise reduction filter it will be reviewed here due to similarities to the low pass filter. Both low and high pass filters remove some frequency information while maintaining others. Spatial and frequency representations of these filters are shown in Figure 5.

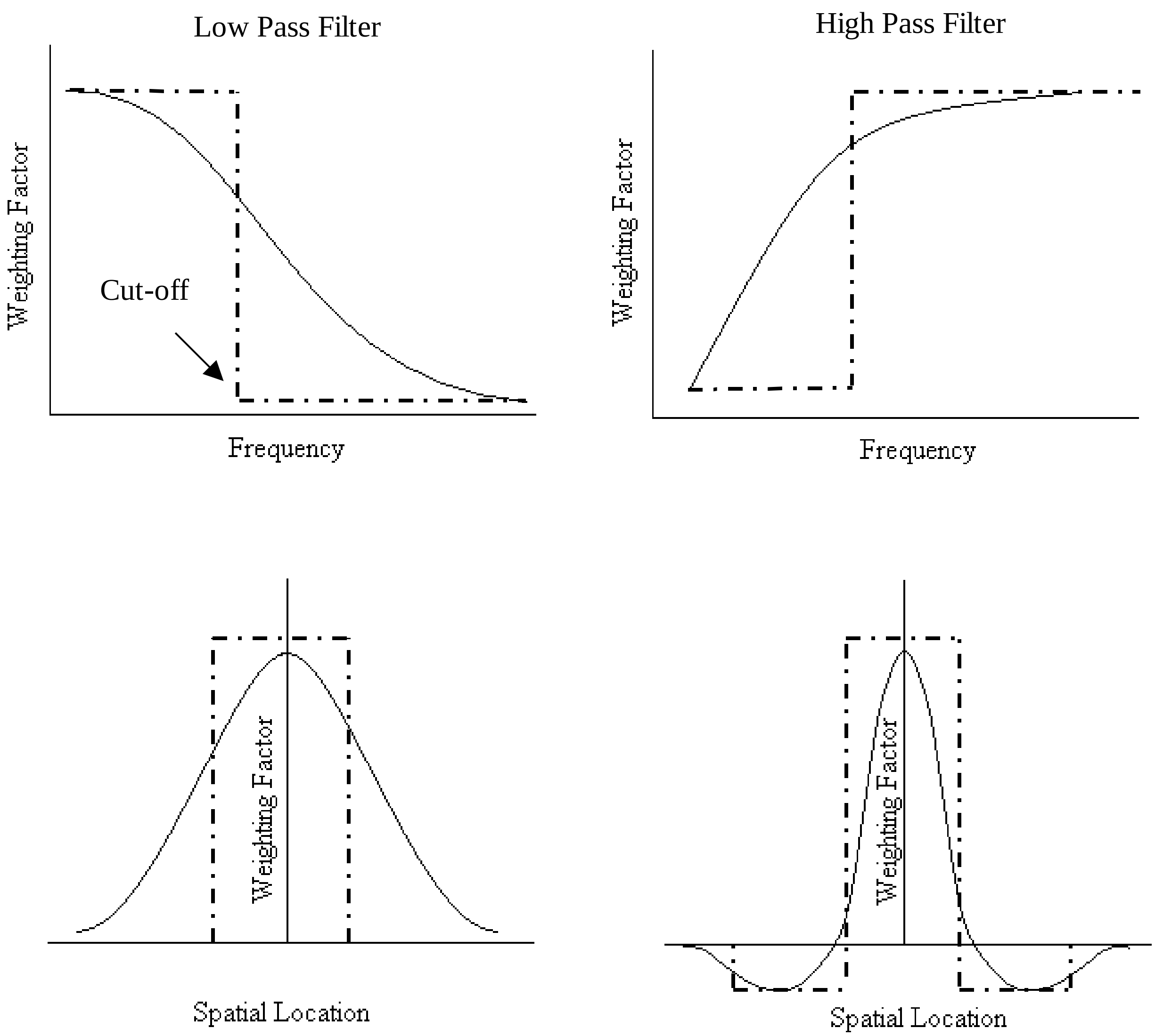

Figure 5

In the frequency representation of the low pass filter the low frequency sine-cosine functions have larger weighting factors and the weighting factors decrease as the frequency increases. The opposite is true for the high pass filter. The spatial domain representation of the low and high pass filter shows the central locations having larger weighting factors for both filters, but the high pass filter includes negative weighting factors as you move away from the central location. The equivalence of the frequency and spatial domain representations of these filters can be shown utilizing the Fourier transformation, although the interpretation of the filter's function is easier using the frequency representation. The application of the low pass and high pass filters in the frequency domain was demonstrated in Figure 4. In this example, the kernel had a circular shape and the neighborhood either included the frequencies inside or outside the circle. The filter used to create Figure 4A utilized a constant weighting factor, i.e. a value of 1.0, for all frequencies within the circular kernel and a weighting value of zero for frequencies outside the circle. In the frequency domain, the point where the weighting factors changes from a large value to smaller values, in this case from one to zero, is termed the cut-off frequency. The choice of the appropriate cut-off frequency to use can be determined based on the desired outcome. For example to remove noise the cut-off frequency can be chosen based on some knowledge of the noise in the acquisition and the required resolution in the results. As stated above, a majority of the frequencies required to represent noise and border definition (i.e. resolution) are both contained in the higher frequency values, so the determination of the cut-off frequency will affect both of these features. If an upper and lower cut-off frequencies are used the filter is termed a band-pass filter since the frequencies between the cut-off values are maintained, i.e. passed into the resultant image. The cut-off frequency is not easily interpreted from the spatial domain representation of these filters. Yet, it can be seen in the application of the low pass filter in the spatial domain that similar results to the frequency domain filter are obtained. This is shown in Figure 6A where a spatial low pass filter is applied to the image shown in Figure 1. Note that both noise and border information are removed while contrast and basic tissue structure are retained, similar to Figure 4A.

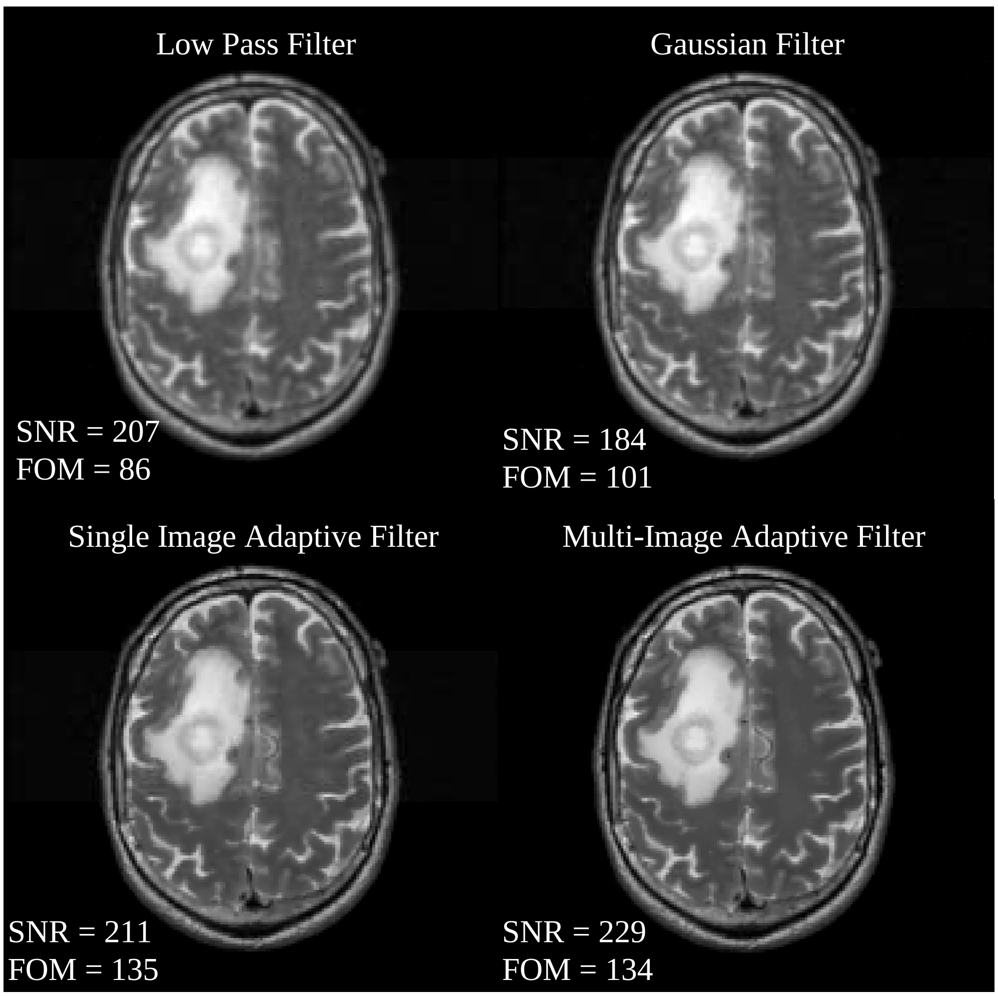

Figure 6

Although in the example of the low and high pass filters the cut-off frequency was chosen as the point where no frequencies above or below this value where included in the kernel the cut-off point can be a smooth or irregular transition between frequencies. This is done by varying the weighting factors in a pre-defined manner as you move from low to high frequencies (see Figure 5). Alternatively, a smooth transition can be generated for the weighting factors. One way to define this transition is to use a slowly varying function. For example, the kernel can be made to resemble a Gaussian distribution where the maximum weighting factor is used for the central value and the weighting factors are reduced following a Gaussian distribution as you move away from the center of the kernel. The application of a Gaussian spatial low pass filter is shown in Figure 6B. Note the similarity of this image to the spatial and frequency low pass filter results, i.e. Figures 4A and 6A respectively, with some improvement in boundary definition seen in the Gaussian filter results. In the frequency domain a common method for creating a smooth transition in weighting factors is the Butterworth filter . This filter can be designed to apply a rapid or gradual transition between high and low frequencies based on a desired cut-off frequency and the order of a polynomial function used to define the transition. The higher the order used the more rapid the transition of the weighting factors to zero and the closer the filter resembles the ideal low pass filter, i.e. abrupt transition to zero.

In conjunction with noise reduction, pre-processing algorithms are used for removal or elimination of other artifacts from the data. One common form of artifact is signal inhomogeneity. Inhomogeneity often appears as a gradual shift in the average pixel value in a specific pattern. The pattern varies depending on the imaging modality and the acquisition parameters or object being imaged. However, since the inhomogeneity is usually a gradual change it is often assumed a low frequency feature. Therefore, a low pass filter using a very low cut-off frequency can be used to remove all image information except the inhomogeneity . This results in a filtered image whose inverse can be subtracted from the original image to remove the inhomogeneity. An example of the application of a low pass filter for removal of inhomogeneity effects are shown in Figure 7. This type of filter can be very robust for removing simple inhomogeneity artifacts, but not all inhomogeneities are simple and methods that are more complex have been developed for correcting inhomogeneities in specific situations.

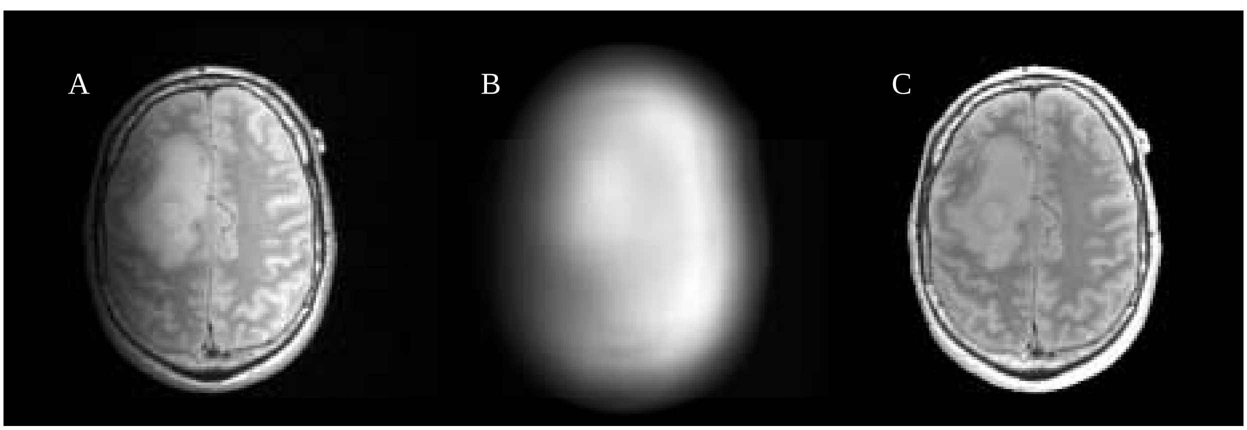

Figure 7

2.1.2. Adaptive Filtering

When developing a filter we have seen that the frequency representation can be used to limit or remove different information in the image. In order to accurately exclude some information, while maintaining other information, a priori knowledge of the data distribution can be used to improve the results. For example assuming inhomogeneity is a gradual change allows the use of a low pass filter to remove its affects on the image. However, unlimited knowledge of all effects that make up the image is not possible so trade-offs are made in the development of filters. This is usually done by compromising on the amount of data lost in one area or feature while maximizing the effects on another. Therefore, to reduce the unwanted affects while maximizing the desirable ones we need to increase the utilization of known information used when developing a filter.

In the filters defined previously the frequency distribution of the data and different cut-off frequencies or distribution shapes were considered in their development. However, the filters were applied to all elements of the neighborhood without considering the pixel values or relative locations of the elements. If an element could be excluded or included in the filter based on some other criteria the filter may be more robust. For example, if the pixel values between the central pixel and any element of a neighborhood varied by more than a specific amount it may be inferred that the tissue within these voxels are not the same (see Figure 1). When a low pass filter is applied to a neighborhood without considering this information the central pixel would be replaced with the average of all elements of the neighborhood reducing the boundary definition between tissues (i.e. smoothing the data). On the other hand, if the filter did not utilize the elements that varied from the central value by a specified amount noise reduction could occur without reducing boundary definition. The application of differing filters or filter elements within a kernel based on additional information obtained during the filter's application is termed adaptive filtering. An example of an adaptive low pass filter applied to the image in Figure 1 is shown in Figure 6C. In this filter an element of the kernel is excluded from the calculation if its pixel value is not within two standard deviations from the central pixel value. The standard deviation is considered an estimate of the noise in the data and is determined from a uniform region of tissue from the image. Note how the average signal inside each tissue in Figure 6C appears similar to the low pass filter results shown in Figures 4A and 6A but how the boundary definition is improved in this image.

An extension of an adaptive filter to remove noise while preserving boundary information can be optimized by using the multi-image data instead of only one image. By using the multi-image data, i.e. pixel vector representation, the information from all images can be integrated in the filter and results that are more robust can be obtained. As in the previous adaptive filter a decision criterion must be developed in order to include or exclude pixels or in this case include or exclude pixel vectors from the filter during its application. One such criterion is to calculate the Euclidean distance of each pixel vector in the neighborhood from the pixel vector in the center and compare the result with a preset threshold value . The threshold value can be found by calculating probabilities of a specific neighbor being included in the filter yet not belonging to the same tissue as the central pixel. Since we are now dealing with vectors, the threshold can use several parameters such as the standard deviation of the pixel values in each image, the contrast between tissues or other criteria. Following the determination of which neighbors are to be included in the filter the average of the contributing pixel vectors can then be determined and the central pixel vector replaced with this vector. This type of filter reduces the noise in all images in the multi-image data set (i.e. all images in the sequence) simultaneously. Figure 6D shows the second fast spin echo image (i.e. Figure 1) following the application of this multi-image adaptive filter to the image sequence shown in Figure 2, again demonstrating noise reduction and edge preservation.

2.1.3. Temporal Smoothing

Noise reduction and smoothing filters can also be applied to multiple images acquired at different times points, i.e. temporal acquisitions. Examples of temporal acquisitions would be first pass contrast bolus tracking, contrast uptake, and activation or functional imaging studies. When these filters are applied to temporal data the neighborhood analysis can be defined three ways. First, the filter can be applied to the neighborhood in each individual image as a standard spatial filter. Second, the filter can be applied to corresponding pixels from the different temporal acquisitions. In this method of temporal analysis the weighting factors of a kernel are applied to a neighborhood consisting of one pixel location from several images and the pixel from the central image would be replaced with the resultant value determined by the filter. The kernel would then be shifted so the central pixel is located on the next image in the sequence until all images are included as the central location in the neighborhood. This would then be repeated for all pixel locations. This is different from the pixel vector analysis given above since the analysis does not consider all images in the operation during each application but instead uses only a subset of images. In addition, the filter modifies only the pixel from the central location of the neighborhood instead of modifying all pixels or elements of the vector simultaneously. The third method of temporal analysis uses a combination of spatial and temporal components in the application. These filters can apply the spatial filter first then the temporal filter, or visa versa, or can use both the spatial and temporal information together. An example of the application of a low pass filter to a temporal signal is shown in Figure 8.

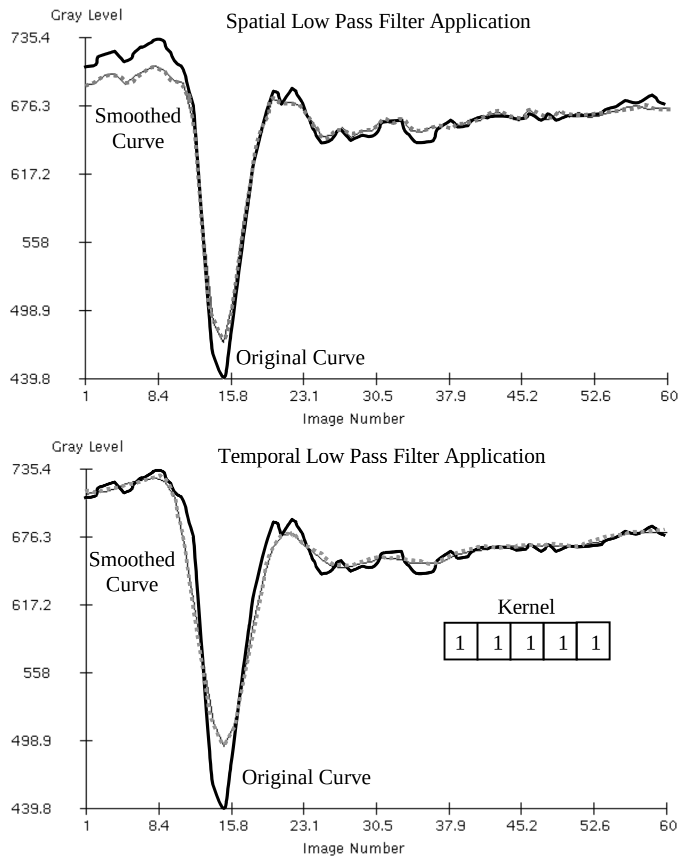

Figure 8

In this example the signal change from a ROI placed in normal white matter from a first pass gadolinium (Gd) gradient echo planar MRI study (TE/TR = 40/1900 ms) is shown. The results after applying either a spatial or temporal low pass filter to the data is over laid onto the original curve. Note the spatial low pass filter reduces the average signal value prior to the bolus arrival and changes the amount the signal decreases during the passage of the Gd bolus. The reduction in the signal is due to the due to the spatial low pass filter smoothing boundaries in each image as explained above. After the passage of bolus, the Gd in the blood pool is distributed evenly reducing the acquired signal for all tissues. This reduces the contrast between white and gray matter in the brain since the blood volume is larger in the gray matter. Because of this, the spatial filter shows less change in the average signal after the bolus as compared to the change seen before the bolus. In the temporal filter application the average signal does not change except during the passage of the bolus. The rapid or sharp changes in the temporal curve that occur when the bolus passes are similar to the edges of tissues seen in spatial images and correspond to high frequency information that is being lost in the smoothing process. Also note in the application of the temporal filter the change in the signal as the bolus passes is exaggerated causing the bolus width to be increased at this point. Therefore the decision to use a spatial, temporal or combination filter to analyze temporal data must consider these effects in order to make the appropriate decisions on how much information is being modified in different regions of a curve during the analysis. In addition, temporal analysis can be optimized using frequency domain filters, modification of filter weighting factors and using prior knowledge or adaptive filters similar to the pre-processing filters discussed above.

2.2. Considerations/Caveats to Preprocessing Methods

In pre-processing and specifically in noise reduction methods the desired results are often to remove artifacts and increase the signal-to-noise ratio (SNR) in the data. The SNR from a uniform region of tissue for the original image and the results obtained for the noise reduction methods described above is shown on each Figure. Note that the application of each filter has increased the SNR by 40% to 75% compared to the SNR from the original image. Yet, noise reduction filters also result in some reduction in boundary definition. It is difficult to quantify a change or preservation of boundary information. One type of quantification would be to measure the standard deviation of the pixel values from a line across a known boundary. This measure can be considered a figure of merit (FOM) for the preservation of boundary information. If the FOM decreases the pixel values across the boundary are becoming more similar, i.e. boundary information is being reduced. If noise was reduced and boundary information maintained, the SNR should increase and the FOM should remain the same. The FOM for the original image and the results from each noise reduction filter across a known boundary is also given on the Figures. Note for the low pass and Gaussian filters the FOM is reduced compared to the original image. The adaptive filters on the other hand show the largest increase in SNR while maintaining the boundary definition as seen in the SNR and FOM for these filters (note, the slight increase in the FOM for these filters is not significant). The adaptive filters demonstrate how the use of a priori knowledge can enhance the results of a filter.

Another aspect that needs to be considered when utilizing neighborhood analysis methods is the kernel size used. In the examples shown above a 3x3 kernel was used for all examples except Figure 6D and Figure 7. If the neighborhood size is increased, the number of data points used in the calculations increases and the results can change dramatically. In Figure 6D, the adaptive filter kernel size was increased to 9x9. This is because the adaptive filter in this example used all the images in the acquired image sequence when making a decision on the pixel vectors to include for each pass of the filter. This allows more pixels from the same tissue to be included in the kernel when determining the new average for the central pixel and therefore the SNR was increased without changing the FOM. When a 3x3 kernel was used, the SNR was only improved by 10% compared to the original image although the FOM remained the same. Although this demonstrates that improved results can be obtained by modifying the kernel size, it is done at a cost. In the case of the adaptive filter the cost was processing time. The increase in the kernel size and the use of the multi-image adaptive filter increased the processing time by approximately a factor of 10. However, the multi-image adaptive filter also processed all five images together, i.e. provided increased SNR for all images simultaneously, so the time difference should consider this fact. Yet, the change in kernel size can also result in decreased processing time while increasing SNR or may be required to provide the desired results. For example, the use of a larger kernel was required in the application of the inhomogeneity correction filter. In order to remove all high frequency components of the data the low pass filter had a kernel size of 33x33. As can be seen in the filtered image (Figure 7B) the boundary definition is completely removed in this application, which is the desired outcome to define only the inhomogeneity.

Therefore, in the application of pre-processing methods it is important to utilize as much a priori information in order to optimize the results and it is important to understand the information that may be lost or modified in the application. If care is taken with the selection and utilization of these filters they can provide an increase in SNR and isolation or removal of unwanted attributes in the data to greatly enhance further analysis.

3. Feature Extraction and Segmentation

As opposed to reducing or removing artifacts it may be desirable to enhance or quantify other features of interest within an image. Features can be objects or tissues or they can be any structure or characteristic of the acquired signal or tissue. For instance, a tumor or lesion can be considered a feature but in addition, the shape or texture of a lesion can also be considered a feature. Of course, artifacts and noise are also features of an image and many of the methods presented next may segment or define many features in the data, including noise. Therefore, pre-processing filters that reduce noise will assist in obtaining improved results using many of the methods explained next.

Often once a feature is located or defined, specific characteristics of the feature may be desired. These characteristics can be either qualitative or quantitative. Examples of qualitative characteristics are the shape, position, or signal intensity of a feature relative to other features while a quantitative characteristic could be the size of the feature or distance between features. In either case, the evaluation of a feature's characteristics is easier if the feature of interest is segmented from other features. Segmentation and feature extraction are often used interchangeably, since one implies the other. However, note that a feature may be a single attribute of a tissue or object and therefore feature extraction may not segment a tissue. On the other hand following segmentation other features of the object or tissue may be desired requiring further feature extraction methods to be applied to the data.

3.1. Edge Detection

Discontinuities in the signal can represent many features within an image, such as the boundary between tissues. The presence of signal discontinuities implies that a method that can determine signal changes would be good at detection of these features. A simple method for defining discontinuities is by determining the rate of change or gradient of the signal. The rate of change can be found by taking the derivative of the signal change. An example of a changing signal and its derivative is shown in Figure 9. Note that the derivative has a positive value when the signal is increasing, is maximum at the point where the signal change is maximum, and is reduced as the signal returns to a constant value. In neighborhood analysis the derivative in a specific direction can be found by applying a filter where the weighting factors are configured as alternating positive and negative rows and columns. In Figure 10A a filter is applied that has alternating positive and negative column weighting factors therefore creating a filtered image in which the vertical edge between features is enhanced, i.e. detected. When the orientations of the positive and negative weighting factors of the kernel are rotated the filter can detect horizontal and oblique edges also. The application of vertical, horizontal, oblique and the sum of all orientations as applied to the image shown in Figure 1 is shown in Figure 10. Note that the sum of all directions of the gradient filters (i.e. Figure 10D) is similar to the high pass filter results shown in Figure 4B. The difference is that high pass filters allow all high frequency information to remain in the image, including noise, where the gradient filters are less sensitive to enhancing the noise unless the noise in the image is large or structured. In these cases, a pre-processing method that reduces the contribution of noise and inhomogeneities should be applied prior to edge detection.

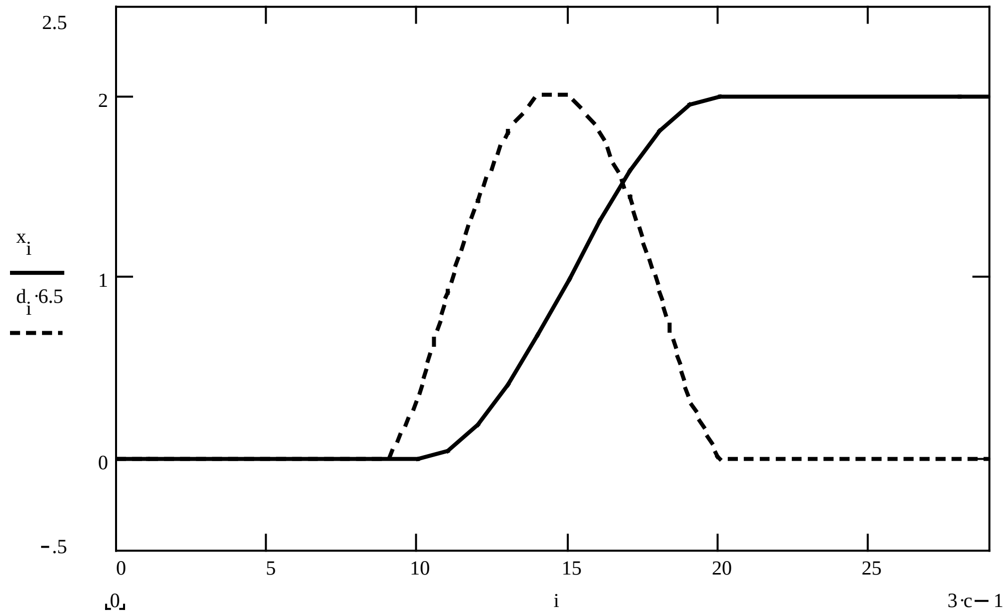

Figure 9

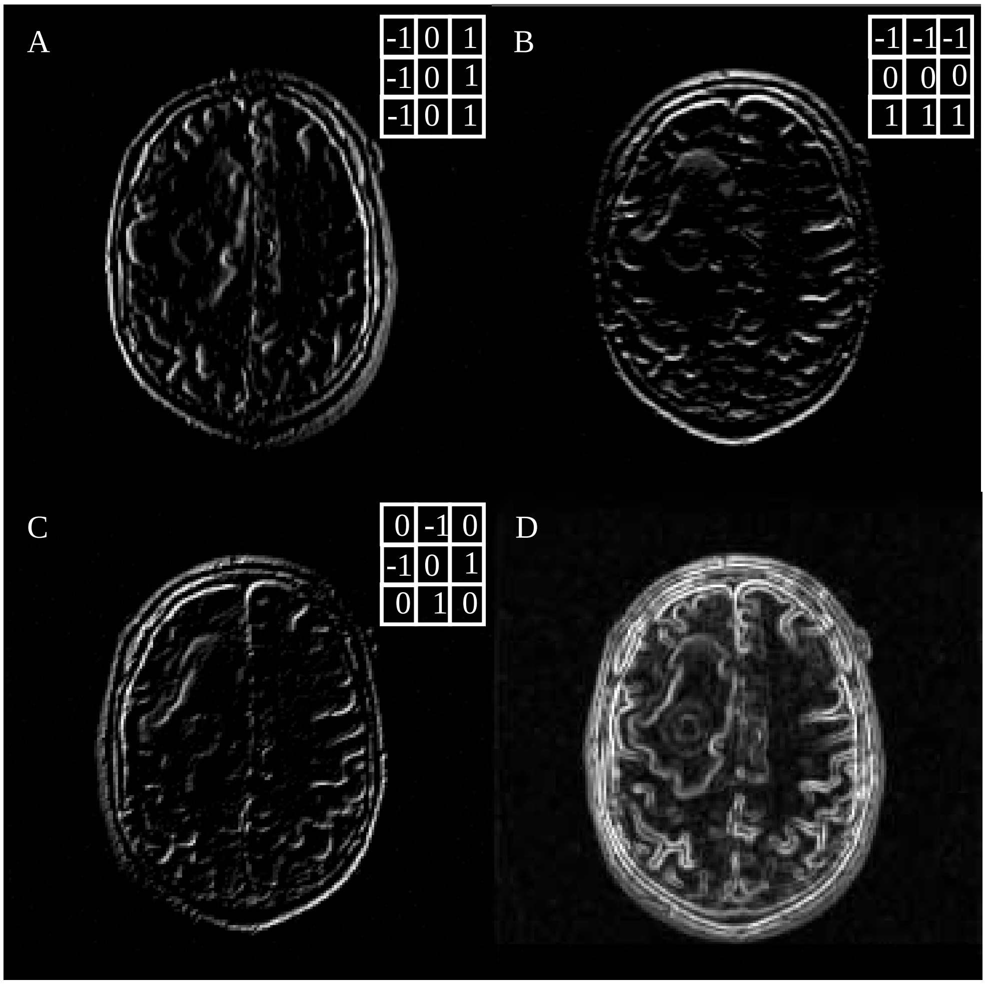

Figure 10

3.1.1. Edge Tracking

The determination of an edge between signal discontinuities doesn't necessarily segment an object. As stated above the edge between different signals in an image can be considered a feature of the image but these signal changes may be due to many different factors, many of which are not the boundary between objects. In addition, the edges for many different objects may be segmented using an edge detection method even when only one object is of interest. In order to use the detected edges as a boundary of a specific object of interest some additional criteria needs to be included in the processing. The criteria for defining an object can be simply the gray level associated with a tissue or structure of interest or could be a pattern segmented by a feature extraction method that defines an object. One method for determining the boundary of an object is to use an edge tracking method. In edge tracking the edge or contour of an object is found by using a predefined edge as a guide. Edge tracking methods locate an initial edge point then follow the edge of the object until it reaches the initial edge point again. The initial edge point is determined from a specified start point, termed a seed point. The seed point is normally input by an operator and can be placed either inside or outside the object of interest. Following the placement of the seed point the image is scanned in a certain direction until it intersects an edge. To follow the edge the pixels in the neighborhood surrounding the initial edge point are examined until another edge pixel is found. This new edge pixel is designated as the second pixel on the edge and the algorithm uses it as the starting point to search for the next edge pixel. This procedure is repeated until all edge pixels are found. The result of an edge tracking scheme used to segment the lesion seen in Figure 1 is shown in Figure 11. In the example the lesion is first segmented based on the range of gray levels separating the lesion from the surrounding normal tissue. This initial segmentation is shown as a ROI overlaid onto the original image in Figure 11A. By defining the object, in this case the lesion, using a ROI allows the edge tracking algorithm to process the data as a binary image. By doing the analysis as a binary process the algorithm is very fast. Although the lesion is clearly included in the range of gray levels chosen to create the ROI, cerebrospinal fluid and extra-cranial tissue are also included in the initial segmentation. To determine the edge of the lesion a seed point for the edge tracking algorithm is placed inside the lesion ROI. From this point the outer edge of the lesion ROI is determined by the algorithm as explained above. A graphical example of how the edge is tracked is shown in the Figure along with the final lesion contour found.

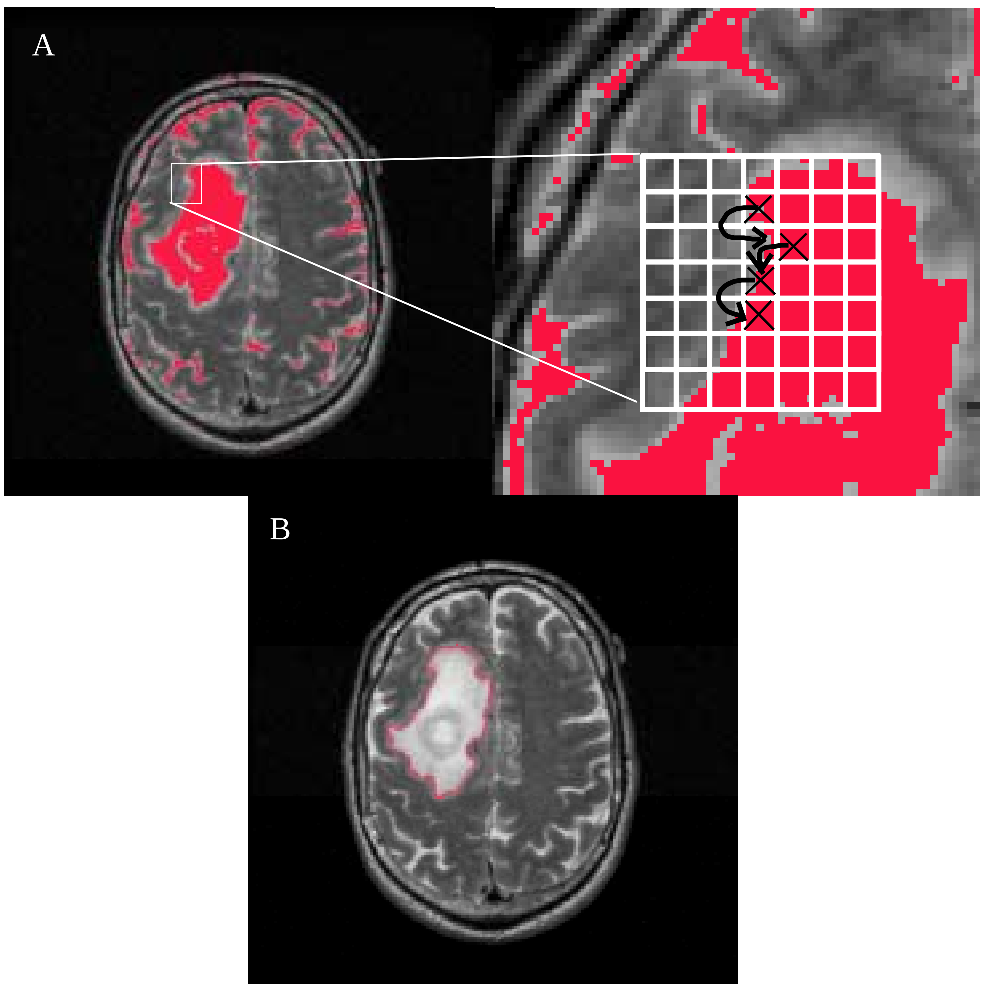

Figure 11

3.2. Connectivity

It is intuitive in human perception that pixels that are not spatially connected may be grouped into the same object (Gestalt reference). Often when performing image segmentation some pixels that are part of an object may not be segmented due to noise, non-uniformities, low contrast between objects (e.g. partial volume effects), or other effects that exclude these pixels as part of the object. In addition, the image processing method may segment pixels with the same gray level values or other attributes defined by the algorithm as the same object yet they may not be considered the same object by the operator. Therefore, the processing methodology must have the ability to combine or exclude pixels as being the same object. The ability to link pixels as similar or as having similar information (i.e. determine the connectivity of objects) is an important concept in image processing. Many factors can be used to determine the connectivity of pixels or objects. These factors can be the spatial relationship (e.g. pixels that are physically close to each other), they can be the similarity of the gray level values or pixel vectors, or other attributes can be defined by the image processing algorithm using more sophisticated criteria to link pixels together as the same object. Note that this definition of connectivity is not limited to the spatial distribution of the pixels but can include many different attributes of the pixels or pixel vectors and therefore the definition of connectivity may be specific to the processing method or desired outcome.

3.2.1. Morphological Operations

Neighborhood analysis using the spatial relationship between pixels with similar values is a standard method of determining connectivity between pixels. For example, if the pixel values in a neighborhood are within a specified range of values, e.g. within two standard deviations of the mean pixel value, they may be considered connected. However using this definition, pixels with similar values from two different objects may be considered the same object as seen in Figure 11. Another example would be to consider the orientation or location of pixels with similar values to determine if they are connected. In this manner only pixels that are within a certain number of pixels, or in a specific direction from the central pixel of a neighborhood, would be considered connected. The concept of using shape or structure to determine the connectivity of pixels is termed Morphology. Morphological processing is geometrically based and is related to the analysis of an image to determine the manner in which the image matches a predefined structure. Morphological operations are more clearly defined for binary images and therefore the application to binary images will be presented here. Extensions of morphological operations to gray level image analysis can be found in several references.

All morphological operations can be defined from two principle operations. These operations are called erosion and dilation. Erosion is used to exclude pixels as belonging to a ROI, while dilation adds pixels to the ROI. The underlying concept for these morphological operations is to erode or dilate a ROI depending on the shape or structure of the applied kernel or structuring element. Using a 3x3 kernel there are two primary structuring elements used (see Figure 12). The first considers only the vertical and horizontal elements of the kernel, termed erode-4 and dilate-4, and the second considers all elements in the kernel, termed erode-8 and dilate-8. Other more complicated and larger kernel and structuring element configurations can be considered if a specific object is being analyzed. To perform erosion or dilation the pixels surrounding the central pixel in the neighborhood are compared to the structuring element and pixels are either removed or added to the ROI based on this comparison.

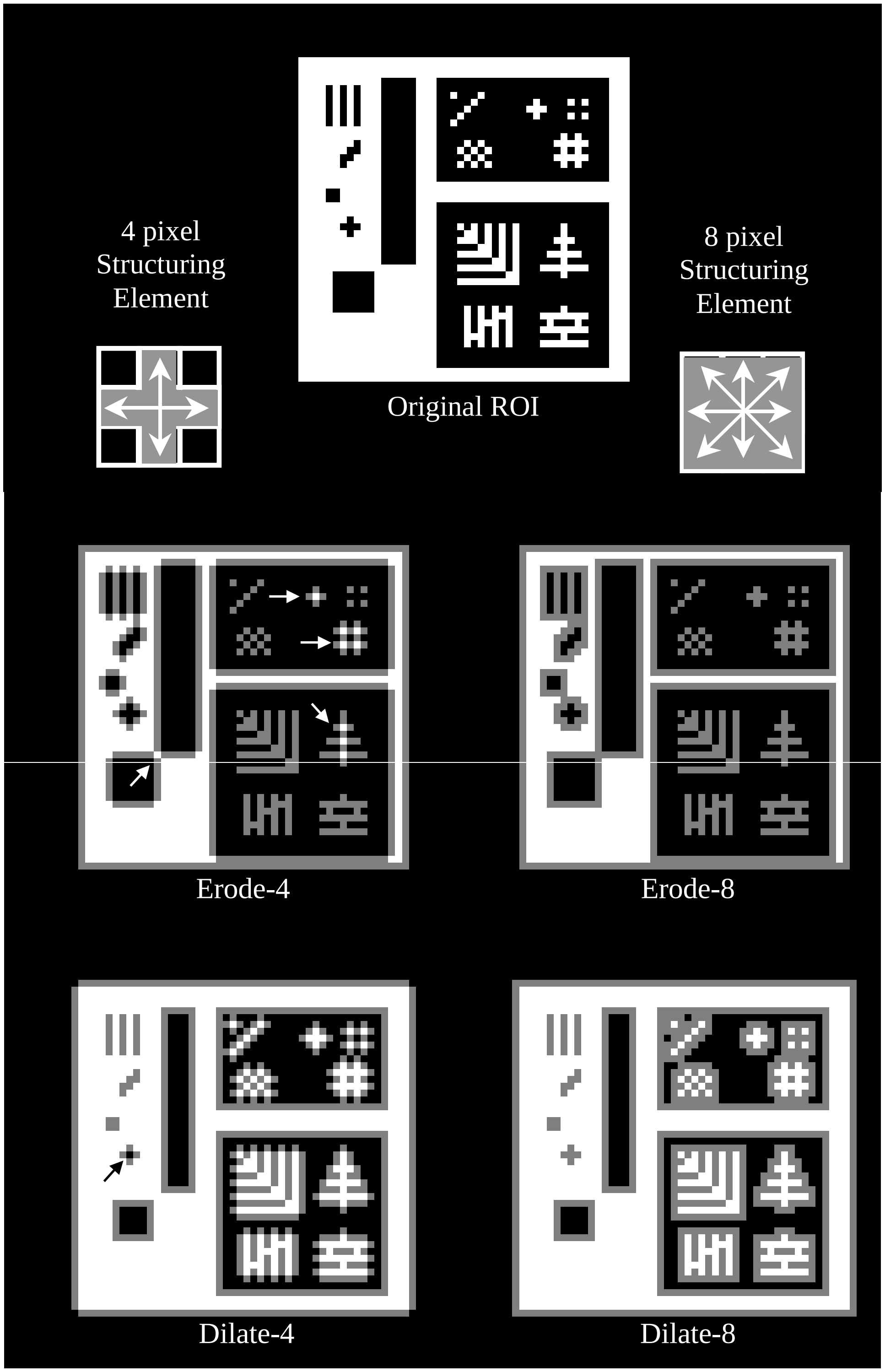

Figure 12

For erosion, if the pixels surrounding the central pixel match the structuring element, the central pixel is not removed from the ROI. If they do not match the structuring element, the central pixel is removed from the ROI. For dilation, if the pixels surrounding the central pixel do not match the structuring element then the pixels corresponding to the structuring element are added to the ROI, i.e. pixels are added to the ROI to make it match the structuring elements. If the pixels match the structuring element, no pixels are added to the ROI. Therefore, it can be interpreted that the erosion and dilation operations assume the central pixel is not on the edge of the feature or object if the surrounding pixels match the structuring element. An example of the erosion and dilation operations for a simple object is shown in Figure 12. Note that when the erode-4 operation is performed the pixels located at the corners of the object and those pixels that are connected to neighboring vertical and horizontal pixels (i.e. those pixels that match the structuring element) are not included in the erosion operation. This results in smoothing of sharp changes in the shape of the ROI (e.g. square corners in the ROI become rounded) and some small objects are reduced to points. The erode-8 operation removes all small objects and all boundary pixels are removed maintaining the general shape of the original ROI. The dilation operation results in larger areas being filled using the dilate-8, whereas the dilate-4 does not fill in all structures (see arrow) and the boundary of the complex objects are smoother following the application of the dilate-8 operation. Using erosion and dilation operations several procedures for removing small clusters of pixels, filling in holes or connecting discontinuous ROI can be performed easily and rapidly. The decision to use either structuring element or to use elements that are more complex depends on the desired results and the shape of the objects or gaps in the original ROI.

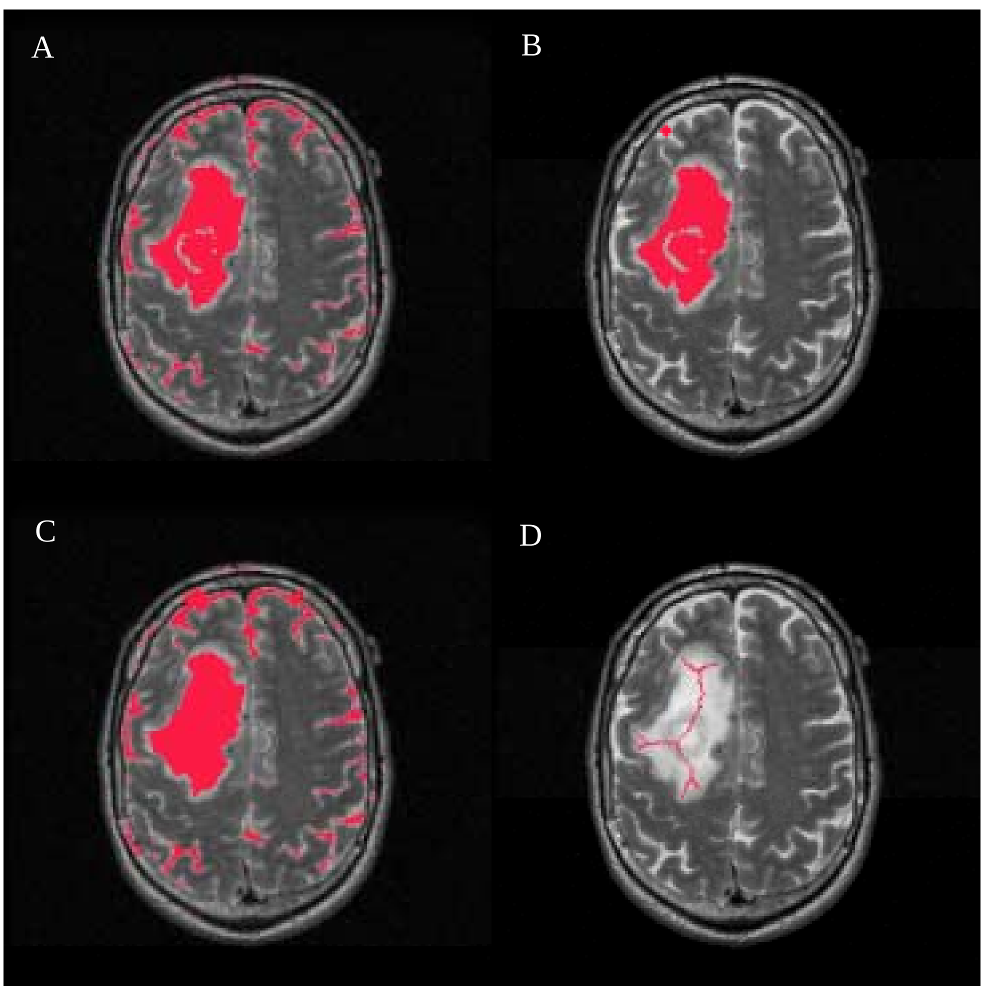

Figure 13

From the erosion and dilation operations many other morphological operations can be derived. These include opening, closing, boundary extraction, skeleton and point operations. Several of these methods are shown in Figure 13. The opening operation generally smooths the boundary of the ROI, widens or increases the size of internal holes or gaps within the ROI, and removes clusters of isolated pixels (see Figure 13B). The opening operation is done by first applying erosion followed by dilation. The amount of smoothing or increased size of openings in or between ROI can be changed by applying multiple erosions followed by dilations. The closing operation is the opposite of opening and consists of dilation followed by erosion. Closing is used to fill gaps in the ROI and connect features. The size of the gaps filled and the distance between objects that are connected can be modified by increasing the number of times the dilation is performed prior to the erosion operations (see Figure 13C). Boundary extraction using morphological operations is done by applying erosion and then finding the difference between the original and eroded ROI. As stated above boundary extraction is useful in many processing methods where the outer contour or edge detection is desired. Using morphological operations boundary definition can be done quickly. The skeleton operation is used to determine the basic geometrical framework underlying the ROI. Skeleton operations are accomplished by performing erosion and opening simultaneously until the ROI is completely removed by the erosion operation. The skeleton of the ROI is then reconstructed by combining the difference between the erosion and opening results from each step (see Figure 13D). The point operation is similar to the skeleton except it is done by applying erosion until all objects are reduced to only one point. If it is assumed the ROI defines a feature or object that is made of a uniform material, the point and skeleton operations can be interpreted as the center of mass and central axis of the ROI, respectively. In addition, the point operation can be used to count objects since following its application each object is reduced to a single point and therefore the number of points remaining equals the number of objects that were present in the image. Many more operations can be developed by combining erosion and dilation operations in varying forms. These methods are always fast and require minimal computer time making them very powerful processing tools.

3.2.2. Multi-Resolution Methods

Other methods for dealing with connectivity of features have been developed using multi-resolution schemes . In these methods discontinuities in a ROI can be removed by changing the overall image resolution to determine the connectivity of objects. This approach uses the fact that the general shape of an object remains at lower resolution and this general shape can be used to guide the connection of objects at the higher resolution. An example of a multi-resolution method is shown in Figure 14. In this method the contour points that define the object are found by an edge-detection or boundary extraction algorithm at several image resolutions, e.g. image sizes of 256x256, 128x128, and 64x64. Then starting from a seed point an edge tracking algorithm is used to define what part of the contour is from the object of interest at each resolution (note, the seed point can be the same for all resolutions). The defined object contour at the lowest resolution is then compared to the contour found at the next higher resolution to determine which pixels at the higher resolution should remain. The comparison can use several different criteria for this comparison. In the example method the criterion used was the distance between contour points at each resolution. If the pixels from the contour at the higher resolution were within a pre-defined distance from the corresponding lower resolution contour points these pixels were kept. If the pixels were not within the distance criteria the pixel was removed from the higher resolution contour. This procedure is repeated using the contour pixels that remain from each resolution until the original resolution contour is corrected. Following the application of this multi-resolution method the contour may have several missing segments due to pixels being excluded in the procedure. These gaps can be filled using straight lines or curves through the adjacent contour points depending on the needs of the analysis.

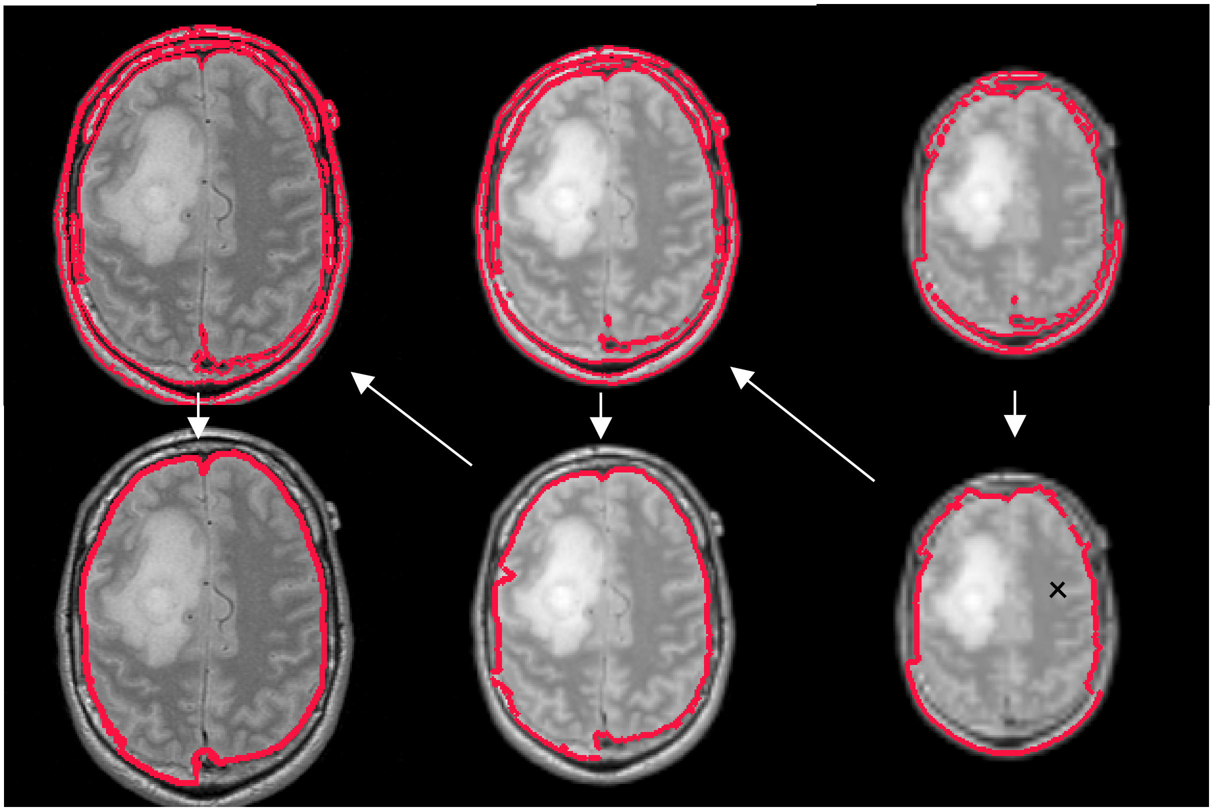

Figure 14

3.3. Histogram Analysis

A gray level histogram of an image is a plot of the number of pixels at each gray level. A histogram is another type of data representation and can be used to infer characteristics related to the contrast between objects. Many image-processing methods use some form of histogram analysis in their procedures. For example, in many of the methods discussed above the analysis included the definition of a range of gray levels to include in the processing. The definition of a range of gray level values is termed gray level slicing. In Figure 15 the gray level histogram for the second fast spin echo image shown in Figure 2 is displayed. By choosing a lower and upper gray level value on the histogram a ROI corresponding to all pixels in that range can be defined (see Figure 15). The gray level values chosen are called threshold values. Many histogram analysis methods are concerned with determining threshold values to create ROI that can be used to segment desired objects in the image. The most direct way of determining threshold values to segment objects is to choose the minima between the peaks seen in the histogram. If the contrast between features is large, each object will be displayed as a distribution of gray levels whose peak values correspond to the average gray level for the object. This assumes the spread of the gray level values around the average values are not too large so they do not obscure the distinction between objects. The spread of the gray level distribution is due to additive noise and partial volume averaging of different features. Since the spread of the gray level distribution is due in part to the noise in the data, pre-processing filters that reduce noise may improve the histogram analysis results. In addition, knowledge of the type and amount of noise in the data is often used to define the histogram analysis method.

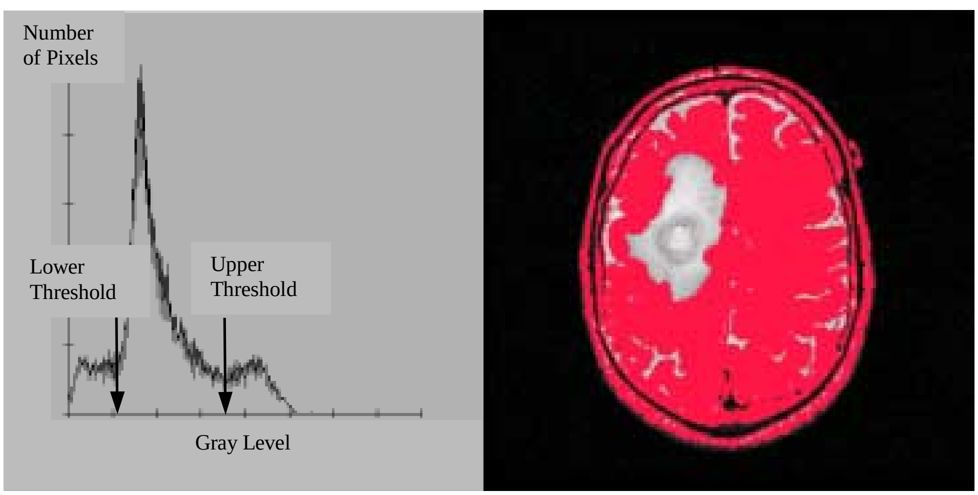

Figure 15

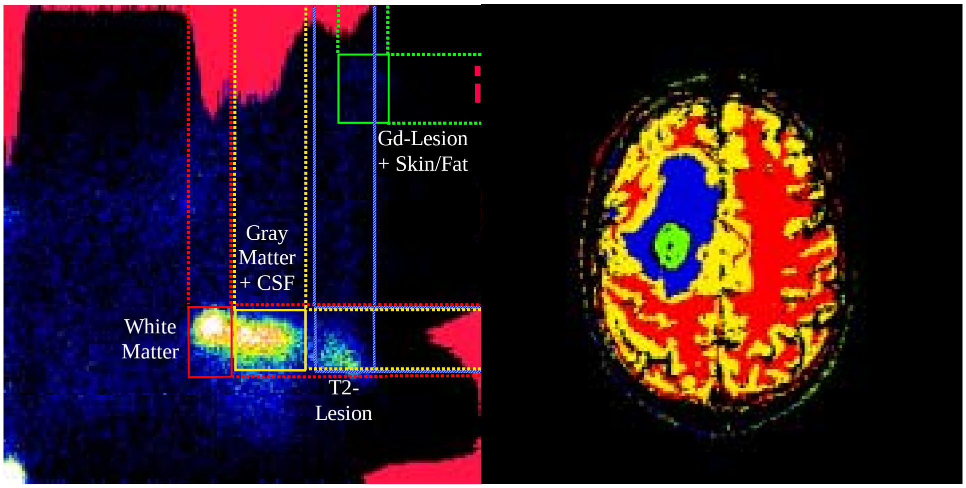

Figure 16

If two or three images are acquired the histogram from each image can be combined to produce a 2D or 3D histogram of the gray level distributions. Figure 16 shows the 2D gray level histogram created by plotting the second FSE image on the x-axis and the post-Gd T1SE image on the y-axis for the example study. The number of pixels at each gray level, in each image, is displayed as intensity in the 2D plot image. Overlaid onto the 2D plot are the 1D histograms for each image. Also shown are threshold ranges chosen to segment different tissue gray level distributions from each image. The ROI segmenting different gray level distributions in the 2D plot are created by finding the intersection from the ROI made on each individual image histogram. Note that three ranges of gray levels are chosen from the FSE image (i.e. x-axis) and two ranges from the post-Gd T1SE image (i.e. y-axis). In this example this provides the segmentation of four objects that are shown as different color ROI in Figure 16. Note that on the 1D histogram from the post-Gd T1SE the Gd enhancement can be localized; yet in the 2D plot it is lost. This is due to the small number of pixels corresponding to the Gd enhancement tissue in the image and the spread of the data. Therefore, using the 2D histogram the Gd tissue may not have been segmented. If three images are used, a 3D histogram plot can be made and objects can be segmented by selecting threshold ranges as cubes. Using three images allows more subtle tissue differences to be segmented from the plot. When more than three images are used the determination of threshold values cannot be done by visual inspection and more advanced methods of analysis most be used.

Another aspect of histogram analysis relates to improving the overall image quality as opposed to segmentation. Most digital image display and printing utilize some form of histogram analysis to map the pixel values to a suitable range for the display or hard copy. For example in Computed Radiography systems histogram analysis is used to enhance the gray level distributions in some regions of the image while suppressing others, e.g. the background signal is suppressed. These types of processing can be termed adaptive histogram modification. Similar to adaptive filtering, adaptive histogram modification applies different analysis methods to different regions of the image based on some a priori knowledge used in the development of the method. In addition to adaptive histogram analysis, digital image display and printing require that data from different gray levels within an image be mapped differently to ensure contrast and details are not lost. This requirement is due to limitations of display systems and the human visual system in distinguishing gray level differences, i.e. visualize contrast at different gray levels. Therefore, all digital image display and printer systems apply a histogram transformation to the digital data to provide the optimal display of the gray level information. In fact, film can be considered an analogue version of a histogram transformation since the log of exposure of a film/screen system maps optical density differently based on the characteristic curve of the film/screen system. Although histogram transformation methods and concepts are extremely important in the interpretation and analysis of digital images it is beyond the scope of this paper to review these methods.

3.3.1. Bayesian Classifiers

From a histogram of the gray levels in an image it can be seen that different objects from the image can be represented by individual distributions and peaks. Therefore, a method for finding peaks and distributions may be good at segmenting the corresponding objects in the image. In Figure 17 the gray level histogram for the first fast spin echo image shown in Figure 2 is displayed and regions corresponding to different tissues are shown as Gaussian distributions overlaid onto the histogram. The decision to use Gaussian shaped curves to define the distributions was a choice made by the operator. Other distributions such as Poisson, Lorenzian, Rician etc, may be more appropriate based on knowledge of the additive noise or other physical aspects of the acquisition. From the distribution chosen the peak value and the spread of the data around the peak can be modeled. In the case of a Gaussian distribution two parameters must be known to create the distributions, these are the mean and standard deviation of the distribution. The mean can be determined from the average gray level of different tissues in the image or as gray levels corresponding to the different peaks visualized in the histogram. If the spread in the data is assumed to be due to noise only, then the standard deviation of the noise may be a good estimate of the distribution spread. Care must be taken when defining the noise from the standard deviation of the acquired signal. First, the region used to define the standard deviation must be from an area where the signal is uniform. If the region includes signals from several tissues, i.e. partial volume effects, this will cause the spread in the data to be larger than the standard deviation due to noise alone. Second, if the distribution of the signal from the region used does not match the distribution being modeled, then the standard deviation determined may again be incorrect. This is observed in MRI where the noise distribution in magnitude images is Rayleigh or Rician, not Gaussian. Although most analysis methods will assume a normal or Gaussian distribution this assumption is justified only for signals from regions that have high SNR. Therefore, in MRI care must be taken to use a region that has a large SNR, e.g. do not use the image background where the signal is close to zero to define noise and a Gaussian model does not apply.

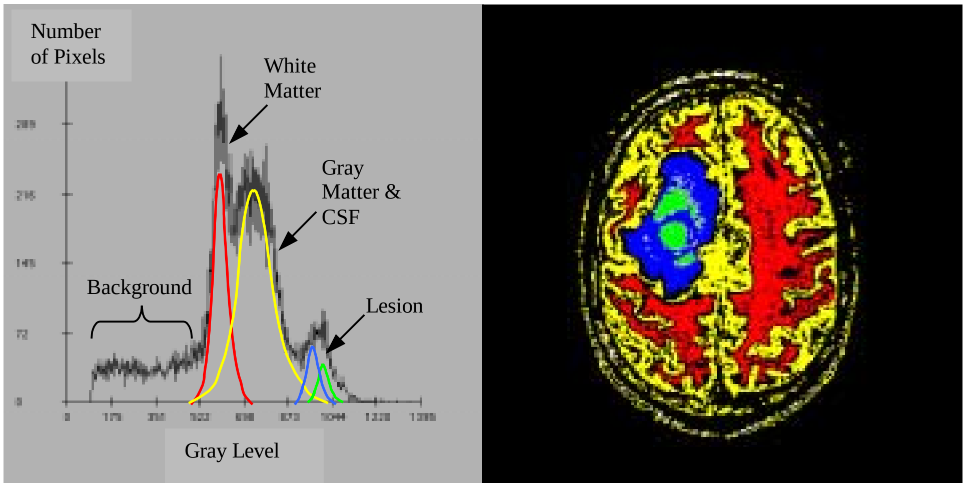

Figure 17

As stated above by using the knowledge of the noise in the data a model of the distribution can be determined. From this the probability of segmenting tissues based on the model can be derived. In many methods segmentation is optimized by minimizing the error in the classification of tissues into each distribution defined. That is, the image analysis method minimizes the error in classifying pixels as belonging to one tissue or another based on the modeled distributions and probability of incorrect classification. One classification method that has been widely used is the Bayes Classifier. The use of a Bayes Classifier is shown in Figure 17 where the distributions were modeled as Gaussian distributions. The intersection of two adjacent Gaussian distributions would be the minimum error in the classification of these tissues and therefore this value would be chosen as the threshold to segment the tissues. As stated above the most direct way of determining threshold values to segment tissues is to choose the minima between the peaks seen in the histogram. Note in this example the distributions found by the Bayes Classifier has intersections between the tissue distributions that are close to the minimum seen in the original histogram, although they are not exactly the same. The exception to this is seen in the two curves defined for the lesion components. The definition of two distributions for these tissues was not visible on the original histogram but the Bayesian Classification segmented these features. Note that the segmentation using the Bayesian Classifier is similar to the manual thresholding of the 2D histogram seen in Figures 16. However, since the Bayesian Classifier did not rely on an operator to define threshold values more reproducible segmentation results would be expected.

More advanced methods for determining the threshold values to segment tissues using Bayes Classifiers or other statistical methods have been developed. In addition, if the assumptions of the distribution are accurate, e.g. the mean and standard deviation determined are accurate, then the partial volume of each tissue can be inferred based on the modeled distributions. A limitation of this method for determining partial volume effects is in the case when more than two tissues are present in a voxel. When three or more tissues are present in a voxel the distinction between peaks and distributions for individual tissues may not be possible and therefore statistical methods utilizing one image may not be able to segment tissues accurately, although incorporating more images into the analysis may allow results that are more robust.

3.4. Multi-Image Filters – Linear Filters

Similar to the use of multiple images in neighborhood analysis many image analysis methods use the information from different acquisitions to perform feature extraction and segmentation. In these filters the information for each pixel are combined to produce one or more composite images where different features or tissues are segmented. In this review linear filters will be discussed since they are used in many applications and the results are easy to interpret. Non-linear filters and other methods such as neural networks and artificial intelligence are also being developed but will not be discussed here.

Often the purpose of acquiring multiple images of the same location is to obtain differing contrast between tissues or structures to aid in the interpretation of the images. Traditionally the person viewing the images segments the different tissues based on these contrast differences by looking at each image and integrating this data in their head. A linear filter integrates images by performing a weighted summation of the images. As with the other filters discussed the weighting factors are determined by the filter being applied and the results varydepending on how the weighting factors are applied. Some common linear filters will be reviewed; these are the Principle Component Analysis, Gram-Schmidt Orthogonalization and k-Means clustering methods.

3.4.1. Principle Component Analysis

Principle Component Analysis (PCA) is a statistical analysis method that results in linear combinations of the acquired images that maximize the variance in a signal. Other terms used for PCA include Karhunen-Loeve Transform, Hotelling Transform and Eigenvector decomposition. The term variance can be thought of as the difference in the signal from tissues, i.e. the contrast between tissues. If the signals for all tissues in an image were averaged, the standard deviation of this average would be a measure of the variance in tissue signals. In this definition tissues that have the maximum contrast will also provide the maximum variance. In the application of PCA a new coordinate system is determined such that the principle axis is along the direction of the maximum variance. Then another axis that is orthogonal to the first is determined such that the signal variance in this direction is maximized. This procedure is continued until the number of axes determined is equivalent to the number of original images input. An example of the application of the PCA using the second FSE image and the post-Gd T1SE image is shown in Figure 18. To demonstrate the application the 2D histogram for these images is plotted. Note in the plot the direction of the maximum signal variance lies along a line labeled PCA-1. This line, or principle axis, is determined by the PCA method and then the second PCA axis is calculated based on this first principle axis and the signal variance in the orthogonal direction to this axis. In two dimensions the second principle axis can only be in one direction, but if three or more images were used the additional principle axes would be determined such that the variance along each axis is also maximized. The PCA images can then be created by plotting the signals onto each of these new axes. The two PCA images created for this example are shown below the plot. Note that the first PCA image has a high SNR for all tissues but the CNR between tissues is not high. The second PCA image shows the Gd enhancement portion of the lesion and the surrounding skin and fat since the signal from these tissues lie well off the PCA-1 axis and the signal from the other tissues are projected close to zero. Also, note that the signal corresponding to the Gd enhancement is included in both PCA images but as in the 2D histogram shown in Figure 16 the signals from this tissue cannot be visualized in the plot due to the small number of pixels corresponding to the Gd enhancement.

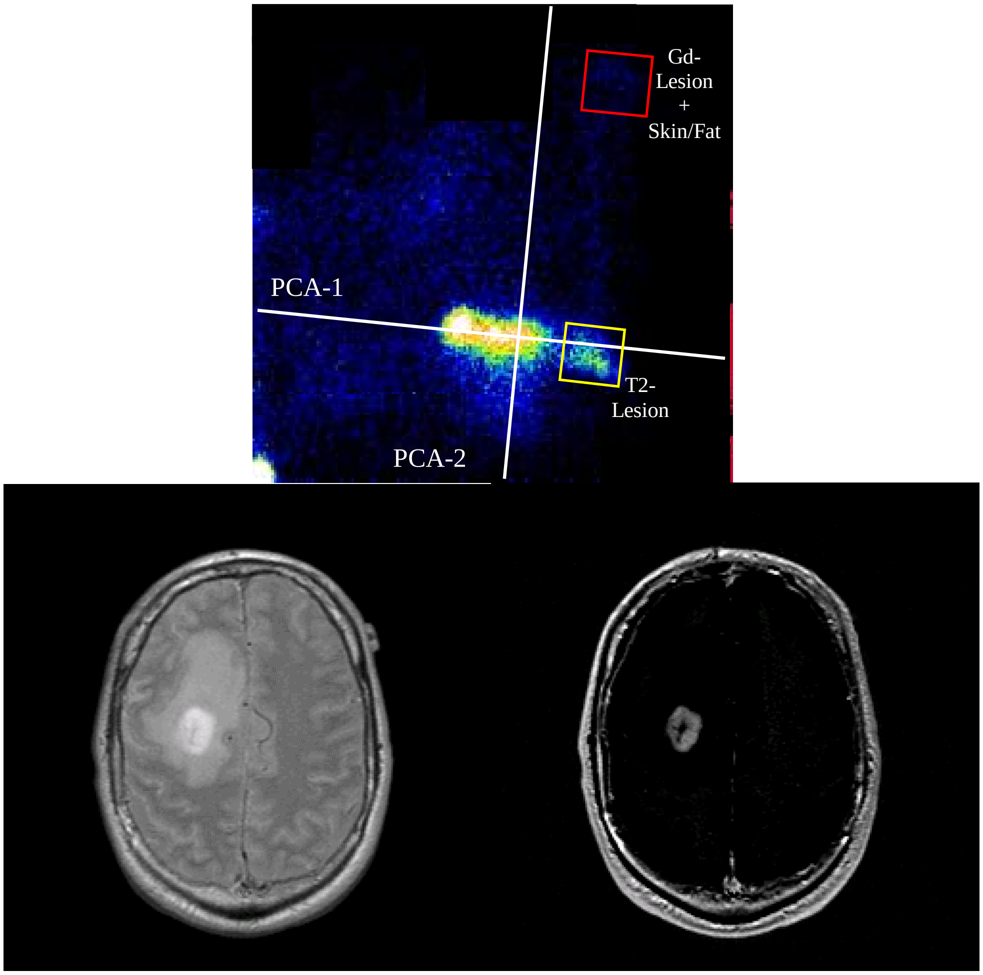

Figure 18

The PCA has been utilized to enhance the visualization of signals in many different modalities but it does not necessarily result in the segmentation of any signal. The segmentation of a signal will occur only if the signal variance is maximized onto one of the PCA axis. There is no way of determining or selecting the tissues that will be segmented and therefore its use as a segmentation method may be limited. On the other hand, a major attribute of the PCA is that it reduces the dimensionality of the data. This is achieved because the first few PCA images have the maximum SNR and they contain the maximum variance information from the original data. That is, if the information contained within the multiple images acquired are correlated then the first few PCA images will account for most of the signal variation and the later PCA images will show little or no information. This is especially true if the images can be reduced to linear combinations of each other. Because of this feature, PCA has been utilized as a compression methodology where the information contained in several input images is reduced to one or two PCA images.

3.4.2. Gram-Schmidt Orthogonalization

Often the size and shape of an object may appear different in the multiple contrast images acquired. This may in part be due to tissues overlapping each other, i.e. partial volume effects, and therefore reduces the visualization of the tissue of interest. If it is assumed the signals from different tissues sum linearly in a voxel, or the signal obtained can be reformatted or processed to ensure a linear relationship, than these different signals can be separated using a filter termed the Eigenimage filter. This filter is also known as the optimal linear transformation for partial volume extraction and the Gram-Schmidt Orthogonalization filter. The Eigenimage filter can maximize the SNR from a desired tissue signal while suppressing the signals within each voxel from other tissues. In this way, the Eigenimage filter has the ability to correct for partial volume averaging effects, allowing the visualization of the partial volume for each tissue to be displayed. The separation of the signals from different tissues within a voxel is accomplished through vector analysis. The Eigenimage Filter is applied by an operator selecting ROI inside tissues or areas that are assumed to be from voxels that contain only one tissue. The average signal change in each image from the ROI defines a signature, or more accurately a signature vector, of the contrast change for that tissue based on the image sequence used. Once the signature vectors for all tissues are defined the Eigenimage Filter applies the Gram-Schmidt Orthogonalization where the tissue to be segmented is used as the last vector in the Gram-Schmidt process. An example of the application of the Eigenimage filter is shown in Figure19. In this example the image sequence shown in Figure 2 was used and ROI were defined for white matter, gray matter, CSF, and two components in the lesion. These components were the Gd enhancement and the surrounding hyper-intensity seen on the Flair image. In Figure 19 images segmenting each of these tissues from the others are shown. Note ROI corresponding to the outer skin and other tissues were not defined by the operator, so their contrast changes in the image sequence were not used in the application and the voxels containing these tissues were not removed from the resulting segmented images.

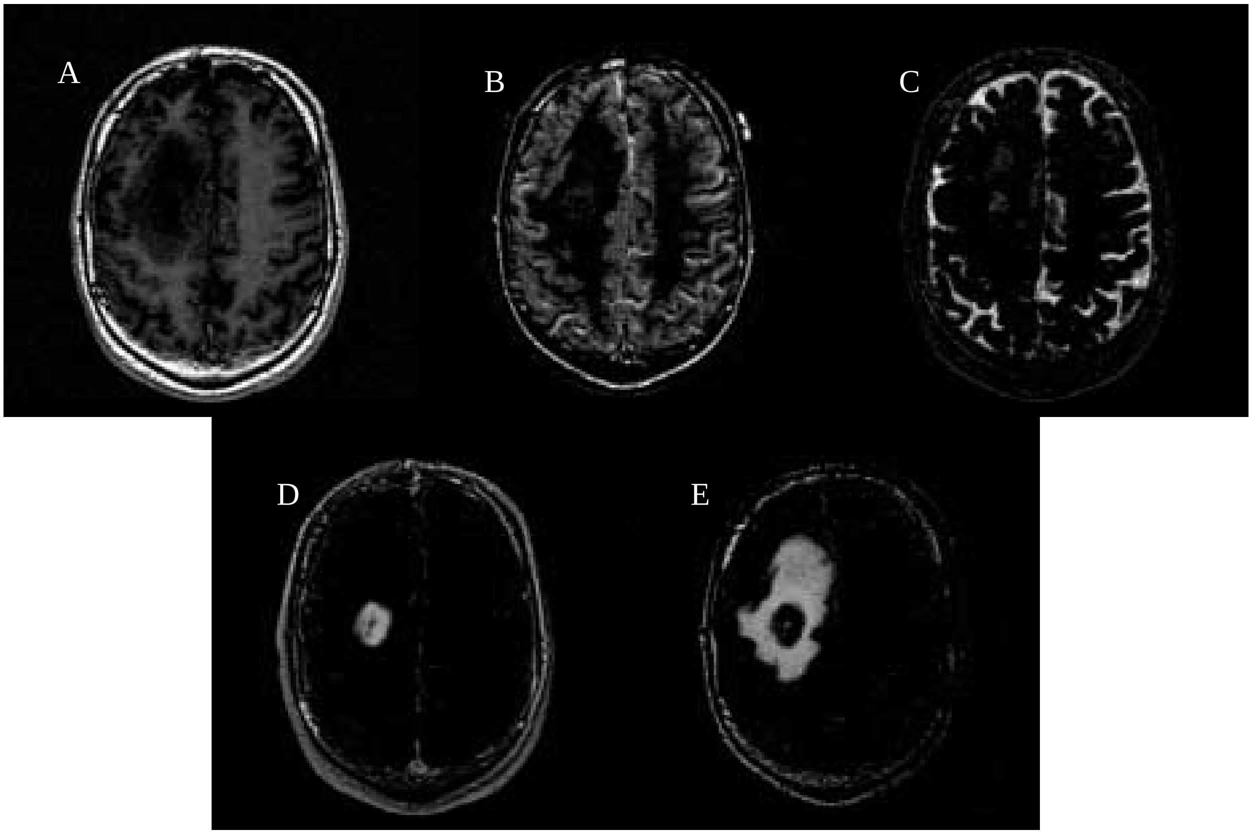

Figure 19

The Eigenimage Filter does not result in a maximum SNR or CNR in all instances, but these quantities can be improved by changing the parameters used to acquire the signal or by processing the signal before or after the application of the transformation. In addition, since the Eigenimage Filter is a linear filter that determines an analytical solution to the segmentation the results are obtained in less than one second/image allowing real time image interpretation.

3.5. Cluster Analysis/Fuzzy Analysis

Examples of K- and C-means cluster analysis results are shown in Figure 20.

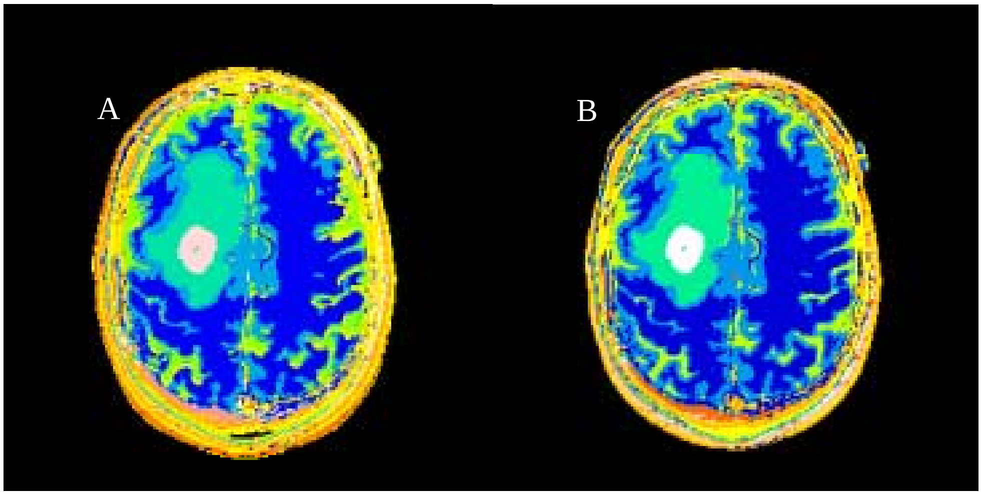

Figure 20

3.6. Considerations/Caveats to Feature Extraction and Segmentation Methods

4. Discussion