**Background for LIDC Analysis**

**Draft, 10/27/2013, Ron Niehaus**

**Part I - Introduction**

1. **Data Sources**

Datasets typically analyzed for the LIDC program consist of data consolidated from three primary sources, as follows:

* XML files that describe each image, and provide the semantic ratings associated with each image.
* DICOM Headers that provide details about each image, and the conditions under which the image was produced.
* Data extracted from the images themselves or from the outlines of the images that were delineated by each radiologist, so called “content-based features.”

1. **Data Relationships**

The following basic relationships are important to understanding the data:

* Each nodule consists of many slices or two-dimensional views.
* Each nodule is given semantic ratings by between one and four radiologists.
* Each slice of each nodule is annotated with an outline by one to four radiologists; therefore the features will be different for each radiologist case, in addition to the semantic ratings being different.
* Radiologists IDs are randomly assigned on each annotation. Therefore, you cannot assume radiologist number 1 for case 2 is the same as radiologist number 1 for case 3.
* One image exists for each annotation of each slice by each radiologist.
* Content-based features are generally calculated for each image (radiologist outline).

1. **Datasets**

There are over 40,864 images in the total dataset representing one image per slice per radiologist per nodule (file: LIDC\_20130817\_AllFeatures2D\_AllSlices.xls). Content-based features are calculated for each image, and are divided into the following general categories – size, shape, intensity, and texture. Semantic ratings are by nodule-radiologist pair. Therefore, all slices for the same radiologist for the same nodule will have the same set of semantic ratings.

There are 6657 unique radiologist-nodule pairs in the LIDC data. For this dataset, feature set aggregation is done by selecting the largest slice by area (in pixels) for each radiologist (FILE: LIDC\_20130817\_AllFeatures2DMaxSlicePerRad\_inLineRatings.xls). Semantic ratings are not aggregated, since each radiologist-nodule pair is associated with one set of semantic ratings.

Datasets generated with one representative slice per nodule will therefore contain one case per nodule, or 2656 cases (file LIDC\_20130817\_AllFeatures2D\_MaxSlicePerNodule\_inLineRatings.xls). Note that this implies that both semantic ratings and feature sets must be aggregated. For the standard dataset, the feature vectors are aggregated by selecting the features calculated for the largest slice by area (in pixels) for each nodule. The semantic ratings are aggregated by selecting the semantic ratings associated with the same slice.

Since many analyst wish to combine semantic ratings in a different way, for example, as probabilities or as the mode of the individual ratings, all of the semantic ratings are provided “in-line” for the aggregated datasets. These dataset file names have the words “inline rating”.

Since many images are two small to be useful (including as small as a single pixel in size), some datasets will have less than the total possible cases, because some nodules are filtered out as outliers based on size. For example, nodules are often clinically defined as between 3mm and 30mm in size, yet some of the outlines are of regions smaller or larger than this size.

**Part II – XML File Reference Information**

XML Files – A Matlab program was created that parses the XML files that exist in the LIDC file structure and produces a comma-separated file that is converted into a spreadsheet, typically with a name like parsedXML. The data elements in this file include both reference information and semantic ratings of the nodule provided by the radiologist. The reference information provided includes:

1. InstanceID – an image instance associated with a radiologist rating of a nodule slice.
2. ImageName – the image names are formed from the instanceID with the DICOM extension, as in 4.DCM.
3. StudyInstanceUID –
4. SeriesInstanceUID –
5. RadiologistID – a number from 1-4, used to distinguish the four radiologists that annotate a nodule.
6. NoduleID – designates a nodule.
7. FilePath – a path to the XML file in the directory structure.
8. ImageZPosition – a number indicating the vertical position of the slice
9. Coords – a delimited list of coordinates that define the outline annotated by the radiologist.
10. RadiologistID – Radiologist ID is a sequence number given to distinguish one radiologist annotation from another. It does not designate a particular radiologist, since the numbers are randomized for each case.
11. NoduleID – Designates the nodule.
12. Subtlety – Semantic ratings of a nodule by a radiologist.
13. Internal Structure – Semantic rating of a nodule by a radiologist.
14. Calcification – Semantic rating of a nodule by a radiologist.
15. Sphericity – Semantic rating of a nodule by a radiologist.
16. Margin – Semantic rating of a nodule by a radiologist.
17. Lobulation – Semantic rating of a nodule by a radiologist.
18. Spiculation – Semantic rating of a nodule by a radiologist.
19. Texture – Semantic rating of a nodule by a radiologist.
20. Malignancy – Semantic rating of a nodule by a radiologist.

**Part III – DICOM Header Information**

DICOM Headers that provide details about each image, and the conditions under which the image was produced. Data extracted from DICOM headers include pixel size and spacing between slices.

**Part IV – Content-Based Features**

1. **Overview**

Content-Based features are extracted from the images themselves or from the outlines of the images that were delineated by each radiologist. Features extracted from the images have included those listed below. In some cases, individual analyst will include additional features than those shown.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Size** | **2D Shape** | **Intensity** | **Texture** | | |
| Area  Perimeter  ConvexPerimeter  EquivalentDiameter  MajorAxisLength  MinorAxisLength | Circularity  Roughness  Compactness  Solidity  Eccentricity  Extent  RadialDistanceSD  SecondMoment | MinimumIntensity  MeanIntensity  SDIntensity  MinIntensityBG  MaxIntensityBG  MeanIntensityBG  SDIntensityBG  IntensityDifference | Haralick  Correlation  Energy  Contrast  Entropy  Inverse Difference  Max Probability  3rd Order Moment  Sum Average  Variance  Cluster Tendency  Absolute Value | Gabor  Gabormean\_0\_0  GaborSD\_0\_0  GaborMean\_0\_1  GaborMean\_0\_2  GaborSD\_0\_2  GaborMean\_1\_0  GaborSD\_1\_0  GaborMean\_1\_1  GaborSD\_1\_1  GaborMean\_1\_2  GaborSD\_1\_2  GaborMean\_2\_0  GaborSD\_2\_0  GaborMean\_2\_1  GaborSD\_2\_1  GaborMean\_2\_2  GaborSD\_2\_2  GaborMean\_3\_0,  GaborSD\_3\_0  GaborMean\_3\_1  GaborSD\_3\_1  GaborMean\_3\_2  GaborSD\_3\_2 | Markov  Markov 1  Markov 2  Markov 3  Markov 4  Markov 5 |

**2. SizeAndShape2D**

Calculates size and shape features for a single image instance. It can be used to calculate any number of size and shape features by commenting or uncommenting section of code. It returns a structure array with the following properties:

Area

Area is the count of the pixels included in the region

Perimeter

Perimeter is calculated by obtaining an ordered list of all boundary pixels, and then calculating the Euclidean distance along the lines connecting the midpoints of the exposed facets of each boundary pixel. See Figure below.

|  |
| --- |
| Facet Midpoint Method of Perimeter Calculation |
| FacetMidpoint.png |

ConvexPerimeter

The convex perimeter is the perimeter of the convex hull of the shape. The convex hull is the minimum convex boundary that encloses the region. The convex hull can be visualized by stretching a rubber band around the shape. The convex perimeter is calculated by obtaining a list of all vertices of the convex hull, and then calculating the Euclidean distance between adjacent vertices.

EquivalentDiameter

Specifies the diameter of a circle with the same area as the region. Computed as sqrt(4\*Area/pi).

MajorAxisLength

Specifies the length of the major axis of the ellipse derived from the eigenvectors of the covariance matrix of the binary mask image.

MinorAxisLength

Specifies the length of the major axis of the ellipse derived from the eigenvectors of the covariance matrix of the binary mask image.

Circularity Index

The circularity index is obtained by dividing the area of the region by the area of a circle with the same convex perimeter. It is a measure of the degree to which the region conforms to a circle. The equation reduces to:

The circularity index varies between 0 and 1, with a perfect circle being a 1.

Compactness

Compactness is a measure of the degree to which the area and perimeter of a region conforms to the relationship between the area and perimeter of a circle. The circle is the most compact shape in the Euclidean plane in terms of minimal perimeter for a given area.

Roughness Index

The roughness index is equal to one minus the ratio of the perimeter of the convex hull to the perimeter of the shape. The convex perimeter should always be less than the shape perimeter, so the index varies between 0 for a very simple shape to 1 for a complex shape with many concavities.

Roughness is related to circularity and compactness by the following equation.

Roughness =

Solidity

Solidity is the ratio of the area of the region to the area of the convex hull. It is a measure of region complexity and low values may indicate complicated regions or a boundary with many concavities.

Extent

Extent is the proportion of pixels in the bounding box that are also in the region. It is a measure of complexity, similar to solidity, only the bounding box is used as the reference instead of the convex hull. The bounding box is the minimal rectangle that contains the region.

Eccentricity

Eccentricity is the ratio of the distance between the focus of an ellipse (that has the same second moments as the region) over the major axis length. In this context the second moment probably refers to the covariance matrix for all points in the region. The major and minor axis of the ellipse are derived from the eigenvalues of this covariance matrix. Elongation and eccentricity are algebraically related by the following equation:

Radial distance SD

Radial Distance SD is the standard deviation of the distance from every boundary pixel to the centroid of the region. The centroid is determined from the Matlab regionprops function.

Second Moment

Second Moment is the sum of the radial distances squared divided by the area.

Elongation

Elongation is the ratio of major axis length to minor axis length. The major and minor axis are in reference to the ellipse with the same second moments as the region. In this context the second moment refers to the covariance matrix for all points in the region. The major and minor axis of the ellipse are derived from the eigenvalues of this covariance matrix. Elongation is algebraically related to eccentricity (see below). Elongation is equal to one over the square root of one minus the eccentricity squared:

**Intensity Features**

Intensity features are calculated by ..

Intensity features are defined in the following table:

|  |  |
| --- | --- |
| MinimumIntensity | The intensity of the lowest intensity pixel value for those pixels within the boundary of the nodule. |
| MeanIntensity | The mean intensity of all pixels within the boundary of the nodule. |
| SDIntensity | The standard deviation of the intensity of all pixels within the boundary of the nodule. |
| MinIntensityBG | The intensity of the lowest intensity pixel value for those pixels in the background (not within the boundry of the nodule). |
| MaxIntensityBG | The maximum value of the intensity values of all background pixels after masking the foreground(nodule) pixels. |
| MeanIntensityBG | The mean value of the intensity values of all background pixels after making the foreground (nodule) pixels. |
| SDIntensityBG | The standard deviation of the the intensity values of all background pixels after masking the foreground(nodule) pixels. |
| IntensityDifference |  |

–

**4. Texture Features**

Texture features attempt to extract periodic patterns in intensity. Texture features include

Haralick Features

Haralick features are based on the co-occurance matrix of the image, which defines the probability of a value repeating at a particular offset and orientation from a specified location. It is the joint probability of two pixels a specified distance apart in a specified direction having the same value. Haralick features include energy, entropy, correlation, Homogeneity, 3rd Order Moment, contrast, sum average, cluster tendency, and maximum probability. They are defined as below.

|  |  |  |  |
| --- | --- | --- | --- |
| *Correlation =* |  | *Energy =* |  |
| *Contrast =* |  | *Entropy =* |  |
| *Inverse Difference =* |  | *Max Probability =* |  |
| *Third Order Moment =* |  | *Sum Average =* |  |
| *Variance =* |  | *Cluster Tendency =* |  |
| *Absolute Value =* |  |  |  |

Markov Features

Markov features are derived from the concept of Markov Random Fields (MRFs). Within the context of a digital image, the Markov property indicates that the probability of a pixel having a particular intensity value is dependent directly only on its immediate neighbors. However, the value of each pixel depends indirectly on its larger neighborhood through propagation.

Gabor Features

Gabor features are interesting for many applications, because they model certain aspects of the human visual system. Gabor features employ a linear filtering technique that detects periodicity (repeating patterns) in intensity. The spatial frequency, the spatial frequency bandwidth, and the orientation of the periodicity are determined by parameters specified for each Gabor feature. Typically, banks of gabor features are used to cover orientations and spatial frequencies of interest to the application. For LIDC, for each of gabor parameter set, the gabor mean and standard deviation is calculated.