**HMS Lung Cancer 1 Marathon Match - Solution Description**

This

approach is



•

•

value



•

by

over

time.

•

slice is

roughly

This

made it

easy

test

•

of

simple

•

my

•

for

**1. Approaches Considered**

Use image matching to improve determination of

lung area.

Make use of

a training step to record tumor

characteristics such as

most probable locations, gray

composition and common shapes.

**2. Final Approach**

Reasons for Selecting Final Approach:

approach was

evolved gradually

adding small

improvements

Fast execution time:

the required time per

5 seconds.

to quickly

changes.

No external

libraries: gives

complete control

the code and makes

deployment

Approach Summary:

Each slice is divided into a grid of 4x4 pixel squares. A square is kept if it is determined to be within the

lungs and the gray value of all pixels within that square are above a certain threshold. Next, adjacent kept

squares are grouped together to form clusters. Groups of clusters are produced by matching clusters

across adjacent slices. Finally, the 'best' group of clusters is chosen as the solution.

Detailed Description:

The code for

found in src/TumorFinder.java

The run method executes

the following steps

each CT scan

Step 1: Determine lung bounds

In this step, both the range of slices containing the lungs and a minimal bounding box enclosing the lungs

for each slice is determined. First the bounding box for the body is found by starting at the edge of the

image and searching inward until the lighter gray pixels of the body are found. The top y value of the body

bounding box is used to find the position of the base of the neck. This is important as a guideline for finding

the start of the lungs. Starting just inside the body bounding box, the lung bounds are found by searching

inward toward the center of the body until the dark area of the lungs is reached. An example of the lung

bounds is shown in the top-left image. The start of the lungs is defined as the first slice with at least 20 dark

pixels in the left or right side of the lung bounds. The end of the lungs is defined to be 8 slices before the

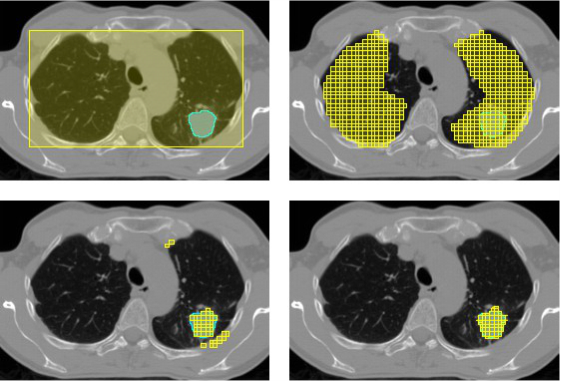
first slice containing less than half the dark pixels of the maximum slice.

Step 2: Find the lung area

The lungs are modeled as a function of slice using a set of 5 pairs of polygons. Depending on slice, the

appropriate pair is stretched to fit the lung bounds.

*Top-Left: The lung bounding box and ground truth (blue)*



*Top-Right: All grid squares within the fitted polygon lung shape*

*Bottom-Left: All clusters formed from squares satisfying gray value requirements*

*Bottom-Right: Final cluster with edges extended to better match the ground truth*

Step 3: Create a grid of 4x4 pixel squares

Squares within the lung shape with a high enough gray value are kept. A small improvement is gained by

removing squares that are near any gray value that is too high. These are likely to be bone and it is rare to

find a tumor near them. The top-right image shows all squares within the lung shape and the bottom-left

image shows squares remaining after gray value requirements.

Step 4: Group kept squares into clusters

All remaining squares are grouped into clusters. The bottom-left image shows 2 large and 3 one square

sized clusters. A score is assigned to give a preference for larger isolated clusters:

score = sqrt(total squares) \* 100 \* (nGood + 1)/(nBad + 1).

nGood is defined as the number of squares near a dark square and nBad is defined as the number of

squares near a non-dark square that isn't kept. For the bottom-left image, this results in a very high score

for the largest cluster and very low scores for the others.

Step 5: Match clusters across slices

3x3 or

the lung shape:

clusters



•



•

•

•

•

Try

5x5,

•

•

output

All clusters with a score of at least 200 are considered in the creation of cluster groups across slices. To be

matched, adjacent clusters must share a reasonable number of squares:

8 \* (total shared squares) ^ 2 >= (squares in cluster 1) \* (squares in cluster 2)

Clusters in subsequent slices are continuously added until a slice with no match is found. The cluster group

score is simply the sum of the scores of the clusters in it. The lung tumor is predicted to be the cluster

group with the highest score.

Step 6: Refine and write out the solution

All squares in the cluster group with the highest score are written out to the solution file. Additionally, the

bounds can be extended by adding parts of adjacent squares if they contain enough non-dark pixels. The

bottom-right image shows the solution for the slice. The addition of partial squares produces a better match

to the ground truth compared to the original shape shown in the bottom-left.

**3. Open Source Resources, Frameworks and Libraries**

No external

resources

are used

**4. Potential Algorithm Improvements**

Improve lung bound determination:

there are often problems

measuring the lung bounding box

in slices

near the beginning or end of the lungs

Improve modeling of

◦ The current model is quite rough with only a single set of 5 pairs of polygons modeling lung shape

vs. slice.

◦ Increase detail and number of polygons used per set

◦ Create multiple sets and select the one which best matches the scan

Use training data

◦ Image matching

◦ Analyze tumor properties

alternate grid square sizes: perhaps

2x2 pixels

Improve formation of

◦ Compare shape to matching clusters from adjacent slices

◦ Split clusters suspected of actually being two smaller joined clusters

◦ Round out edges of clusters flattened from being at the edge of the lung shape

Improve creation of

polygons

◦ Reduce total number of polygons created

◦ Create a more natural shape that better emulates the ground truth

**5. Deployment Guide**



•

1. Make sure Java JDK 1.7 or later is installed

2. The scripts are for Windows, but can easily be modified to run on any machine with Java.

3. Run compile.bat

4. Run test.bat <sourceDir> <solutionPath>

<sourceDir> - directory containing the CT scans with images in PNG format

<solutionPath> - optional name of the solution file ("solution.csv" by default)

**6. Final Verification**

No training step is

required