Visión artificial, Víctor Nathanael Badillo Aldama

Segmentación del área mamaria en mamografía lateral

1.- Introducción

La mamografía es una técnica médica esencial para la detección temprana del cáncer de mama, utilizando rayos X de baja dosis para capturar imágenes detalladas del tejido mamario. Estas imágenes son fundamentales para identificar anomalías y apoyar el diagnóstico clínico. Dado su impacto en la salud, el desarrollo de métodos computacionales para analizar mamografías de manera automática es un área activa de investigación en visión por computador.

En este proyecto, nos centraremos en el análisis de mamografías laterales con dos objetivos principales:

- Segmentación del área mamaria: Delimitar la región de interés correspondiente al tejido mamario, eliminando elementos no deseados, como la zona del músculo.
- Clasificación de la densidad del tejido: Clasificar el tejido mamario segmentado en una de las tres categorías de densidad: glandular-denso, glandular-graso y graso.

Estas tareas son cruciales para facilitar el diagnóstico médico y priorizar casos según su complejidad o riesgo. El trabajo se basará en un subconjunto de imágenes del mini-MIAS database, una base de datos reconocida en este campo.

2.- Segmentación del área mamaria.

Para este primer paso del método computacional en el que queremos acabar con la parte mamaria segmentada y nada más en la imagen, debemos eliminar los artefactos existentes. Al observar las imágenes de prueba, se puede observar que unas cuantas tienen algunos números, etiquetas o símbolos. El primer paso será eliminar estos artefactos.

Para ello el primer método consiste en binarizar las imágenes. Para ello primeramente binarización mediante el método de Otsu podría parecer una buena opción ya que lo que se quiere obtener es "todo", es decir, objetos y fondo. Al tener cierto nivel de gris superior a 0 y que los bordes de las mamas y las etiquetas tienes niveles de gris parecidos, sucede la problemática de una binarización incorrecta, tratando como objeto parte del fondo. Al querer obtener un todo lo visible, se procede a binarizar mediante un umbral con nivel de gris bajo cercano a 0. Hecho esto, un operador morfológico de apertura es conveniente para separar posibles regiones unidas entre la mama y alguna etiqueta cercana. Se tiene el conocimiento de que la parte de la mama unida al músculo siempre va a ser el objeto más grande en la imagen, por lo que esto se convierte en el criterio para quedarse solo con la parte requerida.

Finalmente se redondean los bordes de este objeto y se interseca con la imagen original para obtener esta sin artefactos.

En este punto ya se pueden empezar con la parte de eliminar el músculo. Antes de aportar la solución desarrollada finalmente, otros métodos se consideraron como utilizar el algoritmo de snake para obtener la región del músculo y eliminar esta posteriormente pero al ser los bordes no muy claros no se

obtenían buenos resultados. Otra solución en este apartado a considerar fue la utilización de bordes para unir estos bordes del músculo sabiendo que siguen una curva. El problema de esta solución recaía en que no todos los músculo tenían la misma curva, algunos son más rectos y otros más curvos y alargados y además los bordes al no ser muy claros no se capturan correctamente. Aunque se podría pensar en aumentar el contraste de bordes esto sigue sin ser viable por la semejanza de brillo entre el músculo y del tejido mamario dependiendo del tipo.

Utilizando el conocimiento sobre el dominio y el área de trabajo, se puede aportar una solución que consiste en obtener la máscara del músculo para después eliminarla a la imagen para quedarse solo con la parte mamaria. Para ello se sabe que la parte muscular tiene más brillo que la mama y valores de gris más homogéneos. El crecimiento de regiones utilizando los niveles de gris es la solución propuesta. Para ello se ha de colocar una primera semilla para la región.

Esto se convierte en otro subproblema, las imágenes de prueba consiste en mamografías laterales desde los perfiles derecho e izquierdo, por lo que la parte del músculo nunca estará en la misma zona. Debido a esto el siguiente método propuesto consiste en orientar todas las mamografías en la misma dirección (perfil izquierdo), para después recortar la imagen por la parte izquierda y colocar la semilla en la esquina superior izquierda, donde estará el músculo siempre. Para conseguir esto, se obtiene el contorno de la imagen en este punto, se aplica un operador morfológico para obtener solo la línea vertical de la mamografía. Se busca el índide del primer pixel encontrado y dependiendo de su posición se gira la imagen o no y se recortan las columnas a la izquierda de la imagen centrada ya que tenemos el conocimiento de que las imágenes están colocadas así.

Ahora ya se podría empezar a emplear el crecimiento de regiones, pero tras probar resultados en este paso se observó que es necesario homogeneizar aún más la parte muscular y a poder ser mantener bordes. Primeramente se intentó conseguir utilizando un suavizado guassiano pero al tener que mantener bordes no parece una buena opción ya que se mezcla con el tejido mamario y puede hacer que el crecimiento de regiones funcione incorrectamente. Un filtro bilateral parece más adecuado ya que permite homogeneizar manteniendo los bordes.

En este punto de aplica el crecimiento de regiones, se restauran las dimensiones originales, y con la máscara obtenida se resta a la imagen. Tras esto, al tener el problema de tener imágenes con tejido mamario con un brillo semejante al músculo, se hace un paso intermedio en el que se redondea la máscara del tejido mamario y se devuelve a su orientación original.

Finalmente se ha obtenido la máscara que se quería, de la que se obtiene el contorno y se superpone sobre la imagen original para observar el resultado tras el procedimiento.Estas imágenes y las imágenes segmentadas se guardan en los subdirectorios de results para estas imágenes.

3.- Clasificación del tejido interior de la región mamaria.

El segundo de los objetivos de este método computacional consiste en clasificar las segmentaciones obtenidas en el paso anterior en uno de los 3 tipos de tejidos: graso, glandular-graso y glandular-denso.

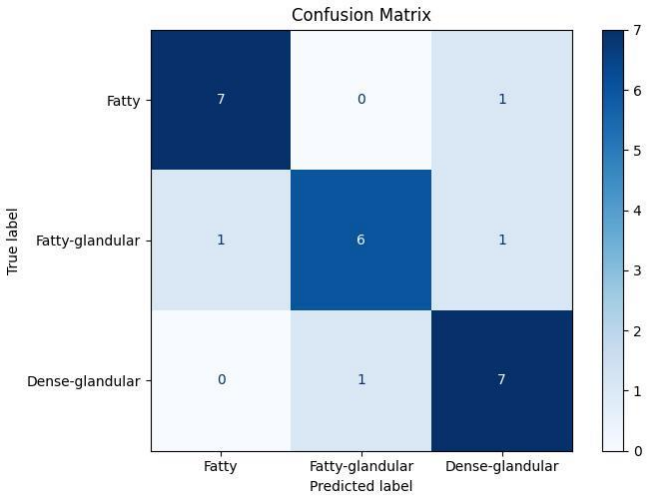
Para clasificar estas segmentaciones se observa la siguiente situación. Existen 24 imagenes de prueba, las cuales son escasas y no se tiene las máscaras de las segmentaciones ideales para el problema, por lo que teniendo en cuenta esto, se ha optado por realizar una clasificación sencilla siguiendo unas condicionales que se basan en los valores de las características extraídas de las segmentaciones.

Estas características consiste en el brillo promedio, refleja la densidad general del tejido, ya que los tejidos más densos, como el glandular-denso, tienden a mostrar valores de brillo más altos en las imágenes mamográficas. La curtosis y la asimetría (skewness) capturan la distribución de las intensidades en la región segmentada, ayudando a identificar si los tejidos son homogéneos (como en el caso del tejido graso) o si presentan colas largas y distribuciones más variables, características comunes en tejidos glandulares. Por último, la proporción de píxeles de alta intensidad mide la prevalencia de regiones brillantes dentro del tejido segmentado, lo cual es indicativo de áreas de mayor densidad.

Las reglas se definen según patrones observados en las características radiológicas: el brillo promedio distingue entre tejidos más claros y más oscuros, reflejando diferencias de densidad tisular; la curtosis y la asimetría ayudan a evaluar la uniformidad y la variabilidad de las intensidades en cada región segmentada, identificando distribuciones más homogéneas en tejido graso y más complejas en los tejidos glandulares; y la proporción de píxeles de alta intensidad permite determinar si hay concentraciones significativas de áreas brillantes, típicas del tejido glandular-denso.

4.- Evaluación de resultados

En cuanto a la clasificación, los resultados obtenidos son los siguientes:



El modelo presenta una precisión general del 83% (accuracy), lo que indica que el 83% de las predicciones fueron correctas. Analizando cada clase individualmente:

Fatty: La precisión es del 78%, lo que significa que, de todas las predicciones realizadas para esta clase, el 78% fueron correctas. Sin embargo, el modelo tiene un recall elevado de 88%, lo que sugiere que el modelo es bastante efectivo para identificar correctamente los casos de esta clase. El F1-score es de 0.82, que equilibra precisión y recall.

Fatty-glandular: Esta clase muestra un buen rendimiento, con una precisión y recall de 88%. Esto indica que el modelo tiene una alta capacidad para predecir correctamente las muestras de esta clase, con un F1-score de 0.88, lo que refleja un buen balance entre precisión y recall.

Dense-glandular: Aquí, la precisión es de 86% y el recall de 75%. Aunque la precisión es buena, el recall es relativamente más bajo, lo que sugiere que el modelo podría estar perdiendo algunos ejemplos de esta clase. El F1-score de 0.80 muestra un rendimiento intermedio.

En términos generales, el modelo tiene un buen desempeño con una precisión equilibrada, aunque hay margen para mejorar el recall en la clase Dense-glandular, lo que podría implicar ajustar algunos parámetros o técnicas de entrenamiento para capturar mejor los ejemplos de esa clase. Algunas de las imágenes presentan dificultades para clasificarse correctamente debido a la naturaleza de estas, como por ejemplo mdb312, que presenta un gran brillo a pesar de ser tipo graso o por ejemplo también mdb057 que presenta una cantidad de brillo y distrbución que se podría confundirse con tipo glandular-graso como sucede en este caso.

En cuanto a las validaciones, el uso del contraste interregión es una opción válida y suficiente para la validación de segmentaciones porque permite evaluar cómo se diferencian las distintas regiones dentro de una imagen, lo que es clave para asegurar que el proceso de segmentación ha sido exitoso. Al comparar las intensidades o características entre las regiones segmentadas, se puede comprobar que las áreas están correctamente separadas y bien definidas, lo que garantiza que las segmentaciones sean coherentes. Además, este enfoque es especialmente útil en imágenes donde las regiones tienen características claramente distintas, como en el caso de imágenes médicas, donde las estructuras que se quieren identificar suelen tener contrastes notables entre ellas. Los resultados obtenidos de las validaciones son:

Imagen_ID	Tipo	Contraste interregión
mdb009	Graso	0.798540
mdb060	Graso	0.877623
mdb271	Graso	0.815207
mdb131	Graso	0.040851
mdb028	Graso	0.655476
mdb312	Graso	0.819484
mdb095	Graso	0.720874
mdb080	Graso	0.779557
mdb033	Glandular-denso	0.931905
mdb107	Glandular-denso	0.825583
mdb320	Glandular-denso	0.606745
mdb003	Glandular-denso	0.879316
mdb004	Glandular-denso	0.803405
mdb037	Glandular-denso	0.904500

mdb057	Glandular-denso	0.745192
mdb130	Glandular-denso	0.887535
mdb118	Glandular-graso	0.744760
mdb072	Glandular-graso	0.939188
mdb021	Glandular-graso	0.689937
mdb122	Glandular-graso	0.866489
mdb023	Glandular-graso	0.890300
mdb008	Glandular-graso	0.755198
mdb045	Glandular-graso	0.754891
mdb115	Glandular-graso	0.505712

El contraste interregión muestra buenos resultados generales en la segmentación, con la mayoría de las imágenes alcanzando valores superiores a 0.7 en los tipos de tejido Graso y Glandular-graso, lo que indica una diferenciación clara entre las regiones segmentadas y el fondo. Este patrón sugiere que las segmentaciones fueron efectivas para la mayoría de las imágenes. Sin embargo, hay algunos casos con valores bajos, como mdb131 (graso) y mdb115 (glandular-graso) por las características de estas imágenes.

Los valores más altos de contraste se observan en el tipo de tejido Glandular-denso, que generalmente muestra valores superiores a 0.8, reflejando una segmentación precisa debido a las características más definidas y homogéneas de este tejido. La imagen mdb320 tiene un valor más bajo (0.6067450.606745) en comparación con las demás.

5.- Consideraciones

Para las subpartes del problema se ha dejado comentado el código que imprime las gráficas con los subpasos de cada uno de los subproblemas en caso de querer visualizar mejor el funcionamiento de cada método.

Justificación de usar un contorno no muy redondeado: si redondeo demasiado con open las partes que esten cerca del borde se pueden pegar demasiado al borde la imagen creando un borde que no corresponde.

extract_ground_truth() permite generar un archivo csv con la información útil con el valor verdadero de tipo de tejido y más proveninente del archivo /data/Info.txt

6.- Indentificación de problemas y posibles mejoras

Los valores del filtro bilateral utilizado para homogeneizar la parte del músculo podría llegar a ajustarse mejor combinado con el umbral de crecimiento, pero al ser una problemática con 4 variables se vuelve una tarea complicada.La clasificación utilizada podría utilizar otras características de las segmentaciones y utilizar diferentes condicionales de clasificación pero al ser un problema simple, se utilzó el método proporcionado.La validación de las segmentaciones podría tratar de aplicar transformaciones a las imágenes y observar si las segmentaciones son iguales a las obtenidas sin ellas pero al utilizar el crecimiento de regiones esto no parece una opción.