Base de Datos de Berkeley y Marco de Referencia 500 (BSDS500)

Visión por Computador, Universidad de los Andes, Bogotá, Colombia

March 17, 2016

María Alejandra Bravo Sarmiento
ma.bravo641@uniandes.edu.co

Lina María Mejía López
lm.mejia11@uniandes.edu.co

Abstract

La segmentación y la detección de bordes pueden ser vistas como el mismo problema. En el presente laboratorio, se evaluaron cuatro métodos de segmentación (k-means, gmm, jerárquico, watersheds) en la base de datos BSDS para evaluar su desempeño como detectores de bordes. Adicionalmente, se compararon con el estado del arte: el método ucm2. Se encontró que el método con mejor desempeño fue el jerárquico, seguido por k-means. No obstante, ucm2 supera todos nuestros métodos por un gran margen.

1. Introducción

Tradicionalmente, la visión artificial se ha centrado en tres problemas: clasificación, segmentación y reconstrucción 3D. Uno de los enfoques para llevar a cabo estos objetivos es comenzar con un procesamiento de bajo nivel, como la detección de bordes. No obstante, desde finales del siglo XX, se demostró que la segmentación y la detección de bordes pueden ser abordados y evaluados como el mismo problema. Este trabajo fue desarrollado por un grupo de investigación de la Universidad de Bearkley, que además desarrollo una base de datos con 500 imágenes para poder abordar este problema de una manera bien definida (BSDS). Esta base de datos fue uno de los pasos más relevantes en la historia de la visión artificial.
En el laboratorio pasado, se creó una función que permitía segmentar una imagen con los cuatro métodos de clustering básicos: k-means, gmm, agglomerative hierarchy, y watershed. Adicionalmente, era posible escoger el espacio de colores y el número de clustering para cada método. En el presente laboratorio, se analizaron las funciones para optimizarlas y evaluar su desempeño como detector de bordes en la base de datos BSDS

2. Descripción de los Métodos

Los métodos de clustering son métodos de aprendizaje no supervisado que no requieren marcadores pero si datos para su funcionamiento. Muchos de estos métodos han sido usados para detectar o descubrir patrones en diferentes ámbitos, además son muy útiles cuando no se sabe lo que se está buscando. En visión artificial se utilizan principalmente para agrupar pixeles en regiones, lo que se conoce como segmentación. Para realizar cualquier método de clustering es necesario definir la noción de similitud entre datos dado que esta determina los grupos.
Para este laboratorio los datos o vectores de representación se construyen a partir del espacio de color y/o las coordenadas del pixel. De esta manera, el vector puede ser rgb, lab, hsv, rgbxy, labxy, o hsvxy. Los canales de color pueden tener valores entre 0 y 1, mientras que las posiciones tienen valores enteros entre 1 y el ancho/alto de la imagen. Para que todas las dimensiones tengan la misma importancia, es necesario normalizar las posiciones. Normalmente se dividen entre el número de filas/columnas para que queden entre 0 y 1, pero es posible dividir por otros números con el objetivo de que las posiciones tengan más o menos pesos que el color en la clasificación.

2.1. k means

El método de k-means es uno de los más conocidos métodos de clustering. Recibe como parámetros vectores en $\mathbb{R}^n$ de representación para cada dato (pixel). K-means utiliza la distancia euclidiana como medida de similitud, es por esto que los grupos tienden a ser de tamaño parecido. También recibe como parámetro el número k de grupos que se quieren formar, con esto el comienza seleccionando k centroides de los grupos al azar y calcula las distancias de cada punto a estos y determina cual es el más cercano y le asigna ese grupo. Después de esto vuelve a calcular los centroides de cada grupo de tal manera que se minimice la sumatoria de las distancias entre cada dato y su centroide. Itera esto un número finito de veces, idealmente hasta que converja o se estabilice.

2.2. GMM

El método de gmm o Mezcla de Gaussianas es un método que separa los datos en grupos utilizando distribuciones gaussianas. Al igual que el anterior, recibe como parámetros vectores en $\mathbb{R}^n$ de representación para cada dato (pixel). Este método asume que los datos de cada grupo se distribuyen de forma normal por lo que construye k distribuciones gaussianas que se ajustan a ellos, con k igual al número de grupos en la muestra. Este método recibe como parámetros el número k y los datos de cada pixel, con esto el comienza seleccionando k puntos que corresponden a las medias de las distribuciones gaussianas. Luego calcula las distribuciones gaussianas de tal forma que se ajusten a los datos, esto lo hace por medio de modificar la matriz de covarianzas y un coeficiente de mezcla para cada grupo. Luego de esto estima las responsabilidades de cada pixel que corresponden a las probabilidades de que este pertenezca a cada una de las distribuciones normales, esta probabilidad devuelve una resuesta suave, es decir, devuelve la probabilidad de que cada pixel pertenezca a cada grupo. Una vez se obtienen las responsabilidades se vuelven a calcular las medias, las matrices de covarianza y los parámetros de mezcla de cada gaussiana. Itera esto un número finito de veces, idealmente hasta que converja o se estabilice.

2.3. Jerárquico

El método de clustering jerárquico varía con respecto a los métodos anteriores debido a que este realiza una segmentación aglomerativa jerárquica, es decir que contiene diferentes niveles de segmentaciones. La clusterización aglomerativa utiliza un método bottom-up que comienza con una segmentación en la que cada dato (pixel) es su propio grupo. Dada una distancia entre grupos en cada nivel, los dos grupo más cercanos se juntan para crear un nuevo grupo. Esta operación se realiza repetidas veces hasta que se obtiene el número de grupos deseados. A diferencia de k-means y gmm, la segmentación por aglomeración no es una solución aproximada y es invariante a la inicialización. Adicionalmente, puesto que los grupos son formados por la unión de grupos más pequeños, se obtiene una segmentación jerárquica. Por otro lado, dado que es posible escoger diferentes distancias dependiendo de la representación de cada dato, permite una mayor flexibilidad y no asume distribuciones específicas.

2.4. Watersheads

El método de Watersheds o línea divisora de aguas está basado en el concepto topográfico que hace referencia a los puntos en la superficie de una montaña en los que si cae una gota de agua tiene la misma probabilidad de ir en una u otra dirección. Es por esto que para este método se considera la imagen como una superficie que utiliza la intensidad de cada pixel como altura. Se calcula el gradiente de la imagen y se observan las áreas de discontinuidad, idealmente bordes, en los que se tienen valores más altos que el resto de la imagen. Si se considera que el valor es equivalente a la altura en un valle, los bordes pueden ser vistos como las cimas de cordilleras. Si un lago comienza en cada mínimo regional, las líneas de divisora de aguas son los puntos donde los lagos de unirían en el evento de una inundación, cada lago por separado corresponde entonces a un grupo de la segmentación. Para controlar el numero inicial de lago, y por lo tanto de segmentaciones, se utilizan marcadores o h mínimos regionales. Utilizando h mínimos se determina la altura de separación mínima que debe tener dos mínimos regionales para ser considerados como tal. Para que la entreda de la funcion fuera K, no H, se utilizó busqueda binaria para hayar, en cada imagen, el H necesario para obtener el K deseado.
El método de watershed desarrollado en el laboratorio anterior, solo utiliza imágenes en blanco y negro. No obstante, dicha entrada puede ser obtenida a partir de una imagen rgb, hsv o lab.

3. Elección del Método

En este laboratorio no solo se tenían cuatro métodos de donde elegir, pero cada método también tenían factores que se podían variar, como el espacio de color. Para la elección de los mejores métodos de segmentación para la base de datos de Bearkley se realizaron diferentes experimentos. Primero se escogieron los mejores parámetros para cada método por aparte, esto se realizó con 16 imágenes de la base de datos de entrenamiento, y visualmente se escogieron los parámetros que parecían más prometedores. Después de esto se corrieron los métodos con los parámetros seleccionados y se corrieron con las imágenes de test para ver el resultado en la gráfica de PR y determinar el mejor método.

3.1. Elección del espacio de representación

Para la elección del espacio de representación se realizó el experimento con 16 imágenes de entrenamiento. Se realizó con estas imágenes la segmentación con un k de 5 y para poder comparar mejor con la imagen de marcadores se tomó el gradiente de la imagen y se compararon los contornos generados por la segmentación. Esto se hizo ya que es más fácil visualizar las diferencias de los bordes que de la segmentación.

3.1.1 k means

La función k-means se podía correr en seis espacios de representación (rgb, lab, hsv, rgb+xy, lab+xy, hsv+xy). Se fijó k a 5, y en 16 imágenes, se corrió la segmentación de los diferentes espacios. En la figura 1 se encuentran algunos

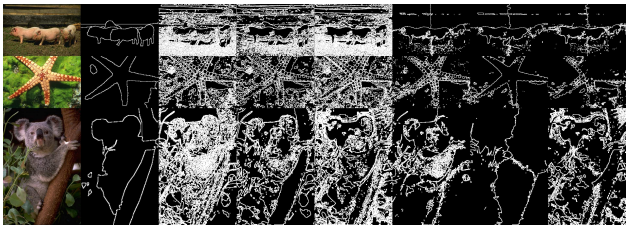

Figure 1. Experimentos con k-means. En orden: original, anoraciones, rgb, lab, hs, egb+xy, lab+xy, hsv+xy, con un K fijo de 5

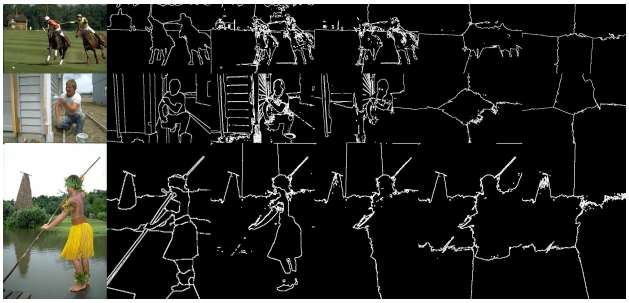

Figure 2. Experimentos con k-means, variansdo normalizacion de xy. En orden: original, anotaciones, 0.6, 0.75, 1, 1.3, 1.5. con un K fijo de 5

de los resultados. Se puede ver claramente que es necesario incluir las dimensiones de la posición para obtener resultados más limpios. Entre las tres opciones restantes, se escogió lab+xy porque La*b* es un color perceptualmente uniforme, por lo que se espera que sea mejor. Además, in los experimentos se notaba que era más selectivo (hay menos bordes), lo que debería disminuir el número de falsos positivos durante la evaluación.
A continuación se corrió el algoritmo (con k = 7 y espacio = lab+xy), con diferentes normalizaciones. Es decir, ahora las dimensiones de posición de tenían un rango [0,1], sino que podía ser un poco mayor o menor, para darle menor o mayor importancia a la posición en el momento de clusterizar (Figura 2). Se encontró que, en general, reducir la importancia de la posición parecía disminuir la calidad de los resultados. Por el otro lado, al aumentar la importancia de la posición, en algunas imágenes mejoraban los resultados, pero en la mayoría se perdían parcial o completamente los bordes deseados. Por esta razón, se optó por dejar la normalización original ([0,1]).

3.1.2 GMM

Para el método de gmm se podía escoger entre los seis espacios de representación (rgb, lab, hsv, rgb+xy, lab+xy, hsv+xy). Con el k fijo en 5 se corrió el método en las 16 imágenes de entrenamiento y se observó (figura 3) en los primeros tres canales mucho ruido dado que no se estaba tomando en cuenta las coordenadas de los pixeles sino

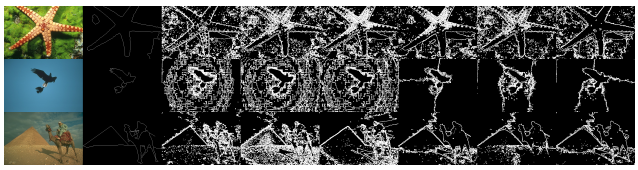

Figure 3. Experimentos con GMM. En el orden: original, anotaciones, rgb, lab, hsv, rgb+xy, lab+xy, hsv+xy

únicamente la intensidad de los mismos. Por otro lado para rgb+xy y lab+xy se puede apreciar mucho ruido en los contornos, en especial para imágenes simples como la primera o segunda en las que en el interior de la estrella de mar y los alrededores del pájaro hay mucho ruido. Finalmente también observamos que en hsv+xy se observan los bordes de sombras que también están marcadas en las anotaciones. Dado que en definitiva en la mayoría de las 16 imágenes de entrenamiento usadas se veían mejor los bordes en el espacio hsv+xy, este se eligió para realizar la prueba con las imágenes de test.

3.1.3 Jerárquico

Para el método de jerárquico fue necesario reducir el tamaño de la imagen dado que el tiempo gastado por imagen era demasiado largo. Se probaron entonces con otros tamaños en el que el lado máximo cambiaba y se conservaba la proporción de la imagen. Para un tamaño de 300 se obtiene un tiempo de 57.07 segundos, para un tamaño de 150 se obtiene un tiempo de 4.23 segundos. Para volver la imagen a su tamaño se realizaba un resize con vecino más cercano para no generar valores diferentes a los de las etiquetas de los grupos. Sin embargo debido a esto, a medida que se realizaba un escalamiento a menor tamaño para optimizar el método, los resultados obtenidos se veían cuadriculados y los bordes también, esto se puede observar en la figura 4 en la que la primera fila corresponde a una segmentación con un reescalamiento de lado máximo de 150 y la segunda fila uno de 300. Finalmente se decidió dejar un escalamiento a un tamaño de máximo 300 dado que a pesar de que los resultados no fueran igual como con la imagen de tamaño original el tiempo de procesamiento era más razonable y no lo redujimos más ya que la diferencia si era notable.

Para la elección del espacio de color se observó en las 16 imágenes de entrenamiento que en los canales en los que no se incluían las posiciones del pixel, auqeyos que no tenían los espacios xy, los resultados tenían mucho ruido. Así mismo observando las imágenes se determinó que los resultados de lab+xy eran los mejores y no producían tanto ruido como los otros, pero si incluían información importante por ejemplo en la figura 5 se observa la sombra para la imagen 2 y el borde del elefante para la imagen 3, bordes que no son tan claros en los otros espacios.

Figure 4. Experimentos de escala con Jerárquico. Las filas corresponden a rescalamiento de 150 pixeles para la primera y de 300 para la segunda. Se observa en todos los espacios de color la cuadricula generada por el rescalamiento para la primera fila. Las columnas corresponden a los espacios de color en el orden: anotaciones, rgb, lab, hsv, rgb+xy, lab+xy, hsv+xy.

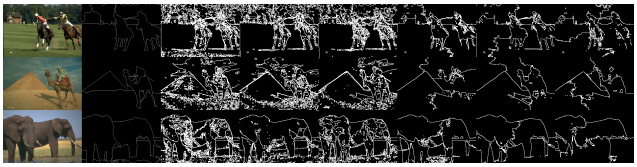

Figure 5. Experimentos con Jerárquico. En el orden: original, anotaciones, rgb, lab, hsv, rgb+xy, lab+xy, hsv+xy

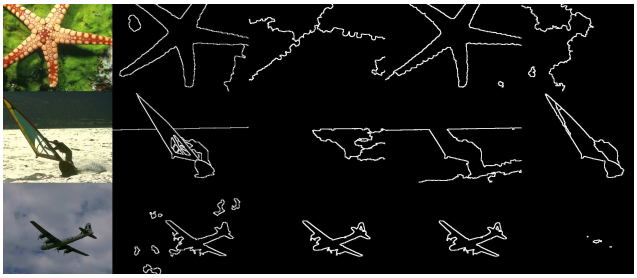

Figure 6. Experimentos con watersheds. En orden: original, anoraciones, rgb, lab, hs, egb+xy, lab+xy, hsv+xy, con un K fijo de 5

3.1.4 Watersheads

El método de watershed se probó con 16 imágenes en los tres espacios de colores (Figura 6). Al comparar visualmente los resultados, se escogió el canal La*b* porque parecía tener los resultados más consistentes. En otros los otros canales, habían imágenes con resultado muy buenos, pero también con resultados muy malos.

3.2. Elección del metodo de cluster

Para la comparación de los métodos con los mejores parámetros se realizaron segmentaciones con cada método de las imágenes de test y luego se corrió el benchmark y se graficaron los resultados. En principio solo se querían escoger dos de los métodos para ser comparados con el estado del arte pero en definitiva lo que se realizó fue una comparación con todos los métodos, esto se hizo para poder compararlos de manera objetiva y basada en una prueba más robusta. Los resultados se muestran en la figura 7 en la que se observa que los mejores métodos fueron primero Jerárquico y después k-means.

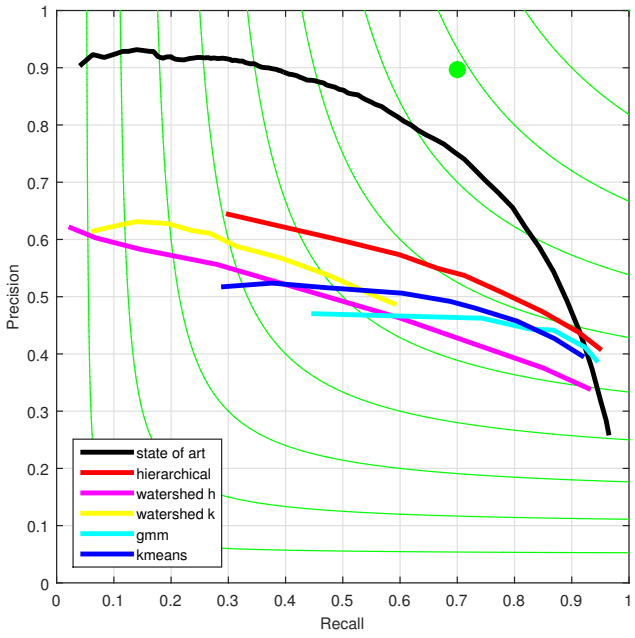

Figure 7. Curve PR

4. Metodología de evaluación

La base de datos de Berkeley consta de 500 imágenes en total de 481x321 pixeles. Considera 200 imágenes de test, 200 de entrenamiento y 100 de validación. La verdad terreno para esta base de datos cosiste en promedio de 5 imágenes de marcadores de bordes hechas por humanos. Estas imágenes son lógicas y tienen valor de 1 en los pixeles de contorno y 0 en el resto. Lo que el marco de referencia medía era el número de pixeles anotados por el algoritmo como pixeles de borde y comparaba esto con cada una de las anotaciones humanas. Esto lo hace contando para cada imagen de anotaciones los verdaderos positivos (TP), falsos positivos (FP), verdaderos negativos (TN) y falsos negativos (FN) con respecto a los contornos generados por el algoritmo. Realiza esto con cada una de las imágenes de anotaciones y va sumando en cada uno de los grupos (TP, FP, TN, FN). Finalmente calcula la precisión (P) y cobertura (R), promedia este valor para todas las imágenes de test, y grafica este punto en la gráfica. Esto lo hace para diferentes umbrales de tolerancia de bordes, o en nuestro caso para los diferentes k al momento de realizar la segmentación, con lo que se obtiene la curva de PC.[1]

5. Comparación de reultados

Los resultados obtenidos se observan en la figura 7 en la que se muestran las curvas de precisión cobertura para cada uno. En esta gráfica se puede apreciar que el método jerárquico es el mejor de los métodos implementados en el laboratorio. Esto se debe a que cruza la curva de nivel

más arriba que las demás curvas. Sin embargo la curva de este método no es muy larga y no alcanza a recorrer todo el gráfico, esto se da debido a que el máximo número de clusters determinado para este método fue de 100 por lo que no llega a una cobertura perfecta y por otro lado ni siquiera al comenzar con dos clusters se logra tener una precisión mayor. El método que le sigue es kmeans que así mismo como el anterior no recorre todo el grafico. El orden de los siguientes métodos es gmm, watershead (variando k) y watershed (variando h). Para estos se puede observar que gmm y kmeans no obtienen precisiones altas aun cuando la cobertura es baja pues la curva no logra continuar hacia la izquierda. De la misma forma watershed k no logra coberturas altas. En cambio watershed h si logra recorrer toda la gráfica pero con los menores resultados. Por otro lado al comparar los métodos del laboratorio con el estado del arte para esa base de datos se puede observar una gran diferencia y superioridad del estado del arte frente a los métodos de clustering.
En la figura 8 se observa una imagen de test segmentada por los diferentes métodos con diferentes k. Se puede observar en la última columna que la imagen de la tercera fila, la del método jerárquico, muestra una muy buena segmentación únicamente con 2 clusters, se ve que es relativamente mejor que las demás. Así mismo a medida que va aumentando el número de clusters se puede apreciar que para los métodos de watersheads el ruido aumenta en gran escala en especial para objetos pequeños como las manos de las personas.

6. Discusión

Los métodos de clustrización no se acercan al estado del arte. No solo es um2 más eficiente, pero también logra llegar a los extremos de la gráfica. Entre los probados, el método jerárquico es el mejor. Este método es bueno pues agrupa similitudes pero no hace ninguna suposición sobre la forma de los clusters, a diferencia de k.means y gmm. Por otro lado este método logra ser robusto dado que sus segmentaciones están anidadas, es decir que para cada k diferente las segmentaciones no varían con parámetros de inicialización y si existe un contorno en el nivel k para el resto de niveles con k mayor este contorno se preserva. Este método además puede variar la distancia usada para realizar el proceso de nion de grupos y esto hace que sea más flexible al momento de dener datos con diferentes distribuciones.
Por otro lado los métodos de kmean y gmm asumen que los datos están distribuidos de manera específica lo cual limita los tamaños de los grupos así como su forma. Además estos dos métodos dependen intrínsecamente de los parámetros de inicialización lo que significa que cada vez que se corra el método con puntos de inicialización diferentes la segmentación obtenida varía. Estos métodos también asumen desde un principio que se conoce el número de clusters dentro de la imagen, hecho que casi nunca es cierto y provoca limitaciones en el uso del método.
Watersheds en cambio es un método que como está basado en marcadores, mínimos regionales, hace que sea genérico e invariante frente a parámetros de inicialización, logrando así segmentaciones anidadas que dan robustez al método. Sin embargo dado que este método no usa información espacial de la misma forma que los otros y le da mayor importancia a la información de intensidad es un método que puede fallar fácilmente cuando la imagen tiene texturas.
En este laboratorio se logró demostrar a partir de experimentos que efectivamente los métodos de clustering sirven para crear segmentaciones de las imágenes. Sin embargo también se puede afirmar que no son suficientes para lograr resultados óptimos y métodos más sofisticados son necesarios para alcanzar el nivel humano. Entre todos, el método Jerárquico fue el mejor, no obstante, cabe resaltar que su eficiencia fue la más baja (incluso después de reescalar) debido a que fue el método que consumió mayor cantidad de tiempo.

7. Mejoras

Claramente, la optimización de los métodos no se realizó de manera objetiva ni con rigor. Por un lado, el K escogido (cinco) fue arbitrario, y con varios K, los resultados habrían podido ser diferentes y se habrían escogido otros parámetros 'óptimos'. La manera correcta de escoger los paramentos habría sido correr el algoritmo para todas las imágenes de entrenamiento y un rango de k, y escoger a partir de las curvas PR obtenidas.
Por otro lado, un reprocesamiento de las imágenes habría podido mejorar los resultados significativamente. Un filtro promedio, en particular, hubiera permitido la eliminación ruido, algunas texturas, y otras informaciones de frecuencia alta. Esto, a su vez, probablemente disminuiría los falsos positivos obtenidos por todos los métodos.
También es posible considerar aumentar el espacio de representación. Un espacio lab+hsv+xy, por ejemplo, puede ser prometedor. Para el método de jerarquía, si se tiene el tiempo y los recursos computacionales, sería mejor realizar la segmentación sin antes reducir el tamaño de la imagen. Esto debería resultar en bordes más precisos, menos cuadriculados, lo que resultaría en una mayor precisión y cobertura.

References

[1] Contour detection and image segmentation resources. `http://www.eecs.berkeley.edu/Research/Projects/CS/vision/grouping/resources.html`. Accessed: 2015-03-14.

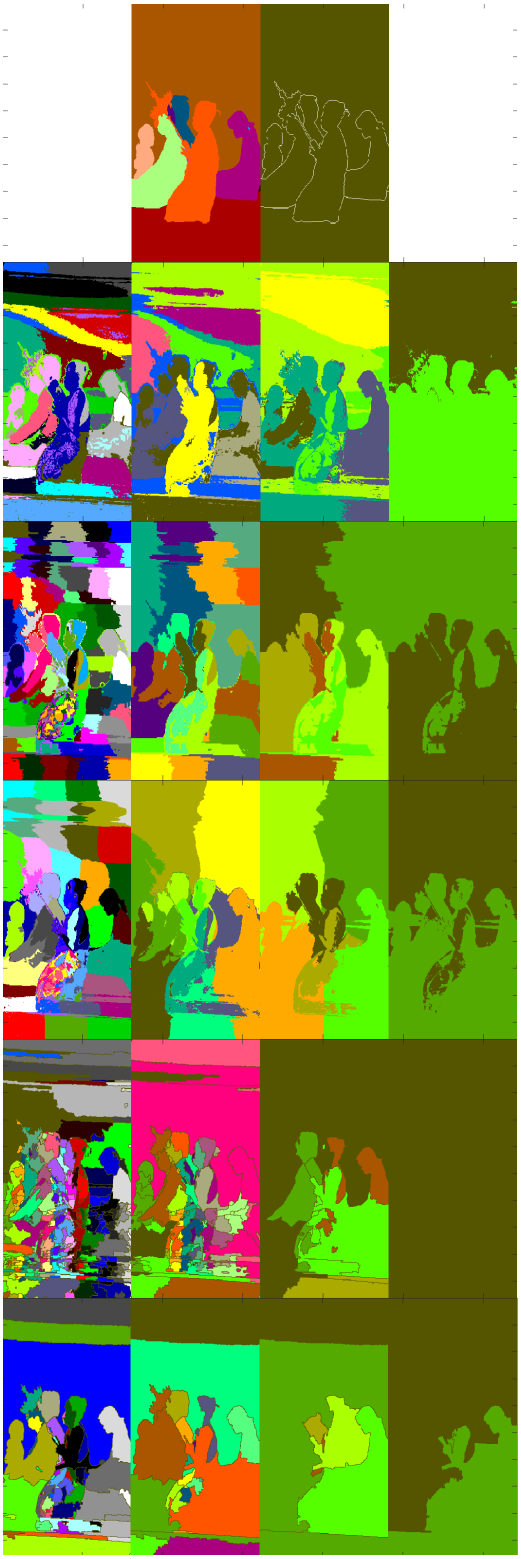

Figure 8. Imagen segmentada de test por los métodos, Fila 1.Anotaciones, 2.gmm, 3.jerárquico, 4.kmeans, 5.watershed h, 6.watershed k