Universidad Nacional de La Plata

Facultad de Informática

Segmentación no supervisada de imágenes RGB-D

Trabajo Final presentado para obtener el grado de

Especialización en Computación Gráfica, Imágenes y Visión por Computador

Alumno: Lic. Luciano Rolando Lorenti

Director: Ing. Armando De Giusti

Agosto de 2019

Índice general

Capítulo 1

Introducción

El propósito de un método de segmentación es descomponer una imagen en sus partes constitutivas. La segmentación es generalmente la primera etapa en un sistema de análisis de imágenes, y es una de las tareas más críticas debido a que su resultado afectará las etapas siguientes. El diseño de un modelo de segmentación es considerado como uno de los problemas más estudiados en análisis de imágenes y visión por computador. El objetivo central de esta tarea consiste en agrupar objetos perceptualmente similares basándose en ciertas características en una imagen.

Tradicionalmente las aplicaciones de procesamiento de imágenes, visión por computador y robótica se han centrado en las imágenes a color. Las cámaras convencionales han mejorado significativamente proveyendo información más valiosa y con mejor calidad. Esto ha permitido un aumento en el rendimiento de las aplicaciones de segmentación. Sin embargo, el uso de la información de color es limitado hasta cierto a punto debido a que las imágenes obtenidas con cámaras tradicionales no pueden registrar toda la información que la escena tridimensional provee. El carácter parcial de la información capturada dificulta la aplicación de algoritmos de segmentación y análisis sobre estas imágenes. Gran parte de las dificultades se deben a la proyección que se realiza del mundo 3D para obtener una representación bidimensional. El proceso de captura pierde la información de profundidad de los objetos. Esto crea ambigüedades entre el tamaño de los mismos y la distancia relativa con respecto al dispositivo sensor. Adicionalmente, las cámaras que no cuentan con una fuente de iluminación activa y dependen de una fuente de iluminación externa, son altamente sensibles a sus variaciones. Esto produce que el mismo

objeto pueda parecer diferente cuando las condiciones de iluminación son distintas.

Una alternativa para afrontar estas dificultades y otorgarle mayor robustez a los algoritmos de segmentación aplicados sobre imágenes obtenidas con cámaras tradicionales es incorporar la información de profundidad perdida en el proceso de captura. Los recientes desarrollos de hardware han permitido la difusión de sensores 3D de bajo costo, y por lo tanto, se ha ampliado el acceso a la información de profundidad de las escenas. La información de profundidad medida por un sensor 3D constituye una nube de puntos en el espacio. Puede ser dada por una grilla rectangular en coordenadas cartesianas $z(x, y)$, o en coordenadas polares $R(\alpha, \phi)$. La forma más simple y a menudo más conveniente de representar y almacenar las coordenadas de una superficie de una escena es utilizar un mapa de profundidad. Un mapa de profundidad M es una matriz donde la posición (x, y) de un punto corresponde a las fila x y a la columna y de la matriz, y la medición de profundidad z correspondiente a ese punto se almacena en $M(x, y)$. Este tipo de información es llamado mapa de profundidad, imagen de profundidad, imagen de rango, o imagen $2\frac{1}{2}$D. Los mapas de profundidad se asemejan a las imágenes en escala de grises generadas por una cámara 2D, con la diferencia que la información de profundidad reemplaza a la información de intensidad.

La disponibilidad de esta información abre la posibilidad de explotar de forma conjunta la información de color y profundidad para el análisis de imágenes. Las imágenes que contienen información de color de la escena, y la profundidad de los objetos se denominan imágenes RGB-D. Los algoritmos tradicionales de visión por computador que anteriormente eran desarrollados para imágenes de color o intensidad han sido mejorados para incorporar información de profundidad [1]. La incorporación de la profundidad como una característica adicional mejora la precisión en la segmentación de imágenes [2] [3] [4] [5]. Un punto clave de los métodos para segmentar imágenes utilizando datos de color y distancia, es determinar cual es la mejor forma de fusionar estas dos fuentes de información con el objetivo de extraer con mayor precisión los objetos presentes en la escena. Un gran numero de técnicas utilizan métodos de aprendizaje supervisado. En el último tiempo han surgido numerosas publicaciones utilizando redes neuronales para segmentar imágenes de color y de profundidad [6] [7] [8].

Sin embargo, en muchos casos no existen bases de datos que permitan utilizar técnicas supervisadas y en caso de existir, los costos de realizar el entrenamiento de estos métodos puede ser prohibitivo. Las técnicas no supervisadas, a diferencia de las supervisadas, no requieren

una fase de entrenamiento a partir de un conjunto de entrenamiento por lo que pueden ser utilizadas en un amplio campo de aplicaciones. En el marco de este trabajo de especialización es de particular interés el análisis de los métodos actuales de segmentación no supervisada de imágenes RGB-D.

En este contexto, mejorar las técnicas de segmentación conduce a un perfeccionamiento en las aplicaciones de visión por computador. Por lo tanto, es importante utilizar métricas que permitan analizar y comprender las ventajas, las desventajas y los compromisos en la determinación del método de segmentación. Se han propuesto diversas medidas de desempeño para analizar la calidad de una segmentación. Las medidas de evaluación supervisadas, es decir, aquellas que comparan los resultados de las segmentación con una base de datos anotada denominada *ground truth*, son el tipo de medidas más comunes [9]. Existen sin embargo, diversas métricas no supervisadas que dan cuenta acerca de la calidad de los segmentos obtenidos de acuerdo a alguna medida de cohesión entre *clusters* y de separación entre los mismos. Un segundo objetivo del presente trabajo es, por lo tanto, analizar las métricas de evaluación que permiten indicar la calidad del proceso de segmentación.

El presente trabajo esta estructurado del siguiente modo: en el segundo capítulo se expone una descripción de una selección de los principales métodos de segmentación no supervisados de imágenes RGB-D. En el tercer capítulo se detallan criterios de evaluación de segmentación supervisados y no supervisados. En el cuarto capítulo se detallan los trabajos experimentales. Por último, el capítulo final presenta las conclusiones y propone líneas de trabajo futuras.

Capítulo 2

Segmentación de imágenes RGB-D

Segmentar una imagen es considerado como uno de los problemas más estudiados en análisis de imágenes y visión por computador [10]. El objetivo central de esta tarea consiste en agrupar objetos perceptualmente similares basándose en ciertas características en una imagen. Este problema ha sido abordado desde muchas perspectivas diferentes y por lo tanto hay disponible un gran número de técnicas en la literatura. Los métodos tradicionales, por lo general, unen regiones en una imagen RGB utilizando la información de color disponible. Sin embargo, el uso de color es a veces poco confiable debido a numerosos efectos causados por la variaciones de iluminación y la presencia de sombras.

En el presente trabajo se analizan 3 de los principales métodos actuales que realizan segmentación de imágenes utilizando información de color y de profundidad abordando el problema con diferentes enfoques:

1. JCSA, Joint Color-Spatial-Axial data [3] [11]: Este método utiliza el algoritmo de *expectation-maximization* (EM) para obtener los parámetros de máxima verosimilitud de una mezcla finita de distribuciones de la familia de exponenciales.

2. GCF, Geometry and color fusion: [12]: Este método utiliza técnicas de *clustering* espectral para obtener una representación de la imagen en un espacio donde agrupar los segmentos es sencillo.

3. RGBD-Felzenszwalb [13]: Este método es una extensión del método clásico de segmentación de Felzenszwalb que utiliza conceptos de teoría de grafos para realizar la segmenta-

ción. La extensión incorpora información de distancia al algoritmo de segmentación.

2.1. JCSA

El método propuesto por Hasnat et. al [3] identifica las posibles regiones de la imagen usando un modelo estadístico de generación de imágenes. Luego une las regiones basándose en estadísticos asociados a la propiedad planar de las mismas. El modelo se basa en tres características diferentes de la imagen RGB-D: el color, la ubicación 3D de los puntos y la normal estimada de las superficies. El modelo generativo propuesto asume que las características son extraídas de forma independiente de una distribución finita de mezcla de tres distribuciones de probabilidad. Cada miembro de la mezcla de distribuciones está compuesto por una distribución gaussiana multivariada para la información de color, una distribución gaussiana multivariada para la posición 3D y una distribución Watson multivariada [14] para los vectores normales a las superficies. El uso de la distribución Watson permite superar la ambigüedad en la dirección relacionada con las superficies normales, y provee estadísticos adecuados para explicar la propiedad planar de las regiones.

Los modelos de mezcla son comúnmente utilizados en aprendizaje no supervisado [15]. En el contexto de análisis de imágenes y segmentación estos modelos han sido empleados utilizando mezclas de gaussianas para generar agrupamientos en imágenes color [16] [17] [18]. Estos agrupamientos se obtienen por medio de la utilización del algoritmo EM que realiza una estimación de máxima verosimilitud de los parámetros del modelo.

El método propuesto por los autores combina un modelo de mezclas de múltiples distribuciones pertenecientes a la familia de exponenciales. Para esto utilizan el método de *clustering* propuesto en [19] que extiende el algoritmo EM para estimar parámetros de modelos de mezclas compuestos por distribuciones de la familia de exponenciales utilizando la divergencia de Bregman.

El modelo generativo en el que se basa el método, supone que la imagen es una mezcla del producto de tres distribuciones:

$$g(\mathbf{x}_i|\phi_k) = \sum_{j=1}^{k} \pi_{j,k} f_g(\mathbf{x}_i^C|\mu_{j,k}^C, \Sigma_{j,k}^C) f_g(\mathbf{x}_i^P|\mu_{j,k}^P, \Sigma_{j,k}^P) W_d(\mathbf{x}_i^N|\mu_{j,k}^N, \kappa_{j,k}^N)$$

donde $\mathbf{x}_i = \left\{\mathbf{x}_i^C, \mathbf{x}_i^P, \mathbf{x}_i^N\right\}$ es un vector característico 9 dimensional del i-ésimo pixel con $i = 1, ..., M$ y M el número de píxeles de la imagen. Los superíndices denotan C para la información

de color, P para la información 3D de la escena y N para la información de los vectores normales. $\phi_k = \left\{\pi_{j,k}, \mu^C_{j,k}, \Sigma^C_{j,k}, \mu^P_{j,k}, \Sigma^P_{j,k}, \mu^N_{j,k}, \kappa^N_{j,k}\right\}_{j=1..k}$ denota el conjunto de parámetros de modelo donde $\pi_{j,k}$ es la probabilidad a priori del conjunto j-ésimo, $\mu_{j,k} = \{\mu^C_{j,k}, \mu^P_{j,k}, \mu^N_{j,k}\}$ son la media del componente j-ésimo, $\Sigma_{j,k} = \{\Sigma^C_{j,k}, \Sigma^P_{j,k}\}$ son las matrices positivas definidas de covarianza del componente j-ésimo, y $\kappa_{j,k}$ es el parámetro de concentración del componente j-ésimo. $f_g(.)$ y $W_d(.)$ son las funciones de densidad de la distribución Gaussiana multivariada y la distribución Watson multivariada, respectivamente. El método JCSA consta de dos sub-tareas: la primera consiste en agrupar los píxeles de la imagen utilizando la información de color, 3D y los vectores normales a las superficies; la segunda etapa consiste en la unión de regiones. La primera tarea realiza un agrupamiento con los datos y genera un conjunto de regiones. La segunda tarea realiza un refinamiento de la segmentación con el objetivo de unir las regiones que son susceptibles a ser sobre-segmentadas.

2.1.1. Método de segmentación

Una función de densidad probabilidad multivariada $f(x|\eta)$ pertenece a la familia de exponenciales si tiene la siguiente forma:

$$f(x|\eta) = \exp(-D_G(t(x), \eta)) \exp(k(x))$$

siendo

$$D_G(\eta_1, \eta_2) = G(\eta_1) - G(\eta_2) - < \eta_1 - \eta_2, \nabla G(\eta_2) >$$

$G(.)$ es el dual de Legendre de la función de normalización estrictamente convexa asociada con una distribución de probabilidad, ∇G es el gradiente de G, $t(x)$ denota el estadístico suficiente y $k(x)$ la medida portadora. La esperanza del estadístico suficiente $t(x)$ con respecto a la función de densidad es el parámetro de esperanza η. D_G es la Divergencia de Bregman (DB) y es calculada a partir de los parámetros de esperanza [3]. La DB puede ser usada para medir la desemejanza entre dos distribuciones de la misma familia de exponenciales definidas por los parámetros de esperanza η_1 y η_2. En el artículo donde es presentado el método, se encuentra una derivación de cada una de las partes de la ecuación para la distribución Gaussiana y Watson.

El modelo generativo propuesto combina distintas distribuciones de la familia de exponenciales basándose en la suposición de independencia. Por lo tanto, la DB del modelo combinado puede ser definido como una combinación lineal de la DB de cada distribución individual:

$$D_G^{comb}(\eta_i, \eta_j) = D_{G,g}^C(\eta_i^C, \eta_j^C) + D_{G,g}^P(\eta_i^P, \eta_j^P) + D_{G,dir}^N(\eta_i^N, \eta_j^N)$$

donde $D_{G,g}(.,.)$ denota la DB utilizando la distribución gaussiana multivariada y $D_{G,dir}(.,.)$ denota la DB usando la distribución direccional Watson. Entonces, es posible definir, con el parámetro de esperanza $\eta = \{\eta^C, \eta^P, \eta^N\}$:

$$G^{comb}(\eta) = G_g(\eta^C) + G_g(\eta^P) + G_{dir}(\eta^N)$$

El método de agrupamiento de Bregman, utiliza la divergencia de Bregman en el algoritmo EM para calcular la estimación de máxima verosimilitud del modelo de mezcla y calcula una probabilidad de pertenencia de las observaciones a cada uno de los grupos.

Con el objetivo de agrupar los datos utilizando el modelo combinado, es necesario estimar los parámetros del modelo ϕ_k para $g(\mathbf{X}|\phi_k)$ tal que:

$$\hat{\phi_k} = \arg\max_{\phi_k} g(\mathbf{X}|\phi_k) \text{ con } g(\mathbf{X}|\phi_k) = \prod_{i=1}^{M} g(\mathbf{x}_i|\phi_k)$$

Donde, $\mathbf{X} = \{x_i\}, i = 1, ..., M$ es el conjunto de observaciones. Sea $\gamma_i = j$ la etiqueta de la observación x_i con $j = 1, ..., k$. El algoritmo de *clustering* de Bregman consiste en un paso de esperanza y un paso de maximización. En el paso de esperanza, la probabilidad posterior es calculada de la siguiente manera:

$$p(\gamma_i = j|\mathbf{x}_i) = \frac{\pi_{j,k} \exp(G^{comb}(\eta_{j,k}) + < t(\mathbf{x}_i) - \eta_{j,k}, \nabla G^{comb}(\eta_{j,k}) >}{\sum_{l=1}^{k} \pi_{l,k} \exp(G^{comb}(\eta_{l,k}) + < t(\mathbf{x}_i) - \eta_{l,k}, \nabla G^{comb}(\eta_{l,k}) >}$$

Donde $\eta_{j,k}$ y $\eta_{l,k}$ denota los parámetros de esperanza para cualquier cluster j y l, dado que el número total de componentes es k. El paso de maximización, actualiza las proporciones de mezcla y los parámetros de esperanza para cada grupo.

$$\pi_{j,k} = \frac{1}{M} \sum_{i=1}^{M} p(\gamma_i = j|\mathbf{x}_i) \quad \text{y} \quad \eta_{j,k} = \frac{\sum_{i=1}^{M} p(\gamma_i = j|\mathbf{x}_i)\mathbf{x}_i}{\sum_{i=1}^{M} p(\gamma_i = j|\mathbf{x}_i)}$$

Luego de la inicialización, iterativamente se aplica un paso de esperanza y un paso de maximización hasta que se satisfacen los criterios de convergencia. Los criterios se basan en un máximo número de iteraciones y una diferencia de umbral entre los valores del logaritmo negativo de la verosimilitud entre dos pasos consecutivos. El procedimiento descripto conduce a la obtención de una probabilidad de pertenencia de las observaciones a cada grupo y a la obtención de

los parámetros para cada componente de modelo propuesto. Finalmente, para cada muestra se obtiene la etiqueta γ_i usando el modelo combinado:

$$\gamma_i = \arg \min_{j=1,...,k} D_G^{comb}(t(\mathbf{x}_i), \eta_{j,k})$$

Luego de este procedimiento se obtiene una segmentación de una imagen RGB-D utilizando el color, la información espacial y direccional. Este algoritmo asume que el número máximo de clusters es conocido. Las regiones obtenidas en esta etapa tienden a ser sobre-segmentadas, por lo tanto es necesario realizar un proceso de unión de regiones para obtener la segmentación final.

2.1.2. Unión de regiones

El primer paso para realizar el procedimiento de unión de regiones consiste en la construcción de un grafo de regiones adyacentes (RAG) [20]. En este tipo de grafos, cada región es considerada como un vértice del grafo en el que cada uno está conectado a sus regiones adyacentes en la imagen. El grafo es recorrido siguiendo un orden determinado, evaluando un predicado entre nodos adyacentes para determinar si las regiones deben unirse o no.

Sea $R = r_i$, $i = 1, ..., M$ el conjunto de regiones que se obtienen en la etapa anterior del método y $G = (V, E)$ un grafo no dirigido que representa el RAG. Cada nodo $v_i \in V$ se caracteriza por los parámetros μ y κ de la distribución Watson asociado a la región r_i. Cada arista e_{ij} consiste de dos pesos: w_d, basado en la desemejanza de los parámetros de esperanza y w_b, basado en la semejanza del contorno entre los nodos v_i y v_j. El peso w_d se define como:

$$w_d(v_i, v_j) = \min(D_{G,dir}^N(\eta_i^N, \eta_j^N), D_{G,dir}^N(\eta_j^N, \eta_i^N))$$

El peso w_b se define como:

$$w_b(v_i, v_j) = \frac{1}{r_i \cup r_j} \sum_{b \in r_i \cup r_j} I_g^{rgbd}(b)$$

donde $r_i \cup r_j$ es el conjunto de los píxeles de contorno que se encuentran entre las regiones r_i y r_j, $|.|$ denota la cardinalidad e I_g^{rgbd} es la magnitud normalizada del gradiente de la imagen (MoG) [10] [3] calculado a partir de la imagen RGB-D .

La estrategia de unión de regiones es un procedimiento iterativo que procede empleando un predicado de unión de regiones entre los nodos adyacentes en un cierto orden. Una vez que dos nodos son unidos, la información relacionada a los nodos es combinada y las aristas son

actualizadas. Este procedimiento continua hasta que no hay candidatos para ser unidos. El predicado de unión de regiones es P_{ij} y se define como:

$$P_{ij} = \begin{cases} \text{unir} & \begin{cases} \text{si } \kappa_j > \kappa_p \text{ y} \\ w_d(v_i, v_j) < th_d \text{ y } w_b(v_i, v_j) < th_b \text{ y} \\ \textit{planar outlier ratio} > th_r \end{cases} \\ \text{no unir} & \text{caso contrario} \end{cases}$$

donde κ_j es la concentración de la región r_j. k_p es un parámetro que define el umbral para la propiedad planar de la región, th_d y th_b son dos umbrales asociados al peso de la distancia w_d y el peso del contorno w_b. th_r es el umbral asociado con el *planar outlier ratio* [21]. La primera condición del predicado de unión evalúa si la región j es lo suficientemente plana para ser unida, la segunda condición evalúa si las regiones a unir son lo suficientemente similares y la tercera condición evalúa si después de unir ambas regiones, la nueva región formada será lo suficientemente plana.

2.2. GCF

La técnica propuesta por Zanuttigh et. al. [12] utiliza la información de distancia y color en un algoritmo de *clustering* espectral [22] para segmentar la escena capturada. Dada una escena descripta por un conjunto de M puntos $p_i = 1, ..., M$ con información de geometría y color, el algoritmo utiliza un vector 6-dimensional para almacenar las componentes RGB y la información de geometría con la componentes x, y, z. Las fuentes de información son unificadas en una primera etapa para poder calcular una desemejanza entre los vectores que sea significativa. La información de color es representada en el espacio de color CIElab para darle una significancia perceptual a la distancia euclídea entre colores. Luego, cada punto es normalizado por la suma de las desviaciones estándar:

$$\begin{bmatrix} \hat{L}(p_i) & \hat{a}(p_i) & \hat{b}(p_i) \end{bmatrix} = \frac{3}{\sigma_L + \sigma_a + \sigma_b} \begin{bmatrix} L(p_i) & a(p_i) & b(p_i) \end{bmatrix}$$

Con el objetivo de independizarse de la escala relativa de la nube de puntos de la geometría 3D, todos los componentes de $p_i^g, i = 1, ..., N$ son normalizados respecto al promedio de las desviaciones estándar de cada coordenada. Si σ_n, σ_y y σ_z las desviaciones estándar de las

componentes x, y, z y los p_i. El promedio de la desviación estándar se define como $\sigma_g = \frac{\sigma_x + \sigma_y + \sigma_x}{3}$. Por lo tanto el vector con información geométrica queda definido como:

$$\begin{bmatrix} \hat{x}(p_i) & \hat{y}(p_i) & \hat{z}(p_i) \end{bmatrix} = \frac{1}{\sigma_g} \begin{bmatrix} x(p_i) & y(p_i) & z(p_i) \end{bmatrix}$$

La relevancia de la información geométrica y la de color en el proceso de segmentación es balanceada mediante un parámetro λ. Valores grandes de λ incrementan la relevancia de la geometría, mientras que valores bajos de λ incrementa la relevancia de la información de color. De este modo el vector que representa la escena es un vector 6-dimensional de la forma

$$\begin{bmatrix} \hat{L}(p_i) & \hat{a}(p_i) & \hat{b}(p_i) & \lambda\hat{x}(p_i) & \lambda\hat{y}(p_i) & \lambda\hat{z}(p_i) \end{bmatrix}$$

Estos vectores 6-dimensionales representan la información de geometría y color de la escena de una forma intuitiva y consistente. Esta representación puede ser utilizada por una de las varias técnicas de agrupamiento posibles. Los métodos que utilizan la semejanza de a pares computadas entre todos los posibles pares de puntos no asumen una distribución de los puntos por lo tanto sus resultados en la práctica tienden a ser mas robustos y precisos. La principal desventaja de estos métodos es que necesitan calcular las semejanzas entre todos los pares posibles de puntos. El algoritmo de cortes normalizados [22] es un ejemplo efectivo de esta familia. En este método se construye un grafo a partir de todos los puntos de la imagen y se computan las semejanzas entre ellos, y luego se particiona mediante las herramientas propuestas por las técnicas de *clustering* espectral.

El criterio de cortes normalizados es utilizado para particionar un grafo con el objetivo de obtener agrupamientos que tengan en cuenta la semejanza de los puntos pertenecientes a cada uno y la desemejanza entre los puntos de segmentos diferentes. El problema de minimización del criterio de cortes normalizados puede ser formulado como un problema de autovalores generalizado. Una variedad de métodos se han propuesto para obtener una aproximación eficiente de los segmentos del grafo para sobrellevar la carga computacional y de memoria. Una solución posible es imponer que no todos los puntos estén conectados haciendo que cada nodo del grafo se conecte con su vecindad más cercana. Esta restricción conduce a sobre-segmentar las imágenes, e implícitamente el método impone algunas suposiciones acerca de la distribución de los puntos.

En el método propuesto por Fowlkes et.al. [23], un conjunto de puntos de la imagen es muestreado aleatoriamente, este subconjunto es segmentado de acuerdo al criterio de cortes

normalizados y luego los resultados son propagados al conjunto total de puntos utilizando una técnica para la solución numérica del problema de autofunciones, denominada método de Nyström. Los resultados de este método son comparables con los obtenidos con las técnicas de *clustering* espectral tradicionales, pero con costos computaciones y de memoria comparables con los métodos de agrupamiento central como k-medias. Una descripción de la técnica propuesta por Fowlkes se encuentra en el apéndice A.1. El método propuesto por Zanuttigh et. al. utilizan el método propuesto por Fowlkes et. al. para segmentar la imagen. Con el objetivo de determinar el parámetro λ óptimo, definen una medida de evaluación no supervisada $GCFQ$, descripta en el capítulo 3. La segmentación es llevada a cabo variando este parámetro y tomando el valor de aquel parámetro que maximiza la función objetivo $GCFQ$. De esta forma los parámetros del algoritmo son la función de pesos que se utiliza para medir la semejanza entre dos píxeles, la cantidad de autovectores a obtener a partir de la matriz laplaciana del grafo, y los posibles valores del parámetro λ a evaluar.

2.3. RGB-D Felzenszwalb

El algoritmo propuesto por [13] formula el problema de segmentación de imágenes como un problema de partición de grafos. Sea $G = (V, E)$ un grafo en el que los vértices $v_i \in V$ son el conjunto de píxeles a ser segmentados, y las aristas $(v_i, v_j) \in E$ los pares de vértices vecinos. Cada arista $(v_i, v_j) \in E$ tiene un peso correspondiente no negativo $w((v_i, v_j))$, que da cuenta de la desemejanza entre los elementos vecinos v_i y v_j. En el caso de la segmentación de imágenes el peso de una arista es medido como la desemejanza entre los dos píxeles conectados por esa arista, por ejemplo como la diferencia en intensidad, color, movimiento, etc.

Una segmentación S es una partición del conjunto V en componentes tales que cada componente, $C \in S$ corresponde a un componente conectado del grafo $G = (V, E)$. Hay muchas formas distintas de medir la calidad de una segmentación, pero en general, se requiere que los elementos en un componente sean similares entre sí, y los elementos en componentes distintos sean desemejantes. Esto significa que las aristas entre dos vértices pertenecientes al mismo componente deberían tener pesos relativamente bajos y aristas que conectan píxeles en diferentes conjuntos deberían tener pesos relativamente grandes.

Los autores proponen un algoritmo para producir una segmentación relacionada con el al-

goritmo de Kruskal para el cálculo de un árbol recubridor mínimo. Las aristas son consideradas en orden creciente de acuerdo al peso, su vértices son agregados a una región si esta adición no produce un ciclo en el grafo y si los píxeles son similares a los píxeles existentes en la región. El algoritmo retorna múltiples árboles recubridores mínimos disjuntos, en el que cada árbol corresponde a un segmento.

Un enfoque natural para segmentar imágenes RGB-D es extender el algoritmo propuesto por [13] para utilizar como peso de las aristas una combinación pesada de la diferencia en distancia y color:

$$(1 - \alpha)dC(p1, p2) + \alpha dD(p1, p2)$$

donde p_1 y p_2 son dos píxeles, d_C es la diferencia de color y d_D es la diferencia en profundidad, y α es un escalar que pesa la importancia relativa entre las dos fuentes de información.

Capítulo 3

Métricas de evaluación

Mejorar las técnicas de segmentación conduce a un perfeccionamiento en las aplicaciones de visión por computador. En esta tarea la utilización de métricas que permitan a los investigadores analizar y comprender las virtudes y los defectos de los métodos de segmentación resulta de enorme importancia. Se han propuesto diversas medidas de desempeño para analizar la calidad de una segmentación. Las medidas de evaluación supervisadas, es decir, aquellas que comparan los resultados de las segmentación con una base de datos anotada denominada *ground truth*, son el tipo de medidas más comunes [9]. Existen, sin embargo, diversas métricas no supervisadas que dan cuenta acerca de la calidad de los segmentos obtenidos de acuerdo a alguna medida de cohesión entre los *clusters* y de separación entre los mismos.

3.1. Medidas de evaluación supervisadas

La métricas de evaluación supervisadas de segmentos de imágenes determinan la calidad de las particiones por medio de la medición del acuerdo entre los píxeles segmentados y el *ground-truth*.

Las medidas pueden ser clasificadas de acuerdo a la interpretación que se haga de una partición de una imagen. Una posible interpretación consiste en plantear el problema de segmentación como un problema de clasificación de dos clases: una clase esta conformada por los píxeles que pertenecen a los segmentos y la otra clase los píxeles que pertenecen al fondo. Otra posible interpretación considera a cada segmento obtenido como un agrupamiento del conjunto de píxeles, de forma que, cualquier medida utilizada para evaluar algoritmos de *clustering* puede ser

aplicada en este contexto. También, podemos convertir el problema a un problema de *clustering* de dos clases sobre el conjunto de todos los pares de píxeles: aquellos pares pertenecientes a la misma región y aquellos asignados a regiones diferentes. Finalmente, es posible interpretar la segmentación como un problema de detección que intenta distinguir cuáles píxeles conforman la frontera de los objetos y cuáles no.

3.1.1. Generalización de métricas de clasificación binaria

Si se considera la segmentación de una imagen como un problema que apunta a dividir los píxeles de la imagen entre objetos y no-objetos, es posible referir a la clase de los objetos como píxeles positivos y la clase de los no-objetos como píxeles negativos. De acuerdo a esta notación, dado un método de segmentación automático m, la detección de un objeto puede ser escrita como la división de los píxeles de la imagen en dos clases disjuntas $I = P_m \cup N_m$, donde P_m y N_m son los píxeles positivos y negativos. El subíndice m indica el método de segmentación utilizado. De forma equivalente, el objeto etiquetado en el *ground truth* puede ser denotado como $I = P_{gt} \cup N_{gt}$.

Una manera de definir medidas para comparar dos segmentaciones consiste en generalizar las medidas estándares de clasificadores: jaccard, presición-exhaustividad, medida-f, por medio de una comparación región a región. Una descripción de las medidas clásicas de evaluación se encuentran en el apéndice A.2. La principal dificultad consiste en hacer corresponder una región entre las distintas particiones.

El modo más directo de generalizar una medida a un par de particiones S y S' es comparar cada región R en S con su mejor correspondencia local R' en S', y luego obtener la medida final promediando los resultados. Sea M una medida y d la medida global:

$$d(S, S') = \frac{1}{n} \sum_{R \in S} |R| \max_{R' \in S'} M(R, R')$$

donde n es el número de píxeles de la imagen. En la literatura M puede ser definido como, la precisión, la medida F, o el índice de Jaccard.

Precisión

La distancia direccional de Hamming de una partición S hacia otra S' denotada $D_H(S \to S')$ [24] se define de la siguiente manera:

$$D_H(S \to S') = \sum_{R'_i \in S'} \sum_{R_j \neq R_i} |R'_i \cap R_j| = n - \sum_{R' \in S'} \max_{R \in S} |R' \cap R|$$

En [25] ha sido analizada la distancia de Hamming y ha sido propuesta la distancia de partición asimétrica d_{asym} tal que $d_{asym}(S', S) = D_H(S \to S')$. Los autores explican que es el mínimo número de píxeles que deben ser removidos de la partición S para que cada agrupamiento este completamente contenido en un segmento de S'.

Si la d_{asym} es convertida en una medida de semejanza y se normaliza para que sus valores se encuentren en el rango $[0, 1]$.

$$s_{asym}(S, S') = 1 - \frac{1}{n} d_{asym}(S, S') = \frac{1}{n} \sum_{R \in S} |R| \max_{R' \in S'} \frac{|R \cap R'|}{|R|}$$

Si S' es visto como el *ground truth*, la siguiente expresión revela que la medida consiste en una generalización de la media de precisión:

$$s_{asym}(S, S') = \frac{1}{n} \sum_{R \in S} |R| \max_{R' \in S'} \text{Prec}(R, R')$$

Medida F

En el contexto del *clustering* de documentos, en [26] presentaron una medida para generalizar la medida F a un agrupamiento. Es posible aplicar la misma medida al problema de segmentación de imágenes definiendo la siguiente función:

$$L(S, S') = \frac{1}{n} \sum_{R \in S} |R| \max_{R' \in S'} F(R, R')$$

Cobertura de segmentación (SegCov)

La medida presentada en [27] consiste en la cobertura de una partición S por una partición S' que corresponde a la generalización del índice de Jaccard. Formalmente la cobertura de segmentación de una partición S por una partición S' puede ser reescrita como:

$$\text{SegCov}(S, S') = \frac{1}{n} \sum_{R \in S} |R| \max_{R' \in S'} J(R, R')$$

donde J es el índice de Jaccard.

Variación de la información (VoI)

El trabajo propuesto en [28] introdujo un nuevo punto de vista para determinar la calidad de un agrupamiento basándose en resultados de la teoría de la información que son aplicables a particiones. La autora define una variable aleatoria discreta que toma N valores que consisten en elegir de forma aleatoria cualquier pixel de la partición $S = R_1, ..., R_n$ y observar la región a la que pertenece. Asumiendo que cada pixel tiene la misma probabilidad de ser elegido, la probabilidad de cada posible valor de la variable aleatoria es: $p(i) = \frac{|R_i|}{n}$, donde n es el número de píxeles de la imagen. La entropía de esta variable aleatoria puede ser entendida como la incertidumbre de elegir un pixel de la partición, por lo tanto la entropía asociada con la partición S se define como:

$$H(S) = -\sum_{i=1}^{N} P(i)log(P(i)) = -\frac{1}{n}\sum_{R\in S} |R| \log\left(\frac{|R|}{n}\right)$$

De forma similar, cuando tenemos dos particiones S y $S' = R'_1, R'_2, ..., R'_N$ la probabilidad conjunta de un pixel perteneciente al *cluster* i en S y al *cluster* i' en S' se define como $P(i, i') = \frac{|R_i \cap R'_i|}{n}$ y entonces la información mutua entre las partición S y S' se define como:

$$I(S, S') = \frac{1}{n}\sum_{R\in S}\sum_{R'\in S'} |R \cap R'| \log\left(\frac{n|R \cap R'|}{|R||R'|}\right)$$

Este valor puede ser interpretado como la incertidumbre acerca de la región a la que pertenece un pixel en S sabiendo a que región pertenece en S'. La variación de la información [28] [29] se define como

$$VoI(S, S') = H(S) + H(S') - 2I(S, S')$$

Se ha demostrado que $VoI(S, S')$ está acotada por $2\log(\text{máx}(|S|, |S'|))$ de forma que la variación de la información normalizada se define como:

$$NVoI(S, S') = \frac{VI(S, S')}{2\log(\text{máx}(|S|, |S'|))}$$

3.1.2. Segmentación sobre el espacio de pares de píxeles

Una partición de una imagen puede ser vista como una clasificación de todos los pares de píxeles en dos clases: aquellos pares pertenecientes a la misma región y aquellos que provienen de regiones diferentes. Sea $P = \{(p_i, p_j) \in I \times I | i < j\}, |P| = (n^2)$. Dada la partición S y S', dividimos P en cuatro conjuntos diferentes dependiendo la pertenencia del par de píxeles:

- P_{11}: Ambos pertenecen a la misma región en S y en S'
- P_{10}: El par de píxeles pertenece a la misma región en S pero a distintas en S'
- P_{01}: El par de píxeles pertenece a la misma región en S' pero a distintas en S.
- P_{00}: Ambos pertenecen a diferentes regiones en S y S'

Medida F para regiones (F_r)

Si se define el objetivo de la segmentación de imágenes como detectar aquellos pares de píxeles en la misma región y aplicando las medidas de precisión y exhaustividad sobre este punto de vista, la medida presentada en [30] como precisión y exhaustividad se define de la siguiente manera:

$$Prec_r = \frac{|P_{11}|}{|P_{11}| + |P_{10}|} \quad Rec_r = \frac{|P_{11}|}{|P_{11}| + |P_{01}|}$$

Es posible utilizar entonces la medida F para regiones:

$$F_r = \frac{2Prec_r Rec_r}{Prec_r + Rec_r}$$

Probability Rand Index (PRI)

El Probability Rand Index, definido originalmente en [31] es una medida de evaluación que surge naturalmente en este contexto:

$$RI(S, S') = \frac{|P_{00}| + |P_{11}|}{|P|}$$

Medida F para objetos y partes (F_{op})

Motivado por el hecho que la segmentación de imágenes está siendo utilizado cada vez más como etapa preliminar para la detección de objetos, en [9] se propone una medida para determinar la calidad de la segmentación teniendo en cuenta esta perspectiva. Las regiones en una partición son consideradas como posibles candidatos de objetos y son clasificados como correctos o incorrectos dependiendo del grado de solapamiento con las regiones del *ground truth*. Similarmente, las regiones en una sobre-segmentación pueden ser interpretadas como partes de objetos si al unirlas pueden formar un objeto del *ground truth*.

La precisión y la exhaustividad son calculados como una fracción pesada de candidatos respecto al número total de regiones, es decir, candidatos a partes son contados parcialmente.

Formalmente sea $S = R_1, ..., R_n$ una partición de la imagen y G_k un conjunto de particiones del *ground-truth* de la misma imagen. Para cada par de regiones $R_i \in S$, $R'_j \in G$ se calculan las superposiciones relativas como:

$$O_S^{ij} = \frac{|R_i \cap R'_j|}{|R_i|} \qquad O_G^{ij} = \frac{|R_i \cap R'_j|}{|R'_j|}$$

A partir de dos umbrales, γ_o, un umbral para determinar si le región se trata de un objeto, y $\gamma_p < \gamma_o$ un umbral para determinar si la región se trata de una parte de un objeto, clasifican las regiones en ambas particiones como se describe en el algoritmo 1.

forall $R_i \in S, R'_j \in G$ **do**
  if $O_S^{ij} > \gamma_o$ **and** $O_G^{ij} > \gamma_o$ **then**
    $R_i, R'_j \leftarrow$ Candidatos a objetos
  else if $O_S^{ij} > \gamma_p$ **and** $O_G^{ij} > \gamma_o$ **then**
    if R_i *no esta clasificado* **then**
      $R_i \leftarrow$ Candidatos a parte
    end
  else if $O_S^{ij} > \gamma_o$ **and** $O_G^{ij} > \gamma_p$ **then**
    $R'_j \leftarrow$ candidato a parte
end

Algoritmo 1: Clasificación de candidatos a regiones

Sea oc y oc' el número de candidatos a objetos en S y G, respectivamente, y pc y pc' el número de candidatos a partes. Se define la cantidad de fragmentación $fr(R_i)$ de una región $R_i \in S$ como la adición de las superposiciones relativas de los candidatos a partes coincidentes a R_i:

$$fr(R_i) = \sum_j \left\{ O_G^{ij} \quad s.a. \quad O_S^{ij} > \gamma_o \right\}$$

$fr'(R')$ se define de forma equivalente para G. La fragmentación global fr y fr' se calcula sumando la cantidad de fragmentación entre todos los candidatos a fragmentos de S y G, respectivamente.

Podemos definir la precisión-exhaustividad para objetos y partes de la siguiente manera:

$$P_{op} = \frac{oc + fr + \beta pc}{|S|} \quad R_{op} = \frac{oc' + fr' + \beta pc'}{|G|}$$

Intuitivamente, en un resultado sobre-segmentado, la exhaustividad será alta pero la precisión será baja. De forma opuesta, un resultado completamente sobre segmentado tendrá alta precisión, pero baja exhaustividad. Como medida de resumen, proponen usar la medida F entre P_{op} y R_{op}.

3.1.3. Medida para detección de bordes

Medida F para contornos (F_b)

El algoritmo propuesto en [32] calcula la correspondencia entre los contornos obtenidos por medio de un algoritmo de segmentación y los contornos etiquetados por un humano. Convierten el problema de correspondencia en un problema de asignación bipartita de costo mínimo, donde el peso entre los píxeles obtenidos por el algoritmo de segmentación y el pixel obtenido por medio del humano es proporcional a la distancia relativa en el plano de la imagen. El tiempo de complejidad de buscar la mejor correspondencia densa entre los píxeles de ambas imágenes es prohibitivo. Por lo tanto, el algoritmo propuesto realiza una formulación dispersa del problema de asignación. Para esto, incluyen en el grafo solo las aristas con peso $w \leq d_{\max}$. Luego de esta etapa, cualquier nodo aislado puede ser removido del problema de asignación y puede ser contado como una falso positivo. El problema de asignación de mínimo costo requiere especificar el grado de asignación para restringir la búsqueda a soluciones no triviales. Como se desconoce el grado a priori, se debe buscar una coincidencia perfecta, es decir, una coincidencia que involucre a todos los nodos. Sin embargo, el proceso de esparficiación puede remover aristas requeridas para obtener una correspondencia completa. El problema se soluciona agregando aristas espurias a ambos lados del grafo bipartito Todas las aristas incidentes en un nodo espurio tienen un peso mayor que cualquier arista real en el grafo, de esta forma se asegura que solo serán usadas para extender la correspodencia parcial de mínimo costo a una correspondencia completa.

La correspondencia perfecta de costo mínimo en este grafo provee la mejor correspondencia de píxeles entre los mapas de bordes del humano y la maquina, con una tolerancia máxima de d_{dmax}. El cálculo de la correspodencia provee una forma de obtener la precisión P_b y la exhaustividad R_b para una segmentación humana permitiendo acotar el error de localización. Como medida de resumen se utiliza la medida F entre P_b y R_b.

Boundary Displacement Error (BDE)

La métrica denominada Boundary Displacement Error [33] tiene como objetivo evaluar la calidad de la segmentación en términos de la precisión de los contornos de las regiones extraídas.

Sea B el conjunto de los puntos de contorno obtenidos a partir de la segmentación y B_{GT} el conjunto de los puntos de contorno del *ground truth*. Se define la distancia desde un punto arbitrario x en el conjunto $B1$ a $B2$ como la distancia mínima de x a todos los puntos de $B2$, $d(x, B2) = \text{mín}(dE(x, y)) \quad \forall y \in B2$, donde dE denota la distancia euclidea entre los puntos x e y. Es posible medir la discrepancia entre B y B_{GT}, denominada $D_{B,B_{GT}}$ de la siguiente manera:

$$D_{B,B_{GT}} = \frac{\sum\limits_{\forall x \in B} d(x, B_{GT})}{|B|}$$

done $|B|$ denota la cantidad de puntos de contornos en B. Finalmente se define la métrica Boundary Displacement Error de la siguiente manera:

$$\text{BDE}(B, B_{GT}) = \frac{D_{B,B_{GT}} + D_{B_{GT},B}}{2}$$

3.2. Medidas de evaluación no supervisadas

Mientras que los métodos de evaluación supervisados evalúan las imágenes segmentadas con respecto a una imagen de referencia, los métodos de evaluación no supervisados no requieren una. Una imagen de referencia creada manualmente es intrínsecamente subjetiva y obtenerlas es un trabajo muy tedioso que consume mucho tiempo. La habilidad de trabajar sin imágenes de referencia permite trabajar en una gran cantidad de escenarios diversos. Las métricas de evaluación no supervisadas evalúan la segmentación de una imagen basándose en la coincidencia de los segmentos con un amplio conjunto de características de imágenes segmentadas.

3.2.1. Métricas para evaluar imágenes RGB-D

Criterio de Dal Mutto et al. (GCFQ)

En [12] utilizan una métrica no supervisada para optimizar los parámetros del algoritmo de segmentación que proponen. La métrica propuesta extiende la medida presentada en [34] para poder utilizar en forma simultánea una imagen a color I, un mapa de profundidad D y el resultado de un algoritmo de segmentación S, en donde se encuentran K regiones $S_i, i = 1...K$.

El criterio de evaluación supone que una buena segmentación tiene que tener dos propiedades fundamentales:

- Dentro de una región segmentada la imagen debe tener propiedades uniformes, es decir, un color constante o algún patrón o textura repetido.
- Cada par de segmentos diferentes tienen que tener distintas propiedades. Esto asegura que no se produzca una sobre-segmentación de la imagen.

El criterio propuesto otorga un valor más alto para segmentaciones que satisfagan ambos criterios, tanto para la imagen de color como para el mapa de profundidad. Primero consideran el mapa de segmentación y la imagen de color normalizada: la evaluación de la primer propiedad es simple para regiones de color constante, con la que se asocia la desviación estándar dentro de la región segmentada, pero es difícil para regiones altamente texturadas. Como se encuentra disponible la información de profundidad, otorgan mayor importancia al componente de color en la métrica en regiones con poca cantidad de texturas y menor importancia en zonas altamente texturadas donde la información de profundidad puede ser más confiable. Para esto se considera la diferencia entre el desvío estándar de los valores de la imagen y el desvío estándar de la región segmentada:

$$\sigma_t(S_i) = \frac{\sum\limits_{j \in S_i^*} \sigma_w(j)}{|S_i^*|}$$

donde $\sigma_w(j)$ es la desviación estándar local calculada en una ventana de 3×3 centrado en el pixel j; S_i^* es el conjunto de píxeles internos del segmento S_i, es decir, aquellos píxeles en los que la ventana $w(j)$ yace completamente dentro del segmento y $|S_i^*|$ es la cardinalidad de S_i^*

La medida de la disparidad interna D^i_{intra} del segmento S_i puede ser calculado del siguiente modo:

$$D^i_{\text{intra}} = \text{máx}(\sigma(S_i) - \sigma_t(S_i), 0)\frac{S_i}{N}$$

donde $\sigma(S_i)$ es la desviación estándar global del color dentro de la región, $|S_i|$ es el número de puntos en la i-ésima región y N es el número total de píxeles. La medida D_{intra} reduce el peso de las regiones altamente texturadas. La medida D_{intra} para toda la imagen consiste en la suma de los valores de D^i_{intra} de cada región:

$$D^{\text{intra}} = \sum_{i=1}^{K} D_i^{\text{intra}}$$

La disparidad entre regiones segmentadas distintas es calculada como las distancias entre los centroides de los *clusters*:

$$D_{i,j}^{\text{inter}} = |E(S_i) - E(S_j)|$$

Estas disparidades son promediadas por todos los pares de segmentos

$$D^{\text{inter}} = \frac{\sum_{i,j \;\; i \neq j} D_{i,j}^{\text{inter}}}{K(K-1)}$$

y la métrica final para el color es calculada como la diferencia entre la disparidad entre diferentes regiones y la disparidad interna dividido 2:

$$Q^{\text{color}}(I,S) = \frac{D^{\text{inter}} - D^{\text{intra}}}{2}$$

La métrica para la información de geometría se calcula de la misma manera pero sin considerar la desviación estándar local:

$$\begin{aligned}
D_i^{\text{intra}} &= \sigma^D(S_i)\frac{|S_i|}{N} \\
D^{\text{intra}} &= \sum_i D_i^{\text{intra}} \\
D_{i,j}^{\text{inter}} &= |E^D(S_i) - E^D(S_j)| \\
D^{\text{inter}} &= \frac{\sum_{i,j \;\; i \neq j} D_{i,j}^{\text{inter}}}{K(K-1)} \\
Q^{\text{depth}}(D,S) &= \frac{D^{\text{Dinter}} - D^{\text{Dintra}}}{2}
\end{aligned}$$

donde $\sigma^D(S_i)$ es la desviación estándar de los valores de geometría en la región S_i y D^{Dinter} se calcula con respecto a los valores del mapa de profundidad. Finalmente la métrica de segmentación combinada se calcula de la siguiente manera:

$$GCFQ(I,D,S) = Q^{\text{color}}(I,S) + 3Q^{\text{depth}}(D,S)$$

3.2.2. Métricas para evaluar imágenes color

Es posible evaluar una segmentación obtenida utilizando solamente la imagen a color para determinar la calidad de la segmentación.

Criterio de Chen et al. (ECW)

Para evaluar la calidad de una segmentación de una imagen a color, en [35] propusieron el uso de la "diferencia de color visible". Para esto utilizan el espacio de color CIE LAB debido a que conforma un espacio de color cuasi-uniforme. En este espacio, la diferencia entre dos colores, ΔE^*_{ab}, está dada por la distancia euclídea de los valores: $\delta E^*_{AB} = ||(L^*_1, a^*_1, b^*_1) - (L^*_2, a^*_2, b^*_2)||_2$ De acuerdo a [36], el valor ΔE^*_{AB} es perceptualmente análogo a la forma de percibir la variación de color que tienen los humanos. Los valores de la métrica pueden ser clasificados en 3 niveles diferentes para reflejar los grados que la diferencia de color son percibidos por el humano. Cuando la diferencia de color es menor a 3, la diferencia es difícilmente perceptible; cuando la diferencia es entre 3 y 6 la diferencia es perceptible pero tolerable, y se torna inaceptable cuando esta diferencia supera 6. En el artículo se propone que una diferencia de color es visible cuando la diferencia entre los colores es mayor a 6.

Basándose en la "diferencia de color visible", se definen dos medidas para evaluar la calidad de la segmentación de una imagen color. La primera medida, llamada "error visual intra-región", se designa para evaluar el grado de sub-segmentación. En cada región segmentada, estos píxeles con diferencias de color visibles que se alejan del valor medio del color de la región, son considerados como píxeles con errores de color visibles. De forma intuitiva, una región segmentada correctamente debería contener la menor cantidad de píxeles con errores de color visible. Dada una imagen $N \times M$ denotada con f, y una imagen segmentada, $\hat{f}$, que contiene para cada segmento su color medio. El "error de intra-región" se define de la siguiente manera:

$$E_{intra} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M} u(\Delta^*_{AB}(f(x,y), \hat{f}(x,y)) - th)}{N \times M}$$

donde th denota el umbral para la diferencia de color visible y $u(.)$ denota la función escalón

$$u(t) = \begin{cases} 1, & t > 0 \\ 0 & c.c \end{cases}$$

La segunda medida, llamada "error visual entre regiones" es diseñada para evaluar el grado de sobre-segmentación. Dada la segmentación de una imagen color, se toman en cuenta los píxeles de contornos con diferencias de color invisible. Intuitivamente, estos píxeles no deberían ser tratados como contornos. Por lo tanto, el error visual entre regiones de una imagen segmentada es:

$$E_{inter} = \frac{\sum_{i=1}^{R}\sum_{j=1}^{R} w_{ij} \times u(th - ||f_i - f_j||_{L*a*b})}{C \times N \times M}$$

donde R denota el número de regiones segmentadas, w_{ij} denota el tamaño de las regiones unidas i y j y es igual a cero si las regiones no están conectadas, C denota el factor de normalización. Los autores proponen utilizar $C = 1/6$.

Para una imagen segmentada, un valor grande del error visual intra-región significa que muchos píxeles pueden estar agrupados de forma incorrecta, y esta imagen puede estar sub-segmentada. Por otro lado, un valor grande del error visual inter-región significa que muchos píxeles de contorno pueden haber sido generados incorrectamente y la imagen puede estar sobresegmentada. La métrica final surge de promediar los dos errores visuales:

$$ECW = \frac{E_{intra} + E_{inter}}{2}$$

Criterio de Zhang (VE)

Basándose en estudios empíricos, en [37] propusieron un criterio de evaluación basándose en la suposición que una buena segmentación debe maximizar la uniformidad de los píxeles dentro de cada región segmentada, y minimizar la uniformidad entre regiones distintas. En consecuencia, utilizan la entropía para medir el desorden dentro de una región. Sea I una imagen en la que cada pixel de la imagen está caracterizado por tres características, por ejemplo R, G, y B; j una región de la imagen, v una de las características para describir la región j y $V_j^{(v)}$ como el conjunto de todos los posibles valores asociados a la característica v en la región j. Entonces, para la región j de la imagen segmentada y el valor m de la característica v en la región, proponen el uso de $L_j(m)$ para denotar el número de píxeles en la región j que tienen el valor m para la característica v en la imagen original. La entropía de la región j se define como:

$$H_v(R_j) = -\sum_{m \in V_j^{(v)}} \frac{L_j(m)}{S_j} log(\frac{L_j(m)}{S_j})$$

$L_j(m)/S_j$ representa la probabilidad de que un pixel en una región R_j tenga una característica con valor m. Por lo tanto $H_v(R_j)$ es el número de bits necesario por pixel para codificar esa característica en la región R_j. Finalmente, se define la entropía esperada para una región I como la entropía esperada por todas las regiones donde cada región está pesada proporcionalmente a su área. Es decir, la entropía de región esperada de una segmentación I es:

$$H_r(I) = -\sum_{j=1}^{N} \frac{S_i}{S_I} H(R_j)$$

Cuando una región tiene un valor uniforme en una de las características, entonces $H_r(I)$ será pequeño. Cuando todos los píxeles tienen el mismo valor, la entropía será 0. Dado que una imagen sobre-segmentada tendrá una entropía esperada muy baja, incorporan una nueva medida que compensa la tendencia la sobre-segmentación. Mientras que la entropía de región esperada decrece con el numero de regiones, se espera que el numero de bits para especificar una región para cada pixel, una medida que llaman, la entropía de layout, aumente con el número de regiones. Por lo tanto, los dos factores puede ser usados para contrarrestar los efectos de la sobre-segmentación.

$$H_l(I) = -\sum_{j=1}^{N} \frac{S_j}{S_I} \log\left(\frac{S_j}{S_I}\right)$$

Finalmente es definida la función de entropía final como

$$E = H_l(I) + H_r(I)$$

Criterio de Liu (F_{Liu})

Con el objetivo de evaluar imágenes reales y sintéticas de forma local y global en [38], se propuso una función objetivo

$$F(I) = \sqrt{N} \sum_{j=1}^{N} \frac{e_j^2}{\sqrt{S_j}}$$

donde I es la imagen a ser segmentada, N, el número de regiones en la imagen segmentada; S_j el número de píxeles de la imagen de la j-ésima región y e_j, el error de color en la región definido como el error cuadrático medio entre los píxeles de la imagen y el color medio de la región j. El término $\sqrt{N}$ es una medida global que penaliza la sobre-segmentación. El término $\frac{e_j^2}{\sqrt{S_j}}$ es una medida local que penaliza pequeñas regiones con un error de color grande.

Criterio de Borsotti (Q)

El método propuesto en [39] consiste en una mejora de la métrica propuesta en [38], mide el error cuadrático medio del color de los segmentos, penalizando la sobre-segmentación y la sub-segmentación:

$$Q(I) = \frac{1}{10000(N \times M)} \sqrt{R} \times \sum_{i=1}^{R} \left[\frac{e_i^2}{1 + \log A_i} + \left(\frac{R(A_i)}{A_i} \right)^2 \right]$$

donde $N \times M$ es el tamaño de la imagen, R es el número de regiones de la imagen segmentada, A_i es el área de la región i-ésima, e_i es la suma de la distancias euclídeas entre los vectores RGB de los píxeles de la región I y el vector de color atribuido a la región i en la imagen segmentada. El valor $R(A_i)$ representa el número de regiones que tienen área igual a A_i.

Criterios para imágenes en escala de grises

Contraste global (Zeboudj) La medida propuesta en [40] toma en cuenta el contraste interno y externo de las regiones medidas en la vecindad de cada pixel. Si denotamos $W(s)$ la vecindad de un pixel s, $f(s)$ la intensidad de pixel, L el valor máximo posible de intensidad, $c(s,t) = \dfrac{|s-t|}{L-1}$ el contraste entre dos píxeles s y t, el contraste dentro de la región i, I_i y el contraste fuera de la región, E_i son respectivamente:

$$I_i = \frac{1}{A_i} \sum_{s \in R_i} \text{máx}\{c(s,t), t \in W(s) \cup R_i\}$$

$$E_i = \frac{1}{l_i} \sum_{s \in R_i} \text{máx}\{c(s,t), t \in W(s), t \notin R_i\}$$

donde A_i es la superficie y F_i es el borde (de longitud l_i) de la región R_i. El contraste de la región R_i es:

$$C(R_i) = \begin{cases} 1 - \frac{I_i}{E_i} & \text{si } 0 < I_i < E_i \\ E_i & \text{si } I_i = 0 \\ 0 & \text{Caso contrario.} \end{cases}$$

El contraste global es $C = \dfrac{1}{A} \sum_i A_i C(R_i)$. Este criterio aumenta con la calidad de la segmentación. Pero no es muy adecuado para imágenes muy ruidosas o texturadas [41].

Capítulo 4

Resultados experimentales

Con el objetivo de evaluar los métodos de segmentación utilizando las métricas descriptas se utilizó el conjunto de datos provisto por la universidad de Nueva York, denominado NYUD. Este conjunto de datos está formado por imágenes obtenidas utilizando una cámara Microsoft Kinect v1 en múltiples escenarios interiores. Las imágenes presentan distintas condiciones de iluminación, y las escenas cuentan con objetos poco comunes y, en múltiples ocasiones, se encuentran parcialmente ocluidos. Cada imagen cuenta con la información de color, de distancia y con un mapa de segmentación denso que da cuenta de los objetos presentes en la escena. El conjunto de datos consiste de 1449 imágenes RGB-D obtenidas en un amplio rango de edificaciones comerciales y residenciales en tres ciudades estadounidenses. Las etiquetas para los segmentos se obtuvieron utilizando el servicio de Amazon Mechanical Turk. El dataset contiene 35064 objetos distintos.

El conjunto etiquetado está compuesto por pares de imágenes de RGB y profundidad que han sido sincronizadas. El archivo con el conjunto de datos etiquetado es un archivo de Matlab con múltiples variables definidas. Las relevantes para el presente trabajo son las siguientes:

- *depths*. Esta variable contiene un tensor de dimensión $H \times W \times N$ que contiene los mapas de profundidad, donde H y W son la altura y el ancho de cada imagen, respectivamente y N es el número de imágenes del conjunto de datos. Los valores están en metros.

- *images*. Esta variable es un tensor de $H \times W \times 3 \times N$ que contiene las imágenes RGB donde H es la altura y W es al ancho de cada imagen y N es el número de imágenes.

JCSA		GCF		CDNGraph	
Parámetro	Valor	Parámetro	Valor	Parámetro	Valor
kp	5	Número de autovectores	7	Tamaño mínimo de región	1000
th_d	3	λ	[0.0001, 0.001, 0.01, 0.1, 0.5, 0.8]		
th_b	0.2	Proporción de píxeles	0,01		
th_r	0.9				
Tamaño mínimo de región	5				
Máximas iteraciones	500				

Tabla 4.1: Parámetros utilizados para cada algoritmo

- *labels*. Esta variable es un tensor $H \times W \times N$ que contiene el mapa etiquetado en donde H y W son la altura y el ancho de cada mapa respectivamente y N es el número de imágenes. El rango de las etiquetas va desde 1 hasta C donde C es el número total de clases. Si un píxel es etiquetado en 0, se considera que el píxel no tiene etiqueta.

Las imágenes fueron segmentadas utilizando los algoritmos descriptos en el capítulo 2. Los parámetros de cada algoritmo están descriptos en la tabla 4.1. Estos parámetros son los recomendados por los autores de cada método. El número de *clusters* buscado en cada caso fue de 25. Cada imagen del conjunto de datos fue reducida a la mitad de tamaño. Luego del proceso de segmentación, se obtienen los componentes conectados para determinar el número final de segmentos.

4.1. Resultados obtenidos

4.1.1. Tiempo de cómputo

En la figura 4.1 se puede visualizar una estimación de la distribución de tiempos de cómputos para cada algoritmo. La segmentación fue realizada en una computadora de escritorio con un procesador Intel i7-7700 y 16 GB de RAM.

La velocidad de cómputo es menor para CDNGraph. Generalmente GCF requiere un menor tiempo de cómputo que JCSA. El tiempo de GCF proviene del hecho de que se debe realizar una segmentación completa para cada valor del parámetro λ con el objetivo de determinar cual es su valor óptimo. Es posible reducir los tiempos de cómputo de JCSA reduciendo el número

máximo de iteraciones posibles del algoritmo EM.

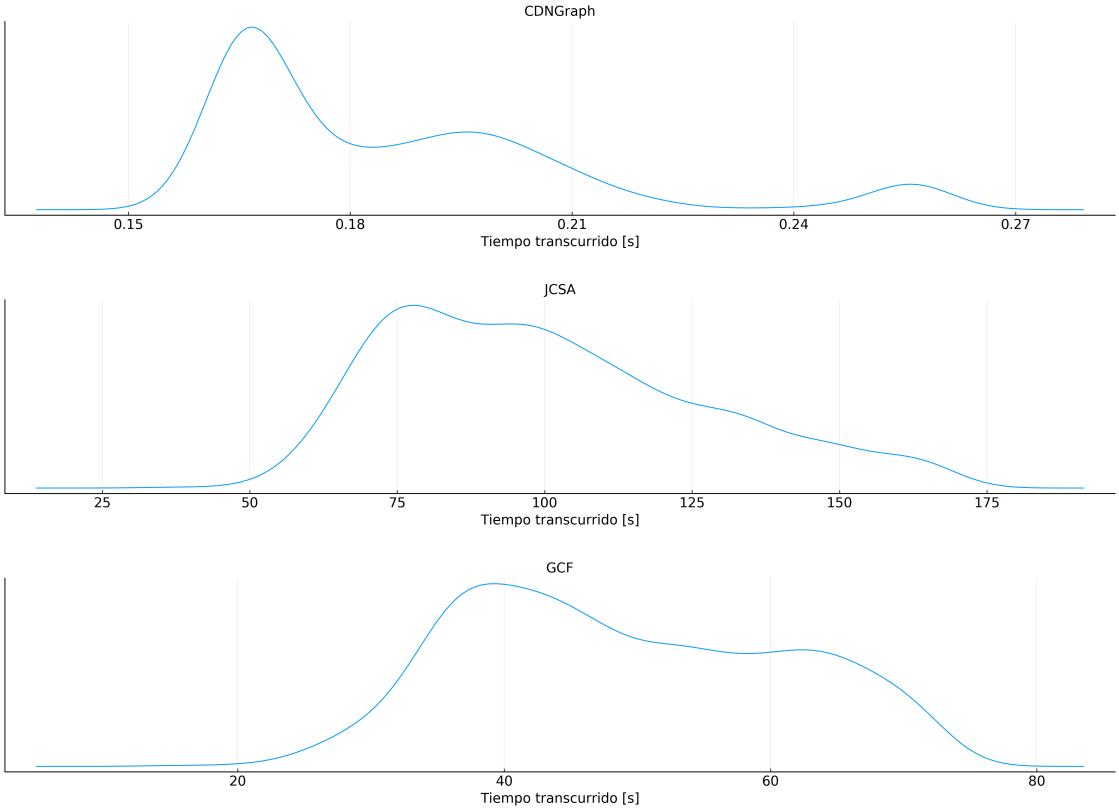


Figura 4.1: Tiempos de cómputo

4.1.2. Número de segmentos obtenidos

La figura 4.2 muestra un histograma del número de segmentos obtenido por cada algoritmo. El algoritmo CDNGraph tiende a sub-segmentar la imagen. En un gran número de imágenes segmentadas, todos los píxeles pertenecen a un único segmento. La varianza del número de segmentos es menor en JCSA y el número de *clusters* finalmente obtenido es más cercano al número de segmentos buscado.

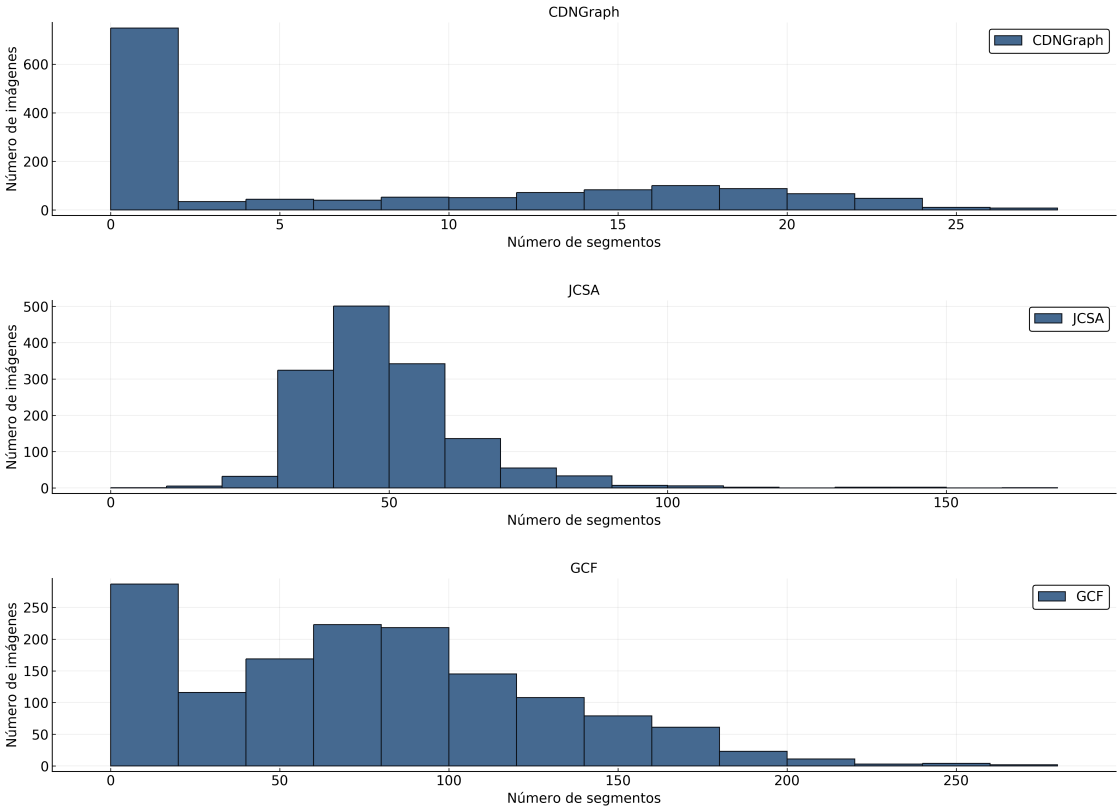
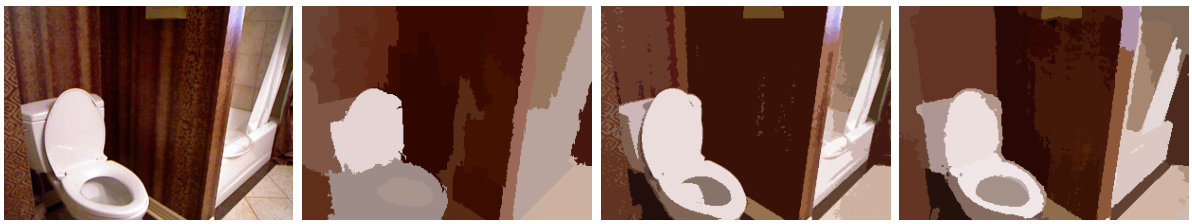


Figura 4.2: Número de segmentos obtenidos

4.2. Análisis cualitativo

Las figuras 4.3, 4.4 y 4.5 muestran tres imágenes reales del dataset y los resultados de aplicar el algoritmo de segmentación sobre ellas. Es posible notar algunas diferencias entre las segmentaciones obtenidas:

CDNGraph tiende a subsegmentar las imágenes y no recobrar adecuadamente las estructuras presentes en la escena, uniendo por ejemplo, el piso y el suelo. JCSA y GCF recobran de una manera mas adecuada las estructuras planas características de las escenas interiores. Sin embargo, GCF tiende a sobresegmentar regiones homogéneas aunque sus resultados son más visualmente agradables. Esto se debe a que realiza una una mejor estimación de los contornos de los objetos. GCF es más sensible a la información de color mientras que JCSA es mas sensible a la información direccional. CDNGraph tiende a perder detalles de las estructuras de la escena y parece ser muy sensitivo a las condiciones lumínicas.



(a) Imagen original (b) CDNGraph N: 20 (c) GC. Ns: 30 (d) JCSA N: 29

Figura 4.3: Segmentación de la imagen 707 del dataset. N es el número de segmentos obtenido

(a) Imagen original (b) CDNGraph N: 21 (c) GCF N: 35 (d) JCSA N: 28

Figura 4.4: Segmentación de la imagen 1166 del dataset. N es el número de segmentos obtenido

(a) Imagen original (b) CDNGraph N: 13 (c) GCF N: 39 (d) JCSA N: 30

Figura 4.5: Segmentación de la imagen 1435 del dataset. N es el número de segmentos obtenido

4.3. Métricas de desempeño

La tabla 4.2 muestra los resultados promedios obtenidos al utilizar las métricas supervisadas para evaluar la calidad de la segmentación obtenida con cada algoritmo. En cada caso los mejores resultados aparecen en negrita. Las métricas Precision, F_b, SegCov, Medida F, F_r, F_{op} y PRI son mejores cuanto más alto es su valor. Por otro lado, BDE y VoI son mejores cuanto más bajo es su valor. Los resultados muestran que la performance del algoritmo JCSA mejora PRI, VoI, SegCov and BDE. La figura 4.6 muestra la distribución de 3 métricas supervisadas. Es

posible notar que F_{op} y F_b permiten discriminar los diferentes algoritmos utilizados. *SegCov* no presenta diferencia significativas entre los algoritmos evaluados.

	CDNGraph	JCSA	GCF
Precision	0.608	**0,809**	0.781
F_b	0.29	**0,472**	0.362
SegCov	0.263	0.256	**0,277**
Medida F	**0,393**	0.373	0.378
BDE	8.884	**5,928**	7.124
F_r	0.291	**0,304**	0.301
F_{op}	0.023	**0,028**	0.021
VoI	**6,953**	7.34	7.015
PRI	0.584477	0.814172	**0,819**

Tabla 4.2: Métricas supervisadas para el dataset NYU

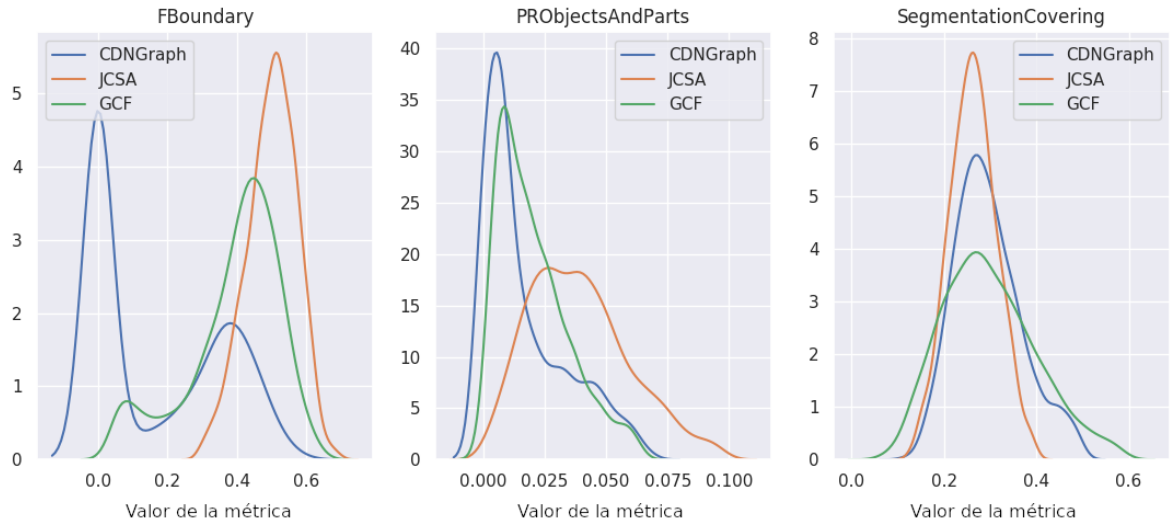


Figura 4.6: Distribución de métricas supervisadas

La tabla 4.3 presenta los valores promedios al utilizar las métricas no supervisadas. Las

métricas no supervisadas $GCFQ$ y $Zeboudj$ indican una mejor calidad de la segmentación cuanto mayor es su valor, Por otro lado, ECW, VE, F_{Liu}, Q son mejores cuanto mas bajo es su valor. Es posible ver que ECW y F_{Liu} no pueden abordar correctamente el problema de sub-segmentación dado que le otorgan mejor puntaje a CDNGraph. La figura 4.7 muestra la distribución de las métricas $Zeboudj$, Q y VE. Es posible ver que los métodos pueden ser distinguidos de acuerdo a estas métricas, aunque la distinción no es tan clara en comparación con las métricas supervisadas.

	CDNGraph	JCSA	GCF
GCFQ	3.579	16.051	**37,019**
ECW	**56,157**	788.338	93.736
VE	30.029	**25,708**	27.007
F_{Liu}	**0,0034**	0.054	0.307
Zeboudj	0.123	0.344	**0,381**
Q	0.0201	**0,008**	0.047

Tabla 4.3: Métricas no supervisadas para el dataset NYU

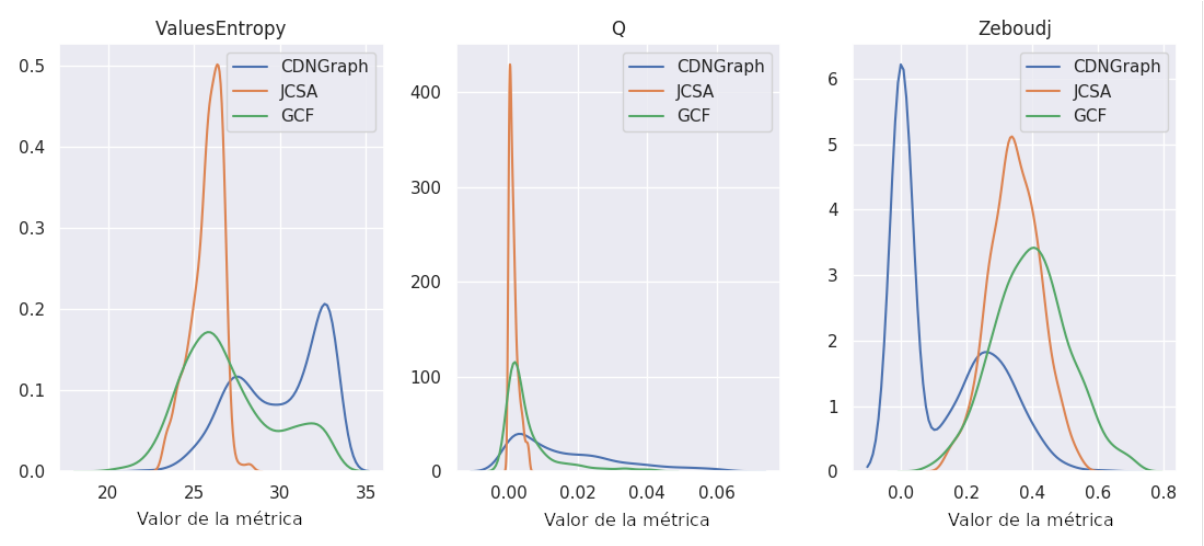


Figura 4.7: Distribución de métricas no supervisadas

Capítulo 5

Conclusiones

En el presente trabajo se compararon algunos de los métodos más relevantes para la segmentación de imágenes RGB-D. Se analizó, a su vez, el uso de métricas supervisadas y no supervisadas para determinar la calidad de las segmentaciones obtenidas. El algoritmo JCSA presenta, en general, mejores resultados que CDNGraph y GCF, tal como demuestran las métricas supervisadas. Sin embargo, su tiempo de cómputo puede hacerlo prohibitivo para ciertas aplicaciones. Las técnicas de *clustering* espectral obtienen una buena aproximación de los contornos de los objetos en un menor tiempo que JCSA. Sin embargo, puede no determinar correctamente el número de segmentos.

Las métricas supervisadas F_{op} y F_b permiten evaluar comparativamente el desempeño de los métodos de segmentación considerados. En cuanto a las métricas no supervisadas, VE y Q permiten discriminar los diferentes algoritmos, aunque de un modo no tan claro como las métricas supervisadas.

Como resultado de este trabajo se han desarrollado dos librerías de acceso público: la primera de ellas, [42], contiene los métodos de segmentación descriptos en el trabajo y las utilidades para realizar la comparaciones descriptas. La segunda, [43], contiene la implementación de los criterios de evaluación supervisados y no supervisados.

Un posible trabajo futuro consiste en la incorporación de la información de distancia en las métricas de evaluación no supervisadas para determinar de una mejor manera la calidad de un método de segmentación en imágenes RGB-D. Otra posible linea de trabajo futuro consiste en la comparación de las técnicas no supervisadas estudiadas con los recientes métodos supervi-

sados de segmentación de imágenes. Finalmente, es posible analizar mejoras en las técnicas de segmentación que utilizan conceptos de *clustering* espectral para estimar de una manera más precisa el número final de segmentos.

Capítulo 6

Bibliografía

[1] J. Han, L. Shao, D. Xu, and J. Shotton, "Enhanced computer vision with microsoft kinect sensor: A review," *IEEE transactions on cybernetics*, vol. 43, no. 5, pp. 1318–1334, 2013.

[2] N. Gupta, W. Xu, and D. Kamboj, "Depth-based segmentation—a review," in *Computing, Communications and Networking Technologies (ICCCNT), 2013 Fourth International Conference on*, pp. 1–9, IEEE, 2013.

[3] M. A. Hasnat, O. Alata, and A. Trémeau, "Unsupervised rgb-d image segmentation using joint clustering and region merging," *J-STSP*, vol. 6, no. 5, pp. 505–521, 2012.

[4] C. Dal Mutto, P. Zanuttigh, G. M. Cortelazzo, and S. Mattoccia, "Scene segmentation assisted by stereo vision," in *3D Imaging, Modeling, Processing, Visualization and Transmission (3DIMPVT), 2011 International Conference on*, pp. 57–64, IEEE, 2011.

[5] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik, "Learning rich features from rgb-d images for object detection and segmentation," in *European Conference on Computer Vision*, pp. 345–360, Springer, 2014.

[6] A. Eitel, J. T. Springenberg, L. Spinello, M. Riedmiller, and W. Burgard, "Multimodal deep learning for robust rgb-d object recognition," in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 681–687, IEEE, 2015.

[7] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[8] C. Couprie, C. Farabet, L. Najman, and Y. LeCun, "Indoor semantic segmentation using depth information," *arXiv preprint arXiv:1301.3572*, 2013.

[9] J. Pont-Tuset and F. Marques, "Supervised evaluation of image segmentation and object proposal techniques," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 7, pp. 1465–1478, 2016.

[10] R. Szeliski, *Computer vision: algorithms and applications*. Springer Science & Business Media, 2010.

[11] M. A. Hasnat, O. Alata, and A. Tremeau, "Joint color-spatial-directional clustering and region merging (jcsd-rm) for unsupervised rgb-d image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 11, pp. 2255–2268, 2016.

[12] C. Dal Mutto, P. Zanuttigh, and G. M. Cortelazzo, "Fusion of geometry and color information for scene segmentation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 5, pp. 505–521, 2012.

[13] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *International journal of computer vision*, vol. 59, no. 2, pp. 167–181, 2004.

[14] S. Sra and D. Karp, "The multivariate watson distribution: Maximum-likelihood estimation and other aspects," *Journal of Multivariate Analysis*, vol. 114, pp. 256–269, 2013.

[15] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*, vol. 1. Springer series in statistics New York, NY, USA:, 2001.

[16] Z. Zivkovic, "Improved adaptive gaussian mixture model for background subtraction," in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, vol. 2, pp. 28–31, IEEE, 2004.

[17] D.-S. Lee, "Effective gaussian mixture learning for video background subtraction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 5, pp. 827–832, 2005.

[18] H. H. Permuter, J. M. Francos, and I. Jermyn, "A study of gaussian mixture models of color and texture features for image classification and segmentation," *Pattern Recognition*, vol. 39, no. 4, pp. 695–706, 2006.

[19] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh, "Clustering with bregman divergences," *Journal of Machine Learning Research*, vol. 6, pp. 1705–1749, 2005.

[20] A. Trémeau and P. Colantoni, "Regions adjacency graph applied to color image segmentation," *IEEE Transactions on image processing*, vol. 9, no. 4, pp. 735–744, 2000.

[21] C. J. Taylor and A. Cowley, "Parsing indoor scenes using rgb-d imagery," in *Robotics: Science and Systems*, vol. 8, pp. 401–408, 2013.

[22] J. Shi and J. Malik, "Normalized cuts and image segmentation," *Departmental Papers (CIS)*, p. 107, 2000.

[23] C. Fowlkes, S. Belongie, F. Chung, and J. Malik, "Spectral grouping using the nystrom method," *IEEE transactions on pattern analysis and machine intelligence*, vol. 26, no. 2, pp. 214–225, 2004.

[24] T. Kanungo, B. Dom, W. Niblack, and D. Steele, "A fast algorithm for mdl-based multiband image segmentation," in *1994 Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 609–616, Jun 1994.

[25] J. S. Cardoso and L. Corte-Real, "Toward a generic evaluation of image segmentation," *IEEE Transactions on Image Processing*, vol. 14, no. 11, pp. 1773–1782, 2005.

[26] B. Larsen and C. Aone, "Fast and effective text mining using linear-time document clustering," in *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 16–22, ACM, 1999.

[27] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, pp. 898–916, May 2011.

[28] M. Meilă, "Comparing clusterings by the variation of information," in *Learning theory and kernel machines*, pp. 173–187, Springer, 2003.

[29] M. Meilă, "Comparing clusterings: an axiomatic view," in *Proceedings of the 22nd international conference on Machine learning*, pp. 577–584, ACM, 2005.

[30] D. R. Martin, J. Malik, and D. Patterson, *An empirical approach to grouping and segmentaqtion.* Computer Science Division, University of California, 2003.

[31] W. M. Rand, "Objective Criteria for the Evaluation of Clustering Methods," *Journal of the American Statistical Association*, vol. 66, pp. 846–850, Dec. 1971.

[32] D. R. Martin, C. C. Fowlkes, and J. Malik, "Learning to detect natural image boundaries using local brightness, color, and texture cues," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 5, pp. 530–549, 2004.

[33] J. Freixenet, X. Muñoz, D. Raba, J. Martí, and X. Cufí, "Yet another survey on image segmentation: Region and boundary information integration," in *European conference on computer vision*, pp. 408–422, Springer, 2002.

[34] C. Rosenberger and K. Chehdi, "Genetic fusion: application to multi-components image segmentation," in *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, vol. 4, pp. 2223–2226, IEEE, 2000.

[35] H.-C. Chen and S.-J. Wang, "The use of visible color difference in the quantitative evaluation of color image segmentation," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 3, pp. iii–593, IEEE, 2004.

[36] J. Hardeberg, *Acquisition et reproduction d'images couleur: approches colorimétrique et multispectrale.* PhD thesis, Télécom ParisTech, 1999.

[37] H. Zhang, J. E. Fritts, and S. A. Goldman, "An entropy-based objective evaluation method for image segmentation," in *Storage and Retrieval Methods and Applications for Multimedia 2004*, vol. 5307, pp. 38–50, International Society for Optics and Photonics, 2003.

[38] J. Liu and Y.-H. Yang, "Multiresolution color image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, no. 7, pp. 689–700, 1994.

[39] M. Borsotti, P. Campadelli, and R. Schettini, "Quantitative evaluation of color image segmentation results," *Pattern recognition letters*, vol. 19, no. 8, pp. 741–747, 1998.

[40] R. Zéboudj, *Filtrage, seuillage automatique, contraste et contours: du pré-traitement à l'analyse d'image*. PhD thesis, Saint-Etienne, 1988.

[41] S. Ouattara, G. L. Loum, A. Clément, and B. Vigouroux, "Analysis of the relevance of evaluation criteria for multicomponent image segmentation," *Journal of Software Engineering and Applications*, vol. 4, no. 06, p. 371, 2011.

[42] L. Lorenti, "RGBDSegmentation.jl." `https://github.com/lucianolorenti/RGBDSegmentation.jl`, 2019. [Online; accedido 28-Mayo-2019].

[43] L. Lorenti, "ImageSegmentationEvaluation.jl." `https://github.com/lucianolorenti/ImageSegmentationEvaluation.jl`, 2019. [Online; accedido 28-Mayo-2019].

[44] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Advances in neural information processing systems*, pp. 849–856, 2002.

[45] P. Jaccard, "Étude comparative de la distribution florale dans une portion des alpes et des jura," *Bull Soc Vaudoise Sci Nat*, vol. 37, pp. 547–579, 1901.

Apéndice A

Apéndice

A.1. Clustering espectral

Dado un conjunto de patrones $X = \{x_1, x_2, ...x_n\} \in \mathbb{R}^m$, y una función de semejanza $d : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$, es posible construir una matriz de afinidad W tal que $W(i,j) = d(x_i, x_j)$. Los algoritmos de agrupamiento espectral obtienen una representación de los datos en un espacio de dimensión inferior resolviendo el siguiente problema de optimización:

$$\begin{aligned} \max_{U \in \mathbb{R}^{n \times k}} \quad & \operatorname{Tr} U^T L U \\ \text{s.a.} \quad & U^T U = I \end{aligned} \tag{A.1}$$

donde $L = D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$ es la matriz laplaciana de W de acuerdo a [44] y D es una matriz diagonal con la suma de las filas de W ubicadas en su diagonal principal. Una vez obtenido U sus filas son consideradas como las nuevas coordenadas de los patrones. En esta nueva representación es más sencillo aplicar un algoritmo de clustering tradicional [22].

Es posible obtener una aproximación a las coordenadas en este nuevo espacio calculando las afinidades de un pequeño conjunto de píxeles y aproximando las afinidades restantes.

Sea $A \subset X$ un subconjunto de patrones muestreados y $B = V - A$, el resto de los patrones no muestreados. W_A es la matriz de semejanza derivada de los datos de A y L_A es la matriz laplaciana de W_A. W_B y L_B son las matrices correspondientes de las afinidades de los puntos de A y B. Es posible definir a L como:

$$W = \begin{bmatrix} W_A & W_B \\ W_B^T & W_C \end{bmatrix} \qquad L = \begin{bmatrix} L_A & L_B \\ L_B^T & L_C \end{bmatrix}$$

Es posible obtener una aproximación de W, denominada $\hat{W}$, solamente a partir de A y B:

$$\hat{W} = \bar{U}\Lambda\bar{U}^T = \begin{bmatrix} A & B \\ B^T & B^T A^{-1} B \end{bmatrix}$$

Con el objetivo de obtener los autovectores de la matriz laplaciana aproximada, $\hat{L} = \hat{D}^{\frac{1}{2}}\hat{W}\hat{D}^{\frac{1}{2}}$, es necesario calcular $\hat{L_A}$ y $\hat{L_B}$:

$$\hat{L_A}_{ij} = \frac{W_{A_{ij}}}{\sqrt{\hat{d}_i \hat{d}_j}} \qquad \hat{L_B}_{ij} = \frac{W_{B_{ij}}}{\sqrt{\hat{d}_i \hat{d}_{j+|A|}}} \tag{A.2}$$

donde $\hat{d} = \hat{W}\mathbf{1}$. Si $\hat{L_A}$ es positiva definida, es posible hallar los autovectores ortogonales aproximados en un solo paso. Sea $S = \hat{L_A} + \hat{L_A}^{-\frac{1}{2}} \hat{L_B} \hat{L_B}^T \hat{L_A}^{-\frac{1}{2}}$ y su diagonalización $S = U_S \Lambda_S U_S^T$, Fowkles et al. [23] demostraron que si la matriz V se define como

$$V = \begin{bmatrix} \hat{L_A} \\ \hat{L_B}^T \end{bmatrix} \hat{L_A}^{-\frac{1}{2}} U_S \Lambda_S^{-\frac{1}{2}} \tag{A.3}$$

$\hat{L}$ es diagonalizada por V y por Λ_S y $V^T V = I$

A.2. Medidas de evaluación supervisadas

A.2.1. Precision y Recall

El objetivo de un método de segmentación es alcanzar una detección perfecta, es decir, $P_m = P_{gt}$. Si esto no sucede es posible analizar los siguientes conjuntos para determinar la calidad de una segmentación:

- Verdadero positivos: Píxeles que son detectados como objetos y son etiquetados de la misma manera en el *ground truth*. $TP = P_m \cap P_{gt}$
- Falsos positivos: Píxeles que son detectados como objetos pero están etiquetados del mismo modo en el *ground truth*: $FP = P_m \cap N_{gt}$.
- Falsos negativos: Píxeles que son clasificados como no-objetos pero se encuentran etiquetados como objetos en el *ground truth*. $FN = N_m \cap P_{gt}$.

El objetivo es maximizar los verdaderos positivos y al mismo tiempo minimizar los faltos positivos y falsos negativos. Un par de medidas ampliamente utilizadas para determinar la calidad de un algoritmo de detección es la siguiente:

- Precisión: Medidas del porcentaje de píxeles detectados que son realmente verdaderos: Precision $= \frac{|TP|}{|P_m|} = \frac{|P_m \cap P_{gt}|}{|P_m|} \leq 1$
- Exactitud: Mide el porcentaje de de puntos positivos en el ground-truth que fueron detectados correctamente: Exactitud $= \frac{|TP|}{|P_{gt}|} = \frac{|P_m \cap P_{gt}|}{|P_{gt}|} \leq 1$

Nuestro objetivo es maximizar ambas medidas, pero por lo general existe un compromiso entre ambas. Por ejemplo, marcar toda la imagen como un objeto nos da la mayor cantidad de verdaderos positivos y falsos negativos y una gran cantidad de falsos positivos. Es decir que nos da un recall perfecto, pero una muy baja precisión. Con el objetivo de medir el trade-off entre las dos medidas, la medida F se define como la media armónica pesada entre precision y *recall*:

$$F_B = (1 + B^2)\frac{Precision \cdot Recall}{Precision + Recall}$$

donde B nos permite pesar la precision y el recall de forma diferente. Si se le otorga la misma importancia a ambas medidas, entonces la medida F es la media armónica.

$$F_1 = 2\frac{Precision \cdot Recall}{Precision + Recall}$$

En terminos de detección de verdaderos y falsos, esta medida puede ser reescrita como:

$$F = \frac{2|TP|}{2|TP| + |FN| + |FP|}$$

A.2.2. Jaccard

El coeficiente de semejanza de Jaccard El índice de Jaccard fue introducido en el contexto de la sociología de la fitosociología en 1901 [45], y en el contexto de la segmentación de objetos es usualmente denominado como la Intersección sobre la unión (IoU por sus siglas en ingles) entre los resultados de un método automático y el ground truth:

$$J(P_m, P_{gt}) = \frac{|P_m \cap P_{gt}|}{|P_m \cup P_{gt}|} = \frac{|TP|}{|TP| + |FN| + |FP|}$$

La medida espacial de calidad propuesta por el comite MPEG7 para determinar la calidad de una segmentación se define de la siguiente manera: $JQM((P_m, P_{gt}) = \frac{|FP| + |FN|}{|P_{gt}|} = 1 - \frac{|TP| - |FP|}{|P_{gt}|}$

Este enfoque no toma en cuenta el tamaño de la región considerada $|P_m|$, de forma que el comportamiento no es consistente para regiones de distinto tamaño. La medida de Jaccard y la medidad F están relacioandas:

$$\frac{F}{2 - F} = J$$