Segmentación de imágenes de tiempo de vuelo vía clustering espectral co-regularizado

Luciano Lorenti, Javier Giacomantone, Armando De Giusti

Instituto de Investigación en Informática (III-LIDI),
Facultad de Informática - Universidad Nacional de La Plata - Argentina.
La Plata, Buenos Aires, Argentina.
{llorenti,jog,degiusti}@lidi.info.unlp.edu.ar

Resumen. Las cámaras de tiempo de vuelo (TOF) generan dos imágenes simultaneas, una de intensidad y una de rango. Esto permite abordar problemas de segmentación donde la información de intensidad o de rango separadamente es insuficiente para extraer los objetos de interés de la escena 3D. En este artículo se presenta un método de segmentación espectral, que combina información de ambas imágenes. Modificando la matriz de afinidad de cada una de las imágenes en función de la otra, se mejora la segmentación de los objetos de la escena. El método propuesto explota dos mecanismos, el primero orientado a reducir la demanda computacional en el cálculo de autovectores de cada matriz, y el segundo destinado a mejorar el rendimiento de la segmentación. Se presentan resultados experimentales sobre dos conjuntos de imágenes reales, que permiten evaluar el método propuesto.

Palabras clave: Segmentación, Imágenes de Rango, Cámaras de Tiempo de Vuelo, Agrupamiento Espectral

1. Introducción

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 afectará a las etapas siguientes [1][2]. Algoritmos de visión por computador, en particular de segmentación, que han sido utilizados con éxito en ambientes industriales, con colores e iluminación controlada, no obtienen resultados similares en contextos diferentes. Una alternativa para abordar problemas en que las condiciones de contorno no permiten una segmentación adecuada es incorporar información de profundidad, es decir, la distancia a la que se encuentran los objetos que conforman la escena respecto al dispositivo de captura [3] [4]. En este contexto, la segmentación de imágenes consiste en utilizar algoritmos que utilicen ambas fuentes de información y no sólo los niveles de intensidad [5][6]. Con esta perspectiva el problema de segmentación puede ser formulado como la búsqueda de formas efectivas para particionar adecuadamente un conjunto de muestras con información de intensidad y distancia. En particular en este trabajo utilizamos una cámara de tiempo

de vuelo,"Time of Flight" (TOF), que nos permite obtener imágenes de rango y de intensidad simultáneamente, la cámara utilizada es la MESA SR 4000 [7]. La SR 4000 es una cámara activa, utiliza su propia fuente de iluminación mediante una matriz de diodos emisores de luz infrarroja modulada en amplitud. Los sensores de la cámara detectan la luz reflejada en los objetos iluminados y la cámara genera dos imágenes. La imagen de intensidad es proporcional a la amplitud de la onda reflejada y la imagen de rango o distancia es generada a partir de la diferencia de fase entre la onda emitida y reflejada en cada elemento de la imagen [8]. Las principales ventajas con respecto a otras técnicas de medición 3D es la posibilidad de obtener imágenes a velocidades compatibles con aplicaciones en tiempo real y la posibilidad de obtener nubes de puntos 3D desde un solo punto de vista [9][10]. Debido a la complejidad computacional requerida por los algoritmos de clustering espectral, recientemente han sido propuestos métodos que facilitan el cálculo de los autovectores de la matriz de afinidad [11] [12] [13]. Han sido utilizadas técnicas de agrupamiento que a partir de múltiples representaciones de los datos permiten mejorar el proceso de agrupamiento [14] [15] [16]. El método propuesto explota dos mecanismos, el primero orientado a reducir la demanda computacional en el cálculo de autovectores de cada matriz, y el segundo destinado a mejorar el rendimiento de la segmentación. La mejora en la demanda computacional se logra mediante la aproximación de los autovectores de la matriz de afinidad derivada de cada una de las imágenes. La segmentación es mejorada con respecto a la utilización de una sola imagen, mediante un mecanismo iterativo que permite obtener el espacio de autovectores óptimo para realizar la segmentación. La evaluación del método propuesto se realiza mediante dos conjuntos de datos de imágenes reales, el primero obtenido por una cámara de tiempo de vuelo, el segundo facilitado por el laboratorio del Laboratorio de Tecnología Multimedia y Telecomunicaciones de la Universidad de Padua [17].
El artículo está organizado del siguiente modo, en la sección 2 se presenta una revisión de los conceptos fundamentales utilizados en el método propuesto. En la sección 3 se expone el método. En la sección 4 se presentan resultados experimentales. Finalmente en la sección 5 se presentan las conclusiones.

2. Agrupamiento 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:

$$\begin{array}{ll} \underset{U \in \mathbb{R}^{n \times k}}{\text{máx}} & Tr\left(U^T L U\right) \\ \text{s.t.} & U^T U = I \end{array} \tag{1}$$

donde $L = D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$ es la matriz laplaciana de W de acuerdo a [18] 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 mas sencillo aplicar un algoritmo de clustering tradicional [19].

Los métodos espectrales de segmentación de imágenes están basados en los autovectores y autovalores de una matriz $N \times N$ derivada de las afinidades entre los píxeles. Es importante destacar que una de las limitaciones principales de ésta clase de algoritmos es la cantidad de memoria requerida debido a que las dimensiones de W crecen cuadráticamente con respecto al número de elementos de la imagen. Un enfoque posible para abordar este problema consiste en utilizar una matriz dispersa que codifique la información local de cada píxel. En esta representación cada elemento es conectado sólo a alguno de sus vecinos más cercanos en el plano de la imagen y todas sus otras conexiones se asumen cero [19] [20]. Otra alternativa posible consiste en calcular las afinidades de un pequeño conjunto de píxeles y aproximar las afinidades restantes [11][12].

2.1. Cálculo aproximado de los autovectores

Una de las propuestas iniciales para definir algoritmos de agrupamiento espectral relaciona la matriz de pesos W con la matriz de incidencia de un grafo y al problema de clustering como un problema de particionado de grafos [19]. Bajo esta perspectiva cada uno de los patrones x_i son considerados como vértices de un grafo pesado no dirigido $G = (V, E)$ y el elemento $W(i, j)$ es el peso de la arista que conecta al vértice i con el vértice j.

Sea $G = (V, E)$ el grafo de semejanza derivado de un conjunto de patrones $X = \{x_1, x_2, ..., x_n\}$, $A \subset V$ un subconjunto de vértices muestreados y $B = V - A$, el resto de los vértices no muestreados. G_A es el grafo que resulta de conectar los vértices de A entre sí y G_B el grafo que resulta de conectar los vértices de A con los vértices de B. Sea W_A la matriz de adyacencia de G_A y L_A la matriz laplaciana de G_A. W_B y L_B las matrices correspondientes de G_B. Podemos entonces formular la matriz de adyacencia de G, que llamaremos W y la matriz laplaciana de G, que llamaremos L, de la siguiente forma:

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

Considerando la diagonalización de $A = U \Lambda U^T$, si se utiliza la extensión de Nystrom [11]: $\bar{U} = \begin{bmatrix} U \\ B^T U \Lambda^{-1} \end{bmatrix}$ como autovectores aproximados de W, es posible obtener una aproximación de W, denominada $\hat{W}$, calculando solamente 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 $\hat{L} = \hat{D}^{\frac{1}{2}}\hat{W}\hat{D}^{\frac{1}{2}}$, es decir, la matriz laplaciana aproximada generada a partir de $\hat{W}$ es posible utilizar la técnica propuesta, que solo requiere calcular $\hat{L_A}$ y $\hat{L_B}$:

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

donde $\hat{d} = \hat{W}\mathbf{1} = \begin{bmatrix} a_r + b_r \\ b_c + B^T A^{-1} b_r \end{bmatrix}$ y a_r representa la suma de las filas de A, b_c representa la suma de las columnas de B y b_r la suma de las filas de B. 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 [11] demostraron que si la matriz V se define como

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

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

2.2. Co-regularización

Cuando el conjunto de datos tiene más de una representación, a cada una de ellas se las denomina vistas. En el contexto de agrupamiento espectral las técnicas de co-regularización intentan fomentar la semejanza de los ejemplos en la nueva representación generada a partir de los autovectores de cada una de las vistas.

Sea $X^{(v)} = \{x_1^{(v)}, x_2^{(v)}, ..., x_n^{(v)}\}$ los ejemplos para la vista v y $L^{(v)}$ la matriz laplaciana creada a partir de X para la vista v. Definimos $U^{(v)}$ a la matriz formada por los primeros k autovectores correspondientes a la matriz $L^{(v)}$ de acuerdo con (1). En [15] fue propuesto un criterio que mide la desemejanza entre dos representaciones:

$$D(U^{(v)}, U^{(w)}) = \left|\left| \frac{K_{U^{(v)}}}{||K_{U^{(v)}}||_F} - \frac{K_{U^{(w)}}}{||K_{U^{(w)}}||_F} \right|\right|_F^2$$

Donde $K_{U^{(v)}}$ es la matriz de semejanza generada a partir de los patrones en la nueva representación $U^{(v)}$ y $||\cdot||_F$ es la norma Frobenius. Si se utiliza como medida de semejanza el producto interno entre los vectores se obtiene $K_{U^{(v)}} = U^{(v)}{U^{(v)}}^T$. Ignorando las constantes aditivas y de escalado, la ecuación anterior puede ser formulada de la siguiente manera:

$$D(U^{(v)}, U^{(w)}) = -Tr\left(U^{(v)}{U^{(v)}}^T U^{(w)}{U^{(w)}}^T\right)$$

El objetivo es minimizar el desacuerdo entre las representaciones obtenidas a partir de cada una de las vistas. Por lo tanto se obtiene el siguiente problema de optimización que combina los objetivos de agrupamiento espectral individuales y el objetivo que determina el desacuerdo entre las representaciones:

$$\begin{array}{ll} \underset{\substack{U^{(v)} \in R^{n\times k} \\ U^{(w)} \in R^{n\times k}}}{\text{máx}} & Tr\left(U^{(v)^T} L^{(v)} U^{(v)}\right) + Tr\left(U^{(w)^T} L^{(w)} U^{(w)}\right) + \\ & \lambda Tr\left(U^{(v)} U^{(v)^T} U^{(w)} U^{(w)^T}\right) \\ \text{s.t.} & U^{(v)^T} U^{(v)} = I \\ & U^{(w)^T} U^{(w)} = I \end{array} \tag{4}$$

El parámetro λ balancea el objetivo de agrupamiento espectral y el de desacuerdo entre las representaciones. El problema de optimización conjunta puede ser resuelto utilizando maximización alternante con respecto a $U^{(v)}$ y $U^{(w)}$. Para un $U^{(w)}$ dado se obtiene el siguiente problema de optimización para $U^{(v)}$:

$$\begin{array}{ll} \underset{U^{(v)} \in R^{n\times k}}{\text{máx}} & Tr\left(U^{(v)^T}\left(L^{(v)} + \lambda U^{(w)} U^{(w)^T}\right) U^{(v)}\right) \\ \text{s.t.} & U^{(w)^T} U^{(w)} = I \end{array} \tag{5}$$

Lo que resulta en un algoritmo de clustering tradicional con la matriz laplaciana modificada $L^{(v)} + \lambda U^{(w)} U^{(w)^T}$

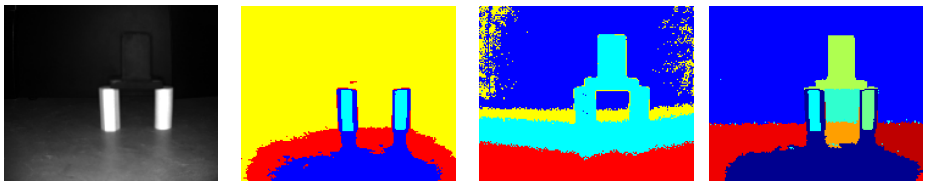
3. Método propuesto

Sea I una imagen de amplitud y R una imagen de rango de dimensión $n \times m$, ambas de la misma escena.

1. A partir de I y R se obtienen las matrices laplacianas aproximadas $\hat{L_I}$ y $\hat{L_R}$ según lo descripto en (2).
2. Sea $\hat{V_I}$ los autovectores aproximados de $\hat{L_I}$ calculados de acuerdo a (3)
3. Se obtienen $\hat{V_R}$, los autovectores de la matriz laplaciana modificada $\hat{L_R} + \lambda \hat{V_I} \hat{V_I}^T$ (5) utilizando el método (2)
4. Se obtiene V_I, los autovectores de la matriz laplaciana modificada $\hat{L_I} + \lambda \hat{V_R} \hat{V_R}^T$ (5) utilizando el método (2)
5. $V = [V_I \quad V_R]$
6. Se aplica un algoritmo de agrupamiento sobre V
7. Se utiliza el criterio propuesto en [13] para evaluar el rendimiento del algoritmo de segmentación. Si la performance mejora ir a 3 sino terminar.

4. Resultados experimentales

El rendimiento del algoritmo de segmentación propuesto fue evaluado sobre 50 imágenes capturadas utilizando la cámara de tiempo de vuelo MESA



(a) (b) (c) (d)

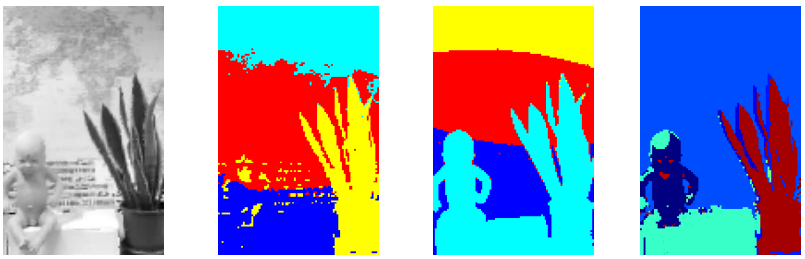
Figura 1. Segmentación aplicada a una imagen capturada con la cámara de tiempo de vuelo SwissRanger SR4000. (a) Imagen de amplitud de una escena real. (b) Método [11] aplicado sobre la imagen de amplitud. $H = 0{,}24$ (c) Método [11] aplicado sobre la imagen de Rango. $H = 0{,}18$ (d) Método propuesto utilizando $\lambda = 3$. $H = 0{,}29$.

SwissRanger SR4000 [7] y el conjunto de datos completo facilitado por el Laboratorio de Tecnología Multimedia y Telecomunicaciones de la Universidad de Padua [17]. La cámara de tiempo de vuelo MESA SwissRanger SR4000 proporciona dos imágenes: una imagen de amplitud y una imagen de rango ambas de 144×176 píxeles. El conjunto de datos [17] contiene imágenes capturadas con una cámara de tiempo de vuelo y una cámara RGB tradicional. El rendimiento del algoritmo de segmentación fue evaluado utilizando el criterio propuesto en [13] y [21] que denominamos H.

La función de semejanza utilizada en todos los casos toma cuenta la disposición espacial de los píxeles en la imagen y la diferencia entre sus valores:

$$W(i,j) = e^{-\frac{\|X_{(i)} - X_{(j)}\|_1^2}{sX}} * e^{-\frac{\|F_{(i)} - F_{(j)}\|_1^2}{sY}}$$

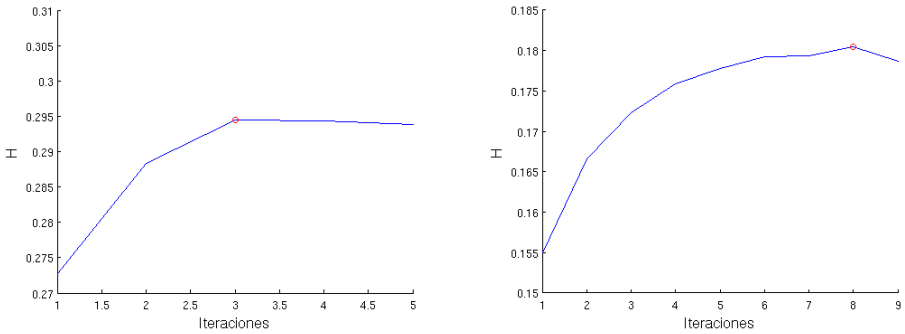
donde $X(i)$ es la ubicación espacial del píxel i, $F(i)$ es el valor del píxel i-ésimo de la imagen, $sX = E(\|X_{(i)} - X_{(j)}\|_1^2) + \frac{3}{4}\sigma(\|X_{(i)} - X_{(j)}\|_1^2)$ de los píxeles dentro del conjunto A y $sY = E(\|F_{(i)} - F_{(j)}\|_1^2)$ de los píxeles dentro del conjunto A. La figura 1 presenta resultados experimentales del método propuesto aplicado a una imagen obtenida con la cámara de tiempo de vuelo MESA SwissRanger SR4000. La imagen de amplitud de la escena capturada 1(a) presenta 3 objetos sobre un fondo negro, todos a la misma distancia. Uno de los objetos tiene un nivel de intensidad similar al del fondo, lo que dificulta su segmentación. Al estar ubicados a la misma distancia poseerán valores de rango similares todos los objetos del frente de la escena. La figura 1(b) y 1(c) presentan el resultado de aplicar el método [11] a la imagen de amplitud y a la imagen de rango respectivamente. La figura 1(d) muestra el resultado de aplicar el método propuesto en el punto óptimo de operación. El método combina correctamente la información de ambas imágenes ruidosas para segmentar los objetos presentes en la escena. La figura 3(a) muestra el rendimiento evaluado en cada iteración del algoritmo. La figura 2 muestra el resultado de aplicar el algoritmo propuesto sobre una escena del conjunto de datos de la Universidad de Padua. La figura 2(a) muestra la imagen de amplitud de la escena. La figura 2(b) y 2(c) muestran el resultado de aplicar el algoritmo [11] a la imagen de amplitud y de rango. Por separado ambas imágenes



(a) (b) (c) (d)

Figura 2. Segmentación de una imagen del conjunto de datos de la Universidad de Padua. (a) Imagen de amplitud de una escena real (b) Método [11] aplicado sobre la imagen de amplitud. $H = 0{,}11$ (c) Método [11] aplicado sobre la imagen de Rango. $H = 0{,}13$ (d) Método propuesto utilizando $\lambda = 3$. $H = 0{,}18$.

no proveen la información necesaria para extraer todos los objetos de la escena. El método propuesto logra mediante la co-regularización extraer la información útil de ambas imágenes, maximizando el rendimiento de la segmentación como se puede ver en la figura 2(d). El rendimiento evaluado en cada iteración del algoritmo es mostrado en la figura 3(b).



(a) (b)

Figura 3. Rendimiento con respecto al número de iteraciones

5. Conclusiones

En este artículo presentamos un método de agrupamiento aplicado a la segmentación de imágenes capturadas con cámaras de tiempo vuelo. Los resultados obtenidos tanto sobre imágenes de intensidad y rango presentan resultados preliminares satisfactorios. El algoritmo combina adecuadamente la información

provista por ambas imágenes incluso en presencia de ruido mediante la utilización de técnicas de co-regularización. El rendimiento resultante de utilizar aprendizaje semi-supervisado con respecto a la utilización de la concatenación de características resultó mejor en todos los casos probados. Una etapa futura de este trabajo prevé la incorporación de información de color al algoritmo de segmentación. Otro aspecto importante sería evaluar la conveniencia de utilizar una medida de disparidad alternativa.

Referencias

1. J. Canny, "A computational approach to edge detection," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. PAMI-8, pp. 679–698, Nov 1986.
2. R. Wu, Z. y Leahy, "An optimal graph theoretic approach to data clustering: theory and its application to image segmentation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 15, pp. 1101–1113, Nov 1993.
3. D. Holz, S. Holzer, R. B. Rusu, and S. Behnke, "Real-time plane segmentation using rgb-d cameras," in *RoboCup 2011: Robot Soccer World Cup XV*, pp. 306–317, Springer, 2012.
4. G. M. Hegde and C. Ye, "A recursive planar feature extraction method for 3d range data segmentation," in *Systems, Man, and Cybernetics (SMC), 2011 IEEE International Conference on*, pp. 3119–3124, IEEE, 2011.
5. A. Bleiweiss and M. Werman, "Fusing time-of-flight depth and color for real-time segmentation and tracking," in *Dynamic 3D Imaging*, pp. 58–69, Springer, 2009.
6. G. Danciu, M. Ivanovici, and V. Buzuloiu, "Improved contours for tof cameras based on vicinity logic operations," in *Optimization of Electrical and Electronic Equipment (OPTIM), 2010 12th International Conference on*, pp. 989–992, May 2010.
7. M. Cazorla, D. Viejo, and C. Pomares, "Study of the sr 4000 camera," in *XI Workshop de Agentes FÃsicos*, 2004.
8. N. Blanc, T. Oggier, G. Gruener, J. Weingarten, A. Codourey, and P. Seitz, "Miniaturized smart cameras for 3d-imaging in real-time [mobile robot applications]," in *Sensors, 2004. Proceedings of IEEE*, pp. 471–474 vol.1, Oct 2004.
9. A. A. Dorrington, C. D. Kelly, S. H. McClure, A. D. Payne, and M. J. Cree, "Advantages of 3d time-of-flight range imaging cameras in machine vision applications," in *Proceedings of 16th Electronics New Zealand Conference*, pp. 95–99, 2009.
10. F. Chiabrando, D. Piatti, and F. Rinaudo, "R-4000 tof camera: Further experimental tests and first applications to metric surveys," in *Proceedings of V Symposium on Remote Sensing and Spatial Information Sciences*, pp. 149–154, 2010.
11. C. Fowlkes, S. Belongie, F. Chung, and J. Malik, "Spectral grouping using the nyström method," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, pp. 214–225, Feburary 2004.
12. N. Arcolano, *Approximation of positive semidefinite matrices using the Nyström method*. PhD thesis, Harvard University, 2011.
13. C. Dal Mutto, P. Zanuttigh, and G. Cortelazzo, "Fusion of geometry and color information for scene segmentation," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 6, pp. 505–521, Sept 2012.

14. A. Kumar, A. Saha, and H. Daume, "Co-regularization based semi-supervised domain adaptation," in *Advances in Neural Information Processing Systems*, pp. 478–486, 2010.
15. A. Kumar, P. Rai, and H. Daume, "Co-regularized multi-view spectral clustering," in *Advances in Neural Information Processing Systems 24* (J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, eds.), pp. 1413–1421, Curran Associates, Inc., 2011.
16. H. Borgdorff, E. Tsivtsivadze, R. Verhelst, M. Marzorati, S. Jurriaans, G. F. Ndayisaba, F. H. Schuren, and J. H. van de Wijgert, "Lactobacillus-dominated cervicovaginal microbiota associated with reduced hiv/sti prevalence and genital hiv viral load in african women," Nature Publishing Group, 2014.
17. M. Technology and U. o. P. Telecommunications Laboratory, "Joint color and depth segmentation datasets," July 2014.
18. 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, MIT Press, 2001.
19. J. Shi and J. Malik, "Normalized cuts and image segmentation," in *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on*, pp. 731–737, Jun 1997.
20. L. Lorenti and J. Giacomantone, "Segmentación espectral de imágenes utilizando cámaras de tiempo de vuelo," in *XVIII Congreso Argentino de Ciencias de la Computación*, 2013.
21. 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. 6, pp. 2223–2226 vol.4, 2000.