A Fuzzy Approach to Evaluate Image Segmentation based on Image Complexity

Luis Madrid-Herrera, *Student, IEEE*, Mario I. Chacon-Murguia, *Senior, IEEE*, and Juan A. Ramirez-Quintana

***Abstract*—Image segmentation evaluation may be carried out by comparing segmentation algorithm results with human ground truths. Its correct evaluation in benchmarks is important to promote the development of new and better segmentation algorithms. It may be partially considered a subjective task because each image may have multiple correct solutions. The evaluation is commonly carried out through crisp metrics, and they fail to generalize the subjectivity of the human criteria present in the human ground truths. Therefore, the interpretation and meaning of the metric values may be ambiguous. Thus, this paper presents a new fuzzy evaluation approach that considers the subjectivity of the human criteria by considering image complexity. This approach leads to an adequate evaluation of this subjective task. The proposed approach demonstrates its advantage when is evaluated using the Peng and SIHD meta-metrics achieving a performance of 0.795 and 0.971, respectively, outperforming the PRI, VI, GCE, and BDE metrics.**

***Index Terms*— Fuzzy evaluation, Image complexity, Image segmentation evaluation.**

I. Introducción

La segmentación de imágenes, SI, es una tarea que tiene un gran impacto en el área de procesamiento digital de imágenes y visión por computadora [1]–[4]. La SI ha sido utilizada en una gran variedad de aplicaciones como la navegación automática, análisis de imágenes médicas, análisis de microorganismos, análisis de cultivos, sistemas de información geográfica, entre otras [5]–[8]. En el estado del arte, comúnmente se encuentran algoritmos de SI sin una aplicación específica. Al no tener una aplicación específica, los algoritmos de SI pueden ser diseñados para intentar segmentar las imágenes de la misma manera en que lo harían algunas personas. Por lo que en la SI pueden existir múltiples soluciones correctas para cada imagen y se considera un problema mal definido [9].

Muchos investigadores se han enfocado en proponer algoritmos más eficientes de SI sin una aplicación específica, con la finalidad de que puedan ser utilizados en múltiples aplicaciones. Sin embargo, no se ha realizado mucha investigación para proponer nuevas maneras de evaluar la SI obtenida por dichos algoritmos [10]. Por lo que en muchos casos se recurre a una evaluación directa de la SI, que consiste en medir el desempeño del algoritmo mediante la comparación de sus resultados contra múltiples segmentaciones humanas.

M. I. Chacon. E-mail of corresponding author: mchacon@ieee.org. The authors are with the Visual Perception Systems Lab, Tecnologico Nacional de Mexico/I.T Chihuahua, Chihuahua 31310, Mexico.

En cambio, la evaluación indirecta consiste en evaluar la SI de acuerdo con su desempeño en una aplicación final [11]. El método propuesto en este trabajo se centra en la evaluación directa de los algoritmos de SI.

La evaluación directa de la SI es mediante métricas que miden la similitud entre segmentaciones bajo diferentes esquemas [12]. Uno de los problemas principales de las métricas existentes de evaluación directa en el estado del arte, es que inducen a los investigadores a buscar el desempeño más alto, sin tomar en cuenta que es prácticamente imposible obtener un desempeño perfecto. Esto es debido a la comparación con múltiples *ground truth* (GT) generados por humanos para cada imagen. Por su parte, muchas de las métricas de evaluación carecen de una interpretación objetiva debido a los rangos de trabajo que tienen y que algunas no cuentan con límites superiores [13]–[15]. Además, la técnica para tomar en cuenta la variación del criterio humano al realizar SI no siempre es la más adecuada. Por ejemplo, la variación del criterio humano es tomada en cuenta mediante la obtención de un promedio contra múltiples GT por imagen. Sin embargo, esto implica que se pierda la generalización del criterio humano al segmentar imágenes, ya que la cantidad de humanos que generan los GT puede ser muy pequeña para representar todas esas múltiples soluciones correctas que puede tener una imagen [11]. Por lo cual es necesario generar un nuevo enfoque de evaluación directa de la SI, que generalice las variaciones en los GT, de acuerdo con una aproximación a la cantidad de soluciones correctas que puede tener una imagen. De esta manera, se podría determinar la calidad de una segmentación hecha por un algoritmo en comparación con el criterio humano de una manera más adecuada.

Debido a la problemática y subjetividad que representa la evaluación directa de la SI, este artículo propone un enfoque difuso basado en evaluación directa que incluye dos contribuciones. La primera parte se enfoca en fusificar las métricas de evaluación de SI del estado del arte. Esta fusificación considera la variación del criterio humano, generalizándolo por imagen en función de la complejidad de imagen, CI. La fusificación facilitará la interpretación de cada métrica y permitirá observar el grado de pertenencia respecto a lo que se considera desempeño humano. La segunda contribución, es una agregación de las métricas difusas para saber si la segmentación que se evalúa es correcta o no, de acuerdo con su consistencia al criterio humano. Los resultados obtenidos del método propuesto demuestran que el enfoque difuso de evaluación de la SI es robusto, confiable y fácil de

interpretar.

Es importante mencionar que, en el enfoque propuesto, así como en la mayoría de otros trabajos del estado del arte publicados, se asume que las imágenes son adquiridas con una buena relación señal ruido. Por lo tanto, el posible problema de ruido no se considera en este trabajo.

La estructura de este artículo es de la siguiente manera. La Sección II presenta las métricas más populares del estado del arte para la evaluación de la SI. También, analiza la relación que existe entre la CI y la variación del criterio humano al segmentar imágenes. La Sección III contiene la metodología llevada a cabo en este nuevo enfoque difuso de evaluación de la SI. La sección IV presenta los resultados obtenidos y su discusión. Por último, la sección V presenta las conclusiones.

II. Métricas de Evaluación y Variación del Criterio Humano

De acuerdo con el estado del arte sobre SI, se puede decir que una segmentación correcta es aquella que tiene buena consistencia con segmentaciones humanas [11]. Para medir esta consistencia, se utilizan métricas de evaluación [12]. La mayoría de las métricas de evaluación de SI, utilizan un esquema rígido, considerando que la mejor segmentación para cierta imagen es cuando se obtiene el más alto desempeño en las métricas respecto a múltiples GT. Sin embargo, no se puede lograr ese nivel de desempeño por la variación implícita en las múltiples segmentaciones correctas que puede tener cada imagen. Por lo tanto, utilizar un esquema rígido en las métricas es poco realista para conocer el grado de desempeño alcanzado por algún algoritmo de SI. Por ejemplo, es posible que un algoritmo obtenga resultados de segmentación similares a los de una persona que realizó un GT con alto desempeño, pero si se compara contra las segmentaciones de otros humanos su desempeño no estaría garantizado.

Por lo tanto, en las siguientes secciones, se presentan las métricas para la evaluación de la segmentación utilizadas en el enfoque difuso propuesto, se analiza la variación de las métricas al evaluar segmentaciones humanas y se describe la motivación del trabajo en utilizar la CI para generalizar el rango de variación humana en cada una de las métricas.

A. Métricas para Evaluación de Segmentación de Imágenes

Las métricas utilizadas en este trabajo para la evaluación de la SI son las de la base de datos BSDS500 [12] debido a su popularidad en el estado del arte, las cuales son; *Probabilistic Rand Index* PRI [16], *Variation of Information* VI [13], *Global Consistency Error* GCE [14], y el *Boundary Displacement Error* BDE [15]. Aunque se tiene conocimiento de otras métricas [10][17], estas son poco consideradas en el estado del arte, por lo que no se consideran en este trabajo.

La métrica PRI compara que tan consistentes son las etiquetas de los pixeles correspondientes entre una segmentación S y un GT. Esta métrica toma valores en el rango [0,1], donde, 1 significa que las etiquetas son totalmente consistentes y 0 que carecen de consistencia. La métrica VI mide la distancia entre dos segmentaciones en términos de su promedio de entropía condicional. La métrica GCE mide como una segmentación S puede verse como un refinamiento de un GT o viceversa, donde 0 significa un desempeño perfecto. Por último, la métrica BDE evalúa el promedio del desplazamiento de los pixeles que son considerados como fronteras entre S y GT.

Las cuatro métricas descritas previamente presentan características, donde dependiendo de la aplicación, algunas métricas son mejores que otras para realizar la evaluación de la SI [18]. Sin embargo, cuando son utilizadas por si solas en la evaluación directa presentan algunas desventajas, las cuales son mostradas en la Fig. 1. Como es mostrado en la Fig. 1, una de las desventajas de PRI es que tiene un pequeño rango dinámico, lo que provoca que los valores obtenidos en la evaluación de las segmentaciones sean comúnmente similares [10]. Además, el PRI es conocido por tener una mayor sensibilidad que especificidad, lo que implica que no se obtienen valores de 0 para evaluaciones donde S y GT son máximamente diferentes, y comúnmente se obtienen valores altos aunque se evalúen segmentaciones muy diferentes al GT [16]. Por su parte, la interpretación de la métrica VI carece de significado cuando se realiza una evaluación con múltiples GT [12], como en el caso de la evaluación directa, además, su medición no es acertada cuando se trabaja con imágenes segmentadas que contienen una cantidad pequeña de regiones [19]. Al igual que VI, GCE tampoco realiza una evaluación acertada cuando S o GT tienen una cantidad pequeña de regiones. Además, GCE no castiga la sobre segmentación de las imágenes [10]. Aunado a esto, GCE carece de significado perceptual cuando S contiene una etiqueta diferente para cada pixel o cuando todos los pixeles tienen la misma etiqueta, ya que en estos casos se obtiene un valor de 0, indicando que la segmentación en términos de GCE es perfecta. En cuanto a BDE, una de sus principales desventajas es que tiene sesgo hacia la sobre segmentación y no cuenta con un límite superior.

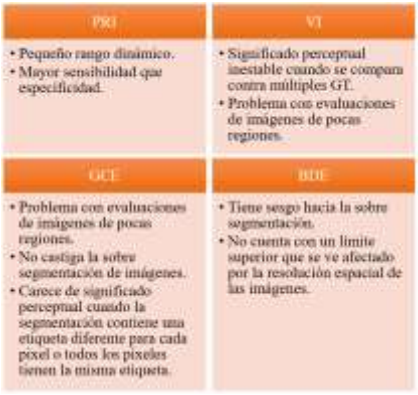

Fig. 1. Principales desventajas de las métricas de evaluación de SI del estado del arte.

De acuerdo con lo anterior, estas métricas presentan diversas situaciones que vuelven complicada la evaluación directa de la SI, debido principalmente a su difícil interpretación por carencia de límites superiores. Por lo cual, es necesario mejorar la manera en la que se lleva a cabo la evaluación directa, facilitando la interpretación de las métricas para que se utilicen de una manera más intuitiva. El enfoque de evaluación directa de SI propuesto mediante la fusificación de las métricas, resuelve el problema de interpretación y mediante la agregación

de las métricas difusas, se combinan todas esas características interesantes que tienen las métricas para determinar de una manera certera e intuitiva el desempeño de los algoritmos de SI.

B. La Complejidad de Imagen y la Variación en el Criterio Humano al Segmentar Imágenes

A continuación, se analiza la variación que existe entre las segmentaciones realizadas por humanos en términos de las métricas $M = \{\text{PRI, VI, GCE, BDE}\}$ en función de la CI. La CI, denotada por Ψ, se utiliza en este trabajo ya que se considera como una medida relativa y subjetiva de la cantidad de información semántica que puede ser percibida mediante el sentido de la vista en una imagen, para un conjunto determinado de imágenes [20]. De acuerdo con [20], Ψ es definida mediante la siguiente ecuación:

$$\Psi_i = \Theta\left(\Xi(I_i), \Xi\left(\{I_{\ddot{\imath}}\}_{\ddot{\imath}=1}^{NI}\right)\right) \qquad \ddot{\imath} \neq i \tag{1}$$

donde I representa una imagen digital, i e $\ddot{\imath}$ representan el índice de una imagen I, Ξ es la cantidad de información semántica percibida mediante el sentido de la vista en una imagen, NI es el número de imágenes bajo análisis y Θ es una relación dependiente del criterio humano para comparar Ξ de las imágenes.

De esta manera, Ψ es utilizada con la finalidad de intentar especificar el rango de desempeño humano en la SI y así poder utilizar este rango para determinar cuándo un algoritmo cumple correctamente con la tarea de segmentación. Dicho análisis se presenta a continuación.

El análisis realizado consiste en dos puntos principales. El primero se enfoca en observar la variación en cada una de las métricas obtenidas de las segmentaciones realizadas por los humanos para cada imagen. La variación humana en cada una de las métricas para cierta imagen i es representada por V_{Mi}. El segundo punto estudia cómo V_{Mi} varía en función de Ψ. Ya que de acuerdo con la definición de Ψ, V_{Mi} puede ser directamente relacionada con la cantidad de interpretaciones que puede tener una imagen. Por ejemplo, una imagen con baja Ψ tiende a tener menor cantidad de interpretaciones (soluciones correctas a la tarea de segmentación) que una imagen con alta Ψ. En la Fig. 2a se muestra una imagen con baja Ψ, así como algunos de sus GTs generados por humanos para esta imagen. Se puede inferir que para esta imagen existe una baja cantidad de interpretaciones o soluciones correctas. Por otro lado, en el caso de la imagen en la Fig. 2b con alta Ψ, el número de soluciones correctas se incrementa considerablemente, desde realizar una segmentación por conjuntos de objetos hasta el caso de asignar una región diferente a cada uno de los objetos de la imagen.

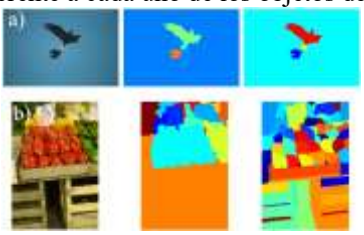

Fig. 2. a) Imagen de baja complejidad con sus respectivos GT, b) Imagen de alta complejidad con sus respectivos GT.

Para el análisis de V_{Mi} en función de Ψ se utilizó la base de datos BSDS500 que cuenta con 500 imágenes y con K número de GTs por cada imagen I_i generados por humanos. El promedio de K en BSDS500 es de 5 GTs por imagen. V_{Mi} fue obtenido para cada I_i de BSDS500 mediante un promedio del valor obtenido para cada M entre todos los GTs de una imagen, es decir:

$$V_{Mi} = \frac{M\left(\text{GT}_{ik,l}, \text{GT}_{ik,m}\right)}{K_i} \qquad l \neq m \tag{2}$$

donde l y m son índices que representan personas que realizaron un GT y k es el índice del GT para una imagen i. Por lo tanto, V_{Mi} indica la variación de la métrica entre los GTs generados por los humanos para cada imagen i.

Para encontrar la tendencia de V_{Mi} en función de Ψ, los resultados fueron ordenados de menor a mayor Ψ y luego procesados con una función de suavizado para observar mejor la tendencia de los datos. Este resultado se muestra en la Fig. 3 para las cuatro métricas en cuestión.

En la Fig. 3 se pueden observar 2 tipos de comportamiento de V_{Mi} en función de Ψ. Las métricas que indican un mejor desempeño entre mayor sea su valor, son denotadas por el conjunto $M\uparrow = \{\text{PRI}\}$ y las métricas de error cuyo desempeño es mejor entre más bajo sea su valor, son denotadas por el conjunto $M\downarrow = \{\text{VI, GCE, BDE}\}$. De acuerdo con Ψ, V_{PRI} decae conforme Ψ incrementa. Para V_{VI}, V_{GCE}, y V_{BDE}, el valor aumenta conforme Ψ incrementa. Esto debido a que a mayor Ψ, mayor número de interpretaciones tiene la imagen segmentada, y por lo tanto se genera una mayor variación en los GT que impacta en el decaimiento del valor de la métrica PRI y el aumento del error de las métricas VI, GCE y BDE.

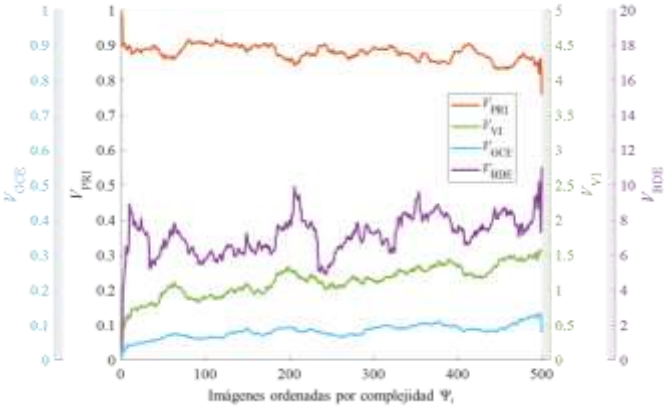

Fig. 3. Variación de las métricas M con respecto a la SI humana en función de Ψ.

Como se observa en los ejes de la Fig. 3, las métricas VI, GCE y BDE no tienen un límite superior, por lo que su interpretación y comparación se vuelve más complicada. Por lo que una de las contribuciones el método propuesto en este trabajo es otorgar un significado a las métricas de evaluación de los algoritmos de SI para que puedan ser interpretadas con mayor facilidad. La interpretación propuesta en este trabajo para las métricas M, radica en analizar y evaluar el grado de consistencia de los valores de las métricas respecto al desempeño humano, en función de la CI. La CI es un parámetro que permite establecer una razón de cambio del número posible

de interpretaciones de segmentación y el rango de variación humana en la generación de GTs. De manera que, a mayor CI, mayor cantidad de interpretaciones, mayor variación humana y por ende más difícil obtener un desempeño alto en las métricas. Por ello, en este trabajo se propone hacer un cambio en la forma en la que se realizan las evaluaciones de los algoritmos de segmentación, para lograr una evaluación más consistente, que sea más estricta con los algoritmos de segmentación cuando se trabaja con imágenes simples que con imágenes complejas. Esto es debido a que en las imágenes simples no se debe permitir tanta variación en las segmentaciones, ya que los GTs generados por humanos son más consistentes que en las imágenes complejas.

III. Metodología del Enfoque Difuso Propuesto

El enfoque difuso que se propone en este trabajo para la evaluación de la SI en función de la CI se realizó utilizando las cuatro métricas populares de BSDS500: PRI, VI, GCE y BDE. Se tiene conocimiento de que existen otras métricas mejor evaluadas en términos de meta-métricas [10][17]. Sin embargo, la finalidad de este trabajo no es proponer nuevas métricas, sino mejorar la interpretación y utilización de las métricas del estado del arte para poder realizar una evaluación más adecuada. Por lo tanto, consideramos que las métricas PRI, VI, GCE y BDE son suficientes para demostrar los beneficios y correcto funcionamiento del enfoque de evaluación propuesto. Sin embargo, el enfoque difuso propuesto es capaz de adoptar otras métricas existentes en la literatura de ser necesario. La Fig. 4 muestra el esquema general del método propuesto, donde, cada métrica M primero se fusifica para obtener μM_{H} y posteriormente las métricas difusas son agregadas para obtener $\mu^{\omega}\mathrm{SEG}_{\mathrm{H}}$ que cuantifica la consistencia de las segmentaciones con el criterio humano.

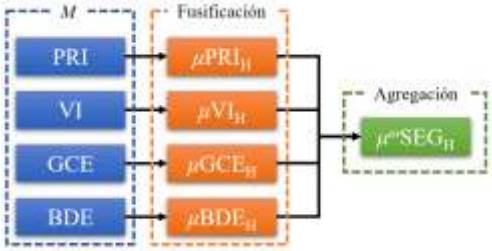

Fig. 4 Esquema general del método propuesto.

A. Fusificación de Métricas de Acuerdo con la Complejidad de Imagen

El primer paso del enfoque de evaluación difuso propuesto se centra en determinar una función de umbral de acuerdo con Ψ y con la variación humana para cada métrica $M = \{\mathrm{PRI}, \mathrm{VI}, \mathrm{GCE}, \mathrm{BDE}\}$. Esta función de umbral representa el valor a partir del cual el valor obtenido de M en una segmentación corresponde a un valor similar al desempeño humano.

Para encontrar esta función de umbral, se realizó un ajuste por medio de regresión lineal en V_{M}, utilizando un filtro de media móvil. Por ejemplo, en la Fig. 5 se muestra el ajuste lineal de V_{PRI} para la métrica PRI. Este ajuste lineal representa la tendencia de V_{PRI} en función de Ψ. Para el punto A mostrado en la Fig. 5, se tiene $\Psi = 0.1890$ y $V_{\mathrm{PRI}} = 0.8977$ y para el punto B, $\Psi = 0.7180$ y $V_{\mathrm{PRI}} = 0.8587$. Por lo que, la función de umbral, VF_M, se obtiene por:

$$VF_M = m_M \Psi + b_M \tag{3}$$

donde, para el caso de la métrica PRI,

$$m_{\mathrm{PRI}} = \frac{V_{\mathrm{PRI}}^{\mathrm{B}} - V_{\mathrm{PRI}}^{\mathrm{A}}}{\Psi^{\mathrm{B}} - \Psi^{\mathrm{A}}} = \frac{0.8587 - 0.8977}{0.7180 - 0.1890} = -0.0737 \tag{4}$$

y

$$b_{\mathrm{PRI}} = V_{\mathrm{PRI}}^{\mathrm{A}} - m_{\mathrm{PRI}} \Psi^{\mathrm{A}} = 0.9116 \tag{5}$$

Así, VF_M se utiliza para determinar una región R_M. A manera general, todos los valores de $M\uparrow$ que superen $VF_{M\uparrow}$ serán considerados dentro de R_M. Para el caso de $M\downarrow$, serían

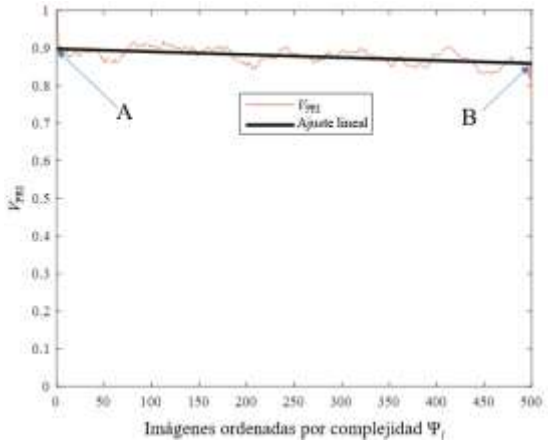

Fig. 5 Ajuste lineal de V_{PRI} en función de Ψ.

todos los valores que queden por debajo de $VF_{M\downarrow}$. Entonces, R_M es una región de valores de M que representa cuando una segmentación es correcta de acuerdo con M, ya que si se encuentra dentro de R_M se tendría un valor igual o mejor al desempeño humano. El procedimiento para la obtención de VF_M para las métricas VI, GCE y BDE, es el mismo que para la métrica PRI pero estas tendrán una pendiente m_M positiva.

Como no siempre se cuenta con suficientes GTs por imagen para generalizar la variación humana, R_M es definida de manera lineal, para así lograr generalizar el comportamiento de V_{Mi} para todas las imágenes utilizando como base a Ψ.

Una vez obtenidas las VF_M y R_M, se fusifica cada métrica M, μM_{H}, con la finalidad de obtener un grado de pertenencia a R_M. Para el enfoque propuesto, se eligieron las funciones S-*Shaped* y Z-*Shaped* debido a su propiedades no lineales y que son comúnmente utilizadas para modelar variables lingüísticas [21]. La función S-*Shaped* es utilizada para fusificar las métricas $M\uparrow$ mientras que Z-*Shaped* es utilizada para las métricas $M\downarrow$. Así, la función $\mu M\uparrow_{\mathrm{H}}$ es representada por:

$$\mu M\uparrow_{\mathrm{H}} = \begin{cases} 0, & M\uparrow < a_M \\ 2\left(\dfrac{M\uparrow - a_M}{VF_{M\uparrow} - a_M}\right)^2, & a_M \le M\uparrow < \dfrac{a_M + VF_{M\uparrow}}{2} \\ 1 - 2\left(\dfrac{M\uparrow - VF_{M\uparrow}}{VF_{M\uparrow} - a_M}\right)^2, & \dfrac{a_M + VF_{M\uparrow}}{2} \le M\uparrow < VF_{M\uparrow} \\ 1, & VF_{M\uparrow} \le M\uparrow \end{cases} \tag{6}$$

donde a_M representa el punto a partir del cual $\mu M\uparrow_{\mathrm{H}} = 0$. Para el

caso de μPRI_H, $a_\text{PRI} = 0$, debido a que se sabe que su límite inferior es 0. En la Fig. 6 se muestra gráficamente la función μPRI_H donde se especifica la localización de los valores $VF_{\text{PRI}i}$ y a_PRI.

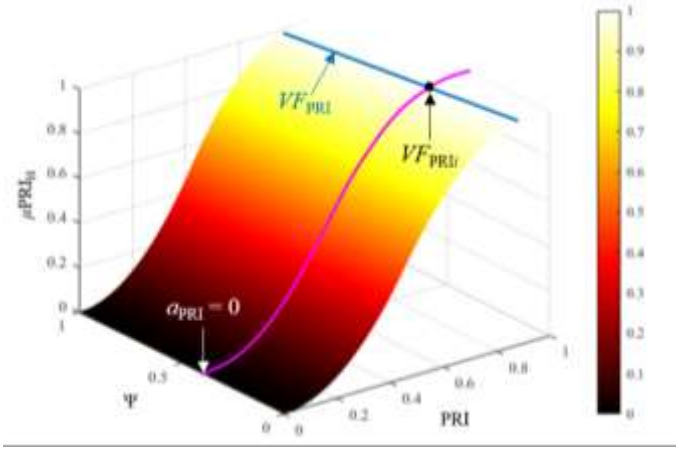

Fig. 6. Representación de la Función μPRI_H, donde se especifica la localización de los valores $VF_{\text{PRI}i}$ y a_PRI.

Las métricas $M\downarrow$ fusificadas son representadas por $\mu M\downarrow_\text{H}$, la cual corresponde a una función Z-*Shaped*. A diferencia de μPRI_H donde $a_\text{PRI} = 0$. Para las métricas VI, GCE y BDE se tiene $a_\text{VI} = 5$, $a_\text{GCE} = 1$, $a_\text{BDE} = 40$. Estos valores fueron definidos experimentalmente, escogiendo valores en los que las métricas ya no representan un desempeño aceptable. Cabe mencionar que estos valores son adecuados para trabajar con las imágenes de BSDS500 que poseen una resolución específica de 321x481 pixeles o 481x321 pixeles dependiendo la relación de aspecto, ya que en algunos casos como BDE su límite superior es afectado de acuerdo con la resolución.

Las métricas M fusificadas μM_H expresan el grado en que una segmentación pertenece al rango de desempeño humano R_M de acuerdo con cada métrica M. Para ejemplificar esto, en la Fig. 7 se muestra gráficamente la función $\mu M\uparrow_\text{H}$. Donde se puede observar la superficie desde una vista superior de la función de membresía S-*Shaped* en función de $M\uparrow$ y Ψ.

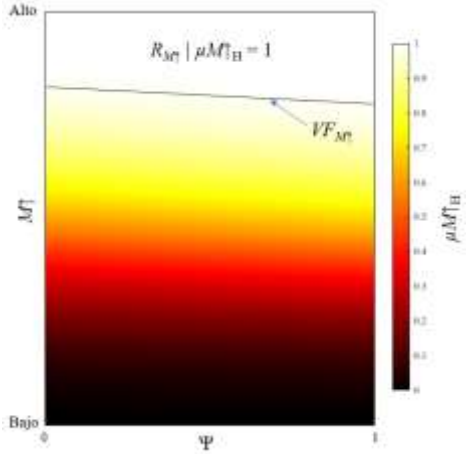

Fig. 7 Representación gráfica de la función $\mu M\uparrow_\text{H}$.

Se puede observar en la Fig. 7 que μM_H es más estricta con los algoritmos de segmentación cuando se trata de imágenes simples que de imágenes complejas. Esto es debido a que de acuerdo con $VF_{M\uparrow}$, existe una menor cantidad de interpretaciones en imágenes con baja Ψ y es esperado que exista una menor variación respecto a los GTs, y conforme incrementa Ψ se tiene que ser menos exigente, ya que incluso entre humanos existe una mayor variación. Por lo tanto, mediante este enfoque de evaluación es posible determinar que tanto una segmentación realizada por un algoritmo de SI se acerca o se encuentra dentro del desempeño humano de acuerdo con cada métrica. Además, este enfoque difuso facilita la interpretación de las métricas, por ejemplo, en el caso de BDE una segmentación de BDE = 30 carece de significado si no se tiene conocimiento del límite superior o de la resolución de las imágenes y no se puede saber si es una segmentación buena o mala. Por lo que, el enfoque propuesto, toma en cuenta estas variables y las utiliza para poder obtener un resultado con mayor significado, ya que μBDE_H: $S\rightarrow[0,1]$, siendo 1 el valor máximo de pertenencia al desempeño humano.

B. Agregación de las Métricas Difusas μM_H

La siguiente contribución presentada en este artículo es un esfuerzo por contestar la pregunta, ¿Qué es una buena segmentación? [11]. Debido a que la respuesta más acertada en el estado del arte menciona que una buena segmentación es aquella que tiene una considerable consistencia con las segmentaciones realizadas por humanos, se trata entonces de cuantificar la consistencia al desempeño humano mediante la agregación de diversas métricas de evaluación. Donde una buena segmentación será aquella en la que en todas las métricas obtengan un desempeño similar al del humano.

Partiendo de la fusificación de las métricas descrita en la sección anterior, es posible realizar su agregación debido a que todas se encuentran dentro del mismo rango, [0,1]. La agregación de métricas se inspira en la métrica *F-Measure* [22], que utiliza una media armónica ponderada. Esta media armónica ponderada fue elegida debido a su sensibilidad con los valores pequeños y por su flexibilidad para otorgar mayor o menor peso en la combinación de métricas, ya que según [18], las métricas tienen diferente contribución dependiendo de la base de datos y los tipos de imágenes. Debido a que el enfoque propuesto no tiene una aplicación específica en cierta base de datos o tipo de imágenes, a todas las métricas se les da el mismo peso, el cual es representado por $\omega_M = 1$.

La función de agregación de M, es definida como:

$$\mu^{\omega}\text{SEG}_\text{H} = \frac{\omega_\text{PRI} + \omega_\text{VI} + \omega_\text{GCE} + \omega_\text{BDE}}{\frac{\omega_\text{PRI}}{\mu\text{PRI}_\text{H}} + \frac{\omega_\text{VI}}{\mu\text{VI}_\text{H}} + \frac{\omega_\text{GCE}}{\mu\text{GCE}_\text{H}} + \frac{\omega_\text{BDE}}{\mu\text{BDE}_\text{H}}} \quad (7)$$

De esta manera, se logra obtener un solo valor en el rango [0,1] que representa el grado de pertenencia, similitud, a una segmentación humana, la cual puede ser considerada como que tan correcta es la segmentación de una imagen.

IV. Resultados

Esta sección presenta los resultados obtenidos mediante el enfoque de evaluación difuso propuesto. Primero, se comparan y analizan las evaluaciones de algoritmos de SI realizadas por medio de M, μM_H y $\mu^{\omega}\text{SEG}_\text{H}$, sección *A*. Luego, la sección *B*

compara a M, μM_H y $\mu^{\omega}SEG_H$ mediante dos meta-métricas. Las meta-métricas son encargadas de medir el desempeño de los procesos de evaluación, respecto al criterio humano.

A. Evaluación de Algoritmos del Estado del Arte Mediante las Diferentes Métricas y Enfoque Propuesto

Para demostrar el correcto funcionamiento del enfoque de evaluación propuesto en este artículo, primero se evaluaron siete algoritmos populares del estado del arte mediante las métricas M para analizar su comportamiento. Luego, los algoritmos fueron evaluados mediante μM_H. Por último, para cada algoritmo se calculó la agregación de las métricas mediante el uso de $\mu^{\omega}SEG_H$, para poder observar cómo es su comportamiento al evaluar la SI.

Los siete algoritmos de SI utilizados fueron SFFCM [3], FRFCM [23], RSSFCA [2], MeanShift [24], MMGR, SLIC-MMGR y LSC-MMGR [1]. Estos algoritmos fueron elegidos de acuerdo con su popularidad y novedad en el estado del arte. Para la evaluación se utilizaron las 500 imágenes de la base de datos BSDS500 y sus respectivos GTs. Para los algoritmos basados en agrupamiento (SFFCM, FRFCM, RSSFCA, MMGR, SLIC-MMGR y LSC-MMGR) se segmentaron todas las imágenes de la base de datos utilizando un número de grupos $2 \leq C \leq 10$, comúnmente utilizado en la literatura, para cada imagen, y con los parámetros descritos por cada autor en los artículos donde se presentan. Luego, para cada métrica se seleccionó el mejor resultado de cada imagen de acuerdo con C. En el caso del algoritmo MeanShift, al no ser un algoritmo basado en agrupamiento se obtuvieron los resultados directamente para cada imagen de acuerdo con los parámetros utilizados en [24]. Los resultados de cada algoritmo de SI se muestran en la

Tabla I y se encuentran ordenados del mejor al peor desempeño en cada métrica.

TABLA I
ALGORITMOS DE SEGMENTACIÓN DE IMÁGENES EVALUADOS POR M Y ORDENADOS DEL MEJOR AL PEOR DESEMPEÑO EN CADA MÉTRICA.

Algoritmo	PRI	Algoritmo	VI
SFFCM	0.802	SFFCM	1.886
MMGR	0.798	MMGR	1.913
LSC-MMGR	0.792	RSSFCA	1.919
SLIC-MMGR	0.792	LSC-MMGR	2.108
MeanShift	0.785	SLIC-MMGR	2.152
RSSFCA	0.782	FRFCM	2.304
FRFCM	0.767	MeanShift	2.731
Algoritmo	GCE	Algoritmo	BDE
RSSFCA	0.034	LSC-MMGR	10.888
MMGR	0.107	SFFCM	11.411
SFFCM	0.133	MMGR	11.681
LSC-MMGR	0.152	SLIC-MMGR	11.702
SLIC-MMGR	0.154	FRFCM	11.754
MeanShift	0.157	MeanShift	12.808
FRFCM	0.194	RSSFCA	12.844

En la

Tabla I se muestra el ranking de los algoritmos para cada métrica. Se observa que cada métrica posiciona de manera distinta a los algoritmos de SI. Por ejemplo, SFFCM es posicionado en primer lugar por la métrica PRI y VI, y para las métricas GCE y BDE ocupan el tercer y segundo lugar respectivamente. Además, aunque se observa algo de consistencia en las posiciones (por ejemplo, SFFCM se mantiene en los primeros lugares y MeanShift en los últimos), también se presenta un caso crítico, donde el algoritmo RSSFCA es posicionado en primer lugar por la métrica GCE, pero en último lugar por la métrica BDE. Por lo tanto, de acuerdo con estas métricas de BSDS500 es complicado tomar una decisión de que algoritmo es mejor, o cual se adapta mejor a la aplicación que se le desee dar al algoritmo.

Una de las desventajas más importantes de utilizar las métricas M para evaluar los algoritmos de segmentación es que carecen de un significado práctico que ayude a interpretar los resultados. Por ejemplo, de la

Tabla I los resultados de GCE = 0.034 por parte de RSSFCA y GCE = 0.194 de FRFCM, no significan que RSSFCA es cerca de 6 veces mejor que FRFCM o que utilizar RSSFCA garantice una buena segmentación. Otro ejemplo similar sucede con la métrica BDE, donde es complicado dar un significado a los valores obtenidos. Se sabe que un valor de cero en la métrica BDE significa una segmentación perfecta, sin embargo, es prácticamente imposible alcanzar ese valor cuando se realiza una comparación con múltiples GTs generados por humanos.

Por lo que, al fusificar las métricas se facilita en gran medida la interpretación de estas, ya que se toma en cuenta el rango de desempeño humano y se limitan sus valores a un rango de [0,1]. En la Tabla II se muestra μM_H para cada algoritmo de segmentación obtenido mediante los resultados de las métricas M en función de la complejidad de imagen Ψ.

Usando el ejemplo anterior, los algoritmos RSSFCA y FRFCM discutidos en la

Tabla I, tienen ahora en la Tabla II los siguientes valores, RSSFCA tiene un $\mu GCE_H = 0.993$, que representa un valor muy cercano al desempeño humano en términos de GCE. Por otro lado, FRFCM obtiene un $\mu GCE_H = 0.947$, disminuyendo cerca de un 5% en desempeño contra RSSFCA, no como en el caso de utilizar M directamente que se observaba una diferencia casi seis veces mayor.

TABLA II
ALGORITMOS DE SEGMENTACIÓN DE IMÁGENES EVALUADOS POR μM_H, ORDENADOS DEL MEJOR AL PEOR DESEMPEÑO

Algoritmo	μPRI_H	Algoritmo	μVI_H
SFFCM	0.965	SFFCM	0.885
SLIC_MMGR	0.965	MMGR	0.877
LSC_MMGR	0.964	RSSFCA	0.867
MMGR	0.964	LSC_MMGR	0.816
RSSFCA	0.957	SLIC_MMGR	0.805
FRFCM	0.946	FRFCM	0.760
MeanShift	0.937	MeanShift	0.611
Algoritmo	μGCE_H	Algoritmo	μBDE_H
RSSFCA	0.993	LSC_MMGR	0.925
MMGR	0.983	SLIC_MMGR	0.917
SFFCM	0.980	SFFCM	0.914
MeanShift	0.977	MMGR	0.908
SLIC_MMGR	0.973	FRFCM	0.905
LSC_MMGR	0.971	RSSFCA	0.895
FRFCM	0.947	MeanShift	0.885

Un punto para destacar cuando se utiliza μM_H, es que el lugar de los algoritmos en el ranking varía en comparación a *M*. Por ejemplo, entre PRI y μPRI$_H$ cinco algoritmos cambiaron su posición, solo SFFCM y LSC-MMGR se mantuvieron en el mismo lugar. Un efecto similar se observa entre GCE y μGCE$_H$, y entre BDE y μBDE$_H$. Esta variación es debida a la inclusión de Ψ para obtener μM_H. Recodando que la variación en el desempeño humano se incrementa conforme Ψ incrementa. Lo cual es modelado mediante la función VF_M.

Para lograr una evaluación con significado, realista y objetiva que facilite la selección de los algoritmos de SI para resolver una tarea específica, en la Tabla III se muestra el ranking generado por μ^{ω}SEG$_H$ de los algoritmos de SI, utilizando ω_M = 1. El valor obtenido por μ^{ω}SEG$_H$ puede ser interpretado como una aproximación a que tan correcta es la segmentación de una imagen considerando el criterio de segmentación humano, ya que este valor se refiere a un grado de pertenencia a la región R_M. Además, μ^{ω}SEG$_H$ se puede extender para combinar más métricas del estado del arte y cada una de ellas puede ser ponderada de acuerdo con la aplicación que realizará el algoritmo de SI.

TABLA III
ALGORITMOS DE SEGMENTACIÓN DE IMÁGENES EVALUADOS POR μ^{ω}SEG$_H$, UTILIZANDO $\omega_M = 1$, ORDENADOS DEL MEJOR AL PEOR DESEMPEÑO.

Algoritmo	μ^{ω}SEG$_H$
SFFCM	0.894
MMGR	0.884
RSSFCA	0.863
LSC_MMGR	0.855
SLIC_MMGR	0.841
FRFCM	0.810
MeanShift	0.756

Es importante resaltar que el ranking generado en la Tabla III coincide con el ranking de la métrica VI mostrado en la

Tabla I. Sin embargo, esto ocurre únicamente para el caso particular de la base de datos BSDS500, no es un comportamiento generalizado. Lo anterior, se abordará en la sección *B* mediante la evaluación de μ^{ω}SEG$_H$ por medio de meta-métricas para demostrar su robustez y confiabilidad sobre las métricas *M* utilizando otra base de datos.

La Fig. 8 muestra un ejemplo de evaluación de dos segmentaciones de una imagen de BSDS500, Fig. 8a, realizadas por un algoritmo con diferentes parámetros para demostrar el uso del enfoque propuesto. La primera segmentación, Fig. 8b, se considera correcta de acuerdo con el criterio humano. Por su parte, la segunda segmentación, Fig. 8c, presenta un error que difícilmente cometería un humano, donde una región uniforme (fondo de la imagen) es fragmentada en dos regiones. Para las segmentaciones de la Fig. 8b y Fig. 8c se muestran los resultados obtenidos por las métricas *M*, las métricas difusas μM_H y la agregación de las métricas difusas μ^{ω}SEG$_H$. Si las segmentaciones de la Fig. 8b y Fig. 8c son comparadas solo utilizando *M*, el resultado de la Fig. 8b sería una mejor segmentación que el de la Fig. 8c. Aunque resultaría complicado interpretar esos valores de *M* para llegar a esa conclusión, y más complicado saber en qué grado es mejor. Esto es debido principalmente a la carencia de límites superiores en VI, GCE y BDE. En cambio, cuando las segmentaciones son comparadas con las métricas difusas μM_H propuestas, es más sencillo determinar que la segmentación de la Fig. 8b es mejor que la de la Fig. 8c y en qué grado es mejor. Esto es debido a que las métricas difusas μM_H propuestas representan un grado de pertenencia al desempeño humano, donde se puede observar que la segmentación de la Fig. 8b toda μM_H = 1 mientras que la segmentación de la Fig. 8c solo obtiene un desempeño aceptable en μVI$_H$ y μGCE$_H$. Estas métricas resultaron altas debido a que μVI$_H$ y μGCE$_H$ no castigan la sobre segmentación presente en el fondo [13][14].

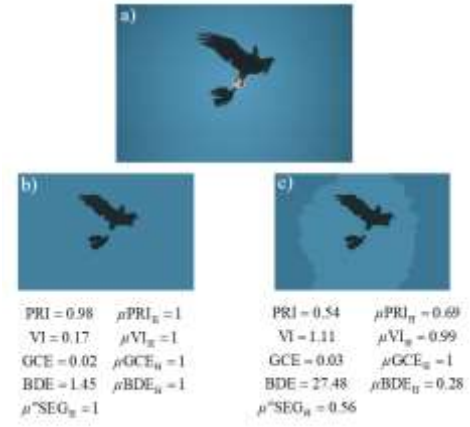

Fig. 8 Ejemplo de evaluación de la segmentación de una imagen. a) Imagen original, b) Segmentación correcta c) Segmentación incorrecta.

Para complementar la evaluación se utiliza la agregación de las métricas difusas μM_H indicada como μ^{ω}SEG$_H$. De manera que μ^{ω}SEG$_H$ representa el grado de pertenencia final con respecto al desempeño humano. μ^{ω}SEG$_H$ determina qué tan coherente es una segmentación considerando el criterio humano. Se puede observar que en la Fig. 8b se obtiene un valor de μ^{ω}SEG$_H$ = 1 mientras que en la Fig. 8c μ^{ω}SEG$_H$ = 0.56. Por lo tanto, se puede determinar fácil e intuitivamente que la segmentación de la Fig. 8b es mejor que la de la Fig. 8c. Lo que indicaría que la segmentación de la Fig. 8b es más similar a la que haría un humano y la de la Fig. 8c sería una segmentación que difícilmente realizaría un humano.

B. Evaluación y Análisis por Medio de Meta-Métricas

Con la finalidad de realizar una evaluación cuantitativa de la mejora obtenida mediante el enfoque propuesto, esta sección muestra los resultados obtenidos por medio de dos meta-métricas. Una propuesta por Peng et al. en [19] y otra llamada SIHD (*Swapped-Image Human Discrimination*) [17]. Las meta-métricas se encargan de evaluar a las métricas *M* respecto al criterio humano bajo una hipótesis aceptada sobre los resultados de una segmentación y analizar qué tan coherente es una medición con dicha hipótesis [17].

1) Meta-métrica Peng

La meta-métrica propuesta por Peng et al. en [19] evalúa que tan coherente es una métrica con el criterio humano al decidir cuál de dos segmentaciones de una imagen es la mejor [10]. Es decir, dada una imagen con sus respectivos GTs y dos segmentaciones de la imagen, se calcula el valor de la métrica

de evaluación de cada segmentación y este valor se compara contra el criterio humano que decidió cuál de las dos segmentaciones era mejor. De esta manera, se obtiene una tasa de exactitud, la cual está dada por la cantidad de veces que la métrica toma la misma decisión que el criterio humano al comparar pares de segmentaciones entre el total de pares comparados.

La base de datos utilizada por Peng et al. [19] consiste en dos partes: Parte_A y Parte_B. La Parte_A fue generada por los autores de [19] conteniendo un total de 200 imágenes y un promedio de 10 GT por imagen, mientras que la parte Parte_B es un extracto de 300 imágenes de la base de datos BSDS500 con un promedio de 5 GT por imagen [12]. Es importante mencionar que los GTs de la Parte_A fueron generados por personas distintas a las que hicieron los GTs de la Parte_B y utilizaron instrucciones diferentes. La

Tabla IV presenta los resultados obtenidos en términos de la tasa de exactitud para $\mu^{\omega}\text{SEG}_{\text{H}}$ y cada métrica *M*. Para la Parte_A el resultado obtenido por $\mu^{\omega}\text{SEG}_{\text{H}}$ supera a los de *M* obteniendo un resultado de 0.795, ya que $\mu^{\omega}\text{SEG}_{\text{H}}$ toma en cuenta la media armónica de todas las métricas fusificadas, lo que ayuda a que se tome una mejor decisión para esta base de datos. Para el caso de la Parte_B, $\mu^{\omega}\text{SEG}_{\text{H}}$ obtiene un resultado de 0.79 el cual solo es superado, aunque no significativamente, por la métrica PRI. En el caso de la Parte_B el resultado de $\mu^{\omega}\text{SEG}_{\text{H}}$ es afectado ya que toma en cuenta a las métricas VI y GCE que obtienen muy bajo resultado. De acuerdo con [19], el resultado obtenido por VI y GCE es muy bajo para la Parte_B, ya que la cantidad de regiones generadas en los GT es baja, por lo que estas métricas, VI y GCE se ven afectadas, y no realizan una selección correcta conforme al criterio humano. Esto demuestra la inestabilidad de VI y GCE ante la manera en que se generan los GTs. Una ventaja de $\mu^{\omega}\text{SEG}_{\text{H}}$ es que se mantiene estable incluso cuando se cambia la manera en la que se generan los GTs lo cual no ocurre con las otras *M*. Otra ventaja del enfoque propuesto es que puede reducir el impacto que tienen las métricas *M*. Por ejemplo, para este caso donde VI y GCE no son buena opción para evaluar la segmentación, se podría reducir el valor de ω_{VI} y ω_{GCE}, lo que permite seguir considerándolas, pero con un menor impacto.

TABLA IV
RESULTADOS OBTENIDOS MEDIANTE LA META-MÉTRICA PROPUESTA POR PENG ET AL. [19].

Parte	$\mu^{\omega}\text{SEG}_{\text{H}}$	PRI	VI	GCE	BDE
Parte_A	**0.795**	0.77	0.745	0.755	0.785
Parte_B	**0.79**	0.803	0.62	0.493	0.783

2) Meta-métrica SIHD

La meta-métrica SIHD [17] se basa en la hipótesis de que dado un par de GTs, una métrica debe de ser capaz de decidir si vienen de la misma imagen con un aceptable refinamiento o de diferentes imágenes con discrepancias inaceptables. Para obtener SIHD, primero se calcula *M* entre todos los pares de GTs seleccionados de la base de datos BSDS500. En este caso se seleccionaron 500 pares de GTs correspondientes a la misma imagen y 500 pares de imágenes diferentes. Luego, para obtener el índice SIHD se utiliza un clasificador que coloca un umbral para separar los dos tipos de pares. SIHD es definido como el porcentaje de clasificaciones correctas de acuerdo con el umbral encontrado por el clasificador. En este caso se utiliza el método de Bayes Risk como clasificador.

En la Fig. 9 se muestran las distribuciones de los pares comparados para el cálculo de SIHD en las métricas *M*. Las gráficas muestran los valores medios de cada distribución y el umbral de decisión encontrado por el clasificador. Se puede observar que los valores SIHD más altos corresponden a la métrica BDE, y VI con 0.938 y 0.936 respectivamente. Para este caso, la métrica PRI es la que obtuvo menor desempeño con un SIHD de 0.87, es decir, que PRI es menos capaz de distinguir cuando dos segmentaciones tienen grandes discrepancias entre sí.

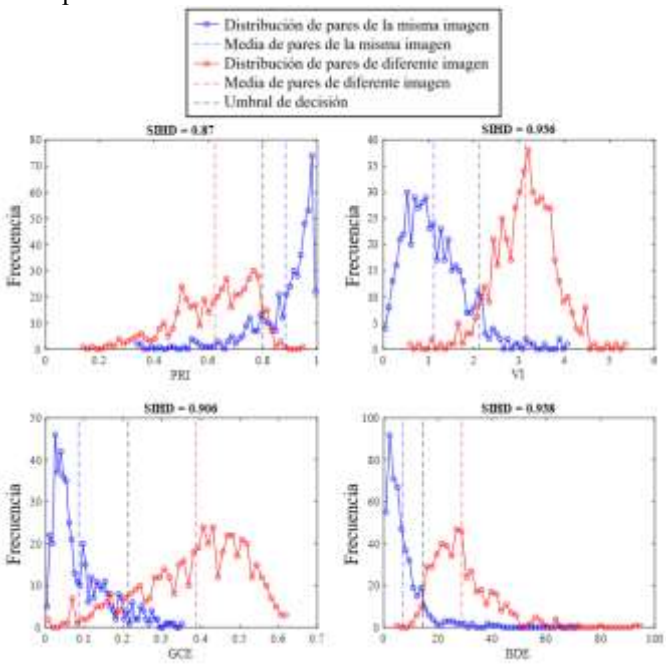

Fig. 9 Gráficas de distribución de pares comparados por medio de *M* para el cálculo de SIHD.

A continuación, en la Fig. 10 se muestra la gráfica de distribución para obtener SIHD utilizando el enfoque propuesto $\mu^{\omega}\text{SEG}_{\text{H}}$, el cual obtiene un valor de SIHD = 0.971, superando considerablemente el valor obtenido por las métricas *M*. En esta gráfica se observa un comportamiento muy diferente al de *M*, esto es debido a que una gran cantidad de pares provenientes de la misma imagen obtienen un desempeño de 1. Esto es esperado ya que se están comparando GTs de la misma imagen realizados por humanos, por lo tanto, los valores obtenidos se encuentran dentro del rango de variación humano, R_M. Por otro lado, el valor medio de $\mu^{\omega}\text{SEG}_{\text{H}}$ para los pares de GTs que provienen de diferentes imágenes se encuentra cerca de $\mu^{\omega}\text{SEG}_{\text{H}} = 0.6$. Estos resultados indican que el enfoque propuesto es capaz de generar una mejor discriminación que las métricas *M*, sobre si los pares comparados provienen o no de una misma imagen.

Debido a que el enfoque propuesto se basa en realizar una evaluación de la SI considerando diversas métricas, es capaz de realizar una evaluación robusta. Esto es debido a que toma en cuenta diferentes aspectos de las segmentaciones mediante la agregación de métricas, algo a lo que es muy difícil llegar utilizando cada métrica por separado. Al agregar las métricas

difusas μM_H se facilita en gran parte la interpretación, ya que, incorpora conocimiento sobre el rango de trabajo de cada una de las métricas M y el rango de desempeño humano. Además, cuando algunas métricas no son tan eficientes para evaluar cierta base de datos, el enfoque propuesto permite la ponderación de las métricas, pudiendo ser modificadas de acuerdo con el problema que se desee resolver mediante los algoritmos de segmentación imágenes.

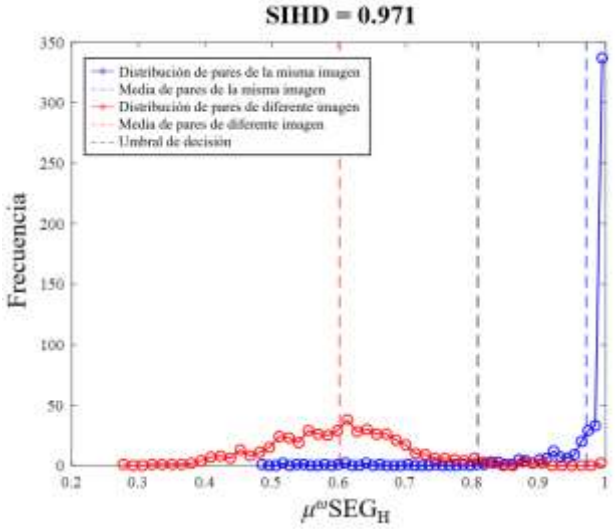

Fig. 10 Gráfica de distribución pares comparados por medio de $\mu^{\omega}SEG_H$ para el cálculo de SIHD.

C. Coste Computacional del Enfoque Difuso Propuesto

Con el fin de complementar los resultados y la contribución del método propuesto se realizó un análisis del coste computacional. El análisis se divide en cuatro etapas: determinación de la complejidad de imagen (Ψ), evaluación de la segmentación por las métricas (M), fusificación de las métricas (μM_H) y agregación ($\mu^{\omega}SEG_H$). Los resultados obtenidos para cada etapa en términos del uso de memoria en KB, tiempo en segundos, FLOPs y porcentaje del uso del CPU, se resumen en la Tabla V. Los resultados mostrados en la Tabla V fueron obtenidos mediante el promedio de procesar cada una de las 500 imágenes de la base de datos BSDS500. El algoritmo se ejecutó en Matlab R2018a en una Laptop Lenovo con procesador Intel i7-10750H @2.60GHz, 16 GB de memoria RAM, sistema operativo Windows 11.

TABLA V
COSTE COMPUTACIONAL PROMEDIO DEL ENFOQUE DIFUSO PROPUESTO PARA UNA IMAGEN DE BSDS500.

Etapa	Uso de memoria	Tiempo	FLOPs	Uso del CPU
Ψ	98412 KB	7.01 s	162	36%
M	8312 KB	0.0894 s	48	4%
μM_H	1660 KB	0.0042 s	0	0%
$\mu^{\omega}SEG_H$	2848 KB	0.0023 s	0	16%
Sumatoria	111232 KB	7.1059 s	210	-

En la Tabla V se puede observar que para evaluar una imagen de BSDS500 el enfoque difuso propuesto se tarda en promedio 7.1 segundos, con un total de uso de memoria de 111,232 KB y 210 FLOPs. Es importante notar que el mayor coste computacional es por parte del cálculo de la complejidad de imagen Ψ, el cual tarda aproximadamente un 98.65% del tiempo total de procesar una imagen. Por lo tanto, en caso de querer optimizar el método de evaluación propuesto habría que comenzar por optimizar la determinación de la complejidad de imagen, ya que el cálculo de M, μM_H y $\mu^{\omega}SEG_H$ es mínimo en comparación al de Ψ. Sin embargo, para fines prácticos de evaluación, este tiempo no es tan significativo, ya que la evaluación no es un proceso que se tenga que realizar en tiempo real, sino es un proceso para ayudar a encontrar cual algoritmo resuelve mejor el problema que se tenga.

V. CONCLUSIONES

Este artículo presentó un nuevo enfoque difuso basado en la CI para evaluar la tarea de segmentación de imágenes. Primero, se fusificaron las métricas más populares del estado del arte para evaluar la SI basándose en la variación del criterio humano. Luego, se propuso una función de agregación de las métricas difusas utilizando la media armónica ponderada. Esto, con la finalidad de obtener una aproximación a que tan consistente es una segmentación con el criterio humano.

El enfoque de evaluación propuesto demostró que otorga un mejor significado e interpretación a las métricas del estado del arte mediante la fusificación. Además, en la agregación de las métricas difusas se permite la ponderación de cada una, en caso de ser conveniente, para poder decidir cuales tendrán mayor contribución en el resultado de evaluación final. También, el enfoque de evaluación propuesto resuelve el problema de límites superiores en algunas métricas causado por falta de estandarización en sus rangos o la resolución espacial de las imágenes.

Con la evaluación por medio de meta-métricas se demostró la robustez y confiabilidad del enfoque propuesto para trabajar con distintas bases de datos generadas por personas e instrucciones diferentes. Esto, debido a que $\mu^{\omega}SEG_H$ obtuvo un desempeño consistente de 0.795 y 0.79 en la meta-métrica propuesta por Peng para la base de datos Parte_A y Parte_B, respectivamente. Por su parte, en la meta-métrica SIHD se obtuvo un desempeño de 0.971 superando considerablemente el valor de las métricas del estado del arte PRI, VI, GCE y BDE. Por lo tanto, el enfoque propuesto establece una mejora considerable respecto a la forma actual de evaluar la SI por medio de métricas, ya que resuelve algunos de sus problemas y ayuda a tomar una mejor decisión sobre que algoritmos son idóneos para resolver algún problema específico, considerando que tan cercanos están sus desempeños al del humano.

AGRADECIMIENTO

This research was supported by the Tecnologico Nacional de Mexico/I.T Chihuahua under grant number 10071.21-P and 14044.22-P .

REFERENCIAS

[1] T. Lei, P. Liu, X. Jia, X. Zhang, H. Meng, and A. K. Nandi, "Automatic Fuzzy Clustering Framework for Image Segmentation," *IEEE Trans. Fuzzy Syst.*, vol. 28, no. 9, pp. 2078–2092, 2020, doi:

10.1109/tfuzz.2019.2930030.
[2] X. Jia, T. Lei, X. Du, S. Liu, H. Meng, and A. K. Nandi, "Robust Self-Sparse Fuzzy Clustering for Image Segmentation," *IEEE Access*, vol. 4, no. 8, pp. 1–14, 2020, doi: 10.1109/ACCESS.2020.3015270.
[3] T. Lei, X. Jia, Y. Zhang, S. Liu, H. Meng, and A. K. Nandi, "Superpixel-Based Fast Fuzzy C-Means Clustering for Color Image Segmentation," *IEEE Trans. Fuzzy Syst.*, vol. 27, no. 9, pp. 1753–1766, 2019, doi: 10.1109/TFUZZ.2018.2889018.
[4] J. D. H. Resendiz, H. M. M. Castro, and E. T. Leal, "A Comparative Study of Clustering Validation Indices and Maximum Entropy for Sintonization of Automatic Segmentation Techniques," *IEEE Lat. Am. Trans.*, vol. 17, no. 8, pp. 1229–1236, 2019, doi: 10.1109/TLA.2019.8932330.
[5] X. Chen and L. Pan, "A Survey of Graph Cuts/Graph Search Based Medical Image Segmentation," *IEEE Rev. Biomed. Eng.*, vol. 11, no. 1, pp. 112–124, 2018, doi: 10.1109/RBME.2018.2798701.
[6] F. Kulwa., "A State-of-the-Art Survey for Microorganism Image Segmentation Methods and Future Potential," *IEEE Access*, vol. 7, no. 1, pp. 100243–100269, 2019, doi: 10.1109/access.2019.2930111.
[7] S. Poudel and S. W. Lee, "Deep multi-scale attentional features for medical image segmentation," *Appl. Soft Comput.*, vol. 109, p. 107445, 2021, doi: 10.1016/j.asoc.2021.107445.
[8] C. Angela, W. Carolina, and C. Carlos, "Medical Image Segmentation Using the Kohonen Neural Network," *IEEE Lat. Am. Trans.*, vol. 17, no. 2, pp. 297–304, 2019, doi: 10.1109/TLA.2019.8863176.
[9] R. Unnikrishnan, C. Pantofaru, and M. Hebert, "Toward objective evaluation of image segmentation algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 929–944, 2007, doi: 10.1109/TPAMI.2007.1046.
[10] M. Simfukwe, B. Peng, and T. Li, "Fusion of measures for image segmentation evaluation," *Int. J. Comput. Intell. Syst.*, vol. 12, no. 1, pp. 379–386, 2019, doi: 10.2991/ijcis.2018.125905654.
[11] Z. Wang, E. Wang, and Y. Zhu, *Image segmentation evaluation: a survey of methods*, vol. 53, no. 8. Springer Netherlands, 2020. doi: 10.1007/s10462-020-09830-9.
[12] P. Arbeláez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 898–916, 2011, doi: 10.1109/TPAMI.2010.161.
[13] S. S. Chai, K. L. Goh, H. H. Wang, and Y. C. Wang, "Comparative evaluation of interactive segmentation algorithms using one unified user interactive type," *Int. J. Comput. Inf. Syst. Ind. Manag. Appl.*, vol. 11, pp. 142–150, 2019.
[14] T. R. Farshi, J. H. Drake, and E. Özcan, "A multimodal particle swarm optimization-based approach for image segmentation," *Expert Syst. Appl.*, vol. 149, pp. 1–13, 2020, doi: 10.1016/j.eswa.2020.113233.
[15] Z. Khan and J. Yang, "Image segmentation via multi dimensional color transform and consensus based region merging," *Multimed. Tools Appl.*, vol. 78, pp. 31347–31364, 2019.
[16] S. Rovetta, F. Masulli, and A. Cabri, "The 'Probabilistic Rand Index': A Look from Some Different Perspectives," *Neural Approaches to Dyn. Signal Exch.*, vol. 151, no. 1, pp. 95–105, 2019, doi: 10.1007/978-981-13-8950-4_10.
[17] J. Pont-Tuset and F. Marques, "Supervised Evaluation of Image Segmentation and Object Proposal Techniques," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 7, pp. 1465–1478, 2016, doi: 10.1109/TPAMI.2015.2481406.
[18] Y. H. Nai., "Comparison of metrics for the evaluation of medical segmentations using prostate MRI dataset," *Comput. Biol. Med.*, vol. 134, no. 104497, pp. 1–12, 2021, doi: 10.1016/j.compbiomed.2021.104497.
[19] B. Peng, L. Zhang, X. Mou, and M. H. Yang, "Evaluation of Segmentation Quality via Adaptive Composition of Reference Segmentations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 10, pp. 1929–1941, 2017, doi: 10.1109/TPAMI.2016.2622703.
[20] L. Madrid-Herrera, "Algoritmo para determinar la complejidad de imágenes aplicado a segmentación," Tesis de doctorado, Instituto Tecnológico de Chihuahua, México, 2022.
[21] M. Versaci and F. C. Morabito, "Image Edge Detection: A New Approach Based on Fuzzy Entropy and Fuzzy Divergence," *Int. J. Fuzzy Syst.*, vol. 23, no. 4, pp. 918–936, 2021, doi: 10.1007/s40815-020-01030-5.
[22] R. Soleymani, E. Granger, and G. Fumera, "F-measure curves: A tool to visualize classifier performance under imbalance," *Pattern Recognit.*, vol. 100, no. 107146, pp. 1–19, 2020, doi: 10.1016/j.patcog.2019.107146.
[23] T. Lei, X. Jia, Y. Zhang, L. He, H. Meng, and A. K. Nandi, "Significantly Fast and Robust Fuzzy C-Means Clustering Algorithm Based on Morphological Reconstruction and Membership Filtering," *IEEE Trans. Fuzzy Syst.*, vol. 26, no. 5, pp. 3027–3041, 2018, doi: 10.1109/TFUZZ.2018.2796074.
[24] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 5, pp. 603–619, 2002, doi: 10.1109/34.1000236.

Luis Madrid-Herrera obtuvo el grado de Ingeniero en Mecatrónica por el Instituto Tecnológico Superior de Nuevo Casas Grandes (2014) y el grado de Maestro en Ciencias en Ingeniería Electrónica por el Instituto Tecnológico de Chihuahua (2018). Actualmente es estudiante de Doctorado en el Instituto Tecnológico de Chihuahua en el laboratorio de percepción visual. Su investigación es en el área de procesamiento digital de imágenes, aprendizaje de máquina, procesamiento de imágenes, visión por computadora y reconocimiento de patrones.

Mario I. Chacon-Murguia recibió los grados de ingeniería (1982) y maestría (1985) en Ingeniería Electrónica por el Instituto Tecnológico de Chihuahua. Obtuvo el grado de Doctorado (1998) en Ingeniería Electrónica en la Universidad Estatal de Nuevo México, Estados Unidos. Es profesor del Instituto Tecnológico de Chihuahua y director del laboratorio de Percepción Visual. Ha publicado más de 218 trabajos y 3 libros. Ha dirigido 45 proyectos de investigación para la industria, e instituciones gubernamentales. Su investigación actual es en el área de percepción visual y procesamiento de imágenes y señales utilizando inteligencia computacional. Es miembro Senior de la IEEE y miembro del Sistema Nacional de Investigadores de México.

Juan A. Ramirez-Quintana obtuvo los grados de ingeniería (2004), maestría (2007) y doctorado (2014) en Ingeniería Electrónica por el Instituto Tecnológico de Chihuahua, México. Del 2008 al 2011 fue investigador y docente en diferentes universidades. Actualmente trabaja como profesor investigador en el Instituto Tecnológico de Chihuahua. Su investigación es en el área de visión por computadora, procesamiento de señales, percepción visual, inteligencia computacional y sistemas embebidos. Es miembro del Sistema Nacional de Investigadores de México.