

Centro de Aplicaciones de Tecnologías de Avanzada



# **Métodos para el reconocimiento automático de objetos combinando modelos de apariencia y relaciones espaciales y jerárquicas**

Tesis presentada en opción al grado científico de  
Doctor en Ciencias Técnicas

**Autora:**

Ing. ANNETTE MORALES GONZÁLEZ-QUEVEDO

**Tutores:**

Dr. EDEL GARCÍA REYES (CENATAV)  
Dr. LUIS ENRIQUE SUCAR SUCCAR (INAOE, México)



Instituto Superior Politécnico José Antonio Echeverría  
Ciudad de la Habana, 2014



*A mis padres,  
porque cada pedacito de lo que soy  
y hago se los debo a ustedes.*

*A Milton,  
porque quiero compartir contigo  
cada pedacito de lo que hago y lo que soy.*



# Agradecimientos

Existen muchas personas a las que quiero agradecer, algunas por tener una participación directa en esta tesis y otras por influir indirectamente.

Quiero agradecer:

*A mi tutor Edel, por la oportunidad, por la guía, por mostrarme el camino a seguir, por darme ánimos cuando los he necesitado, por contribuir a mi formación profesional.*

*Al Dr. Sucar, por el tiempo dedicado a esta investigación, por sus valiosas ideas, por su guía, por sus hospitalidad en México.*

*A Niusvel y Gago, por la parte de la investigación que compartimos, porque fue una experiencia muy valiosa y son compañeros excelentes.*

*A Eric, por su trabajo aplicando los resultados de esta tesis, por su ayuda, su disposición e independencia.*

*A los Drs. Rafael Bello Pérez y Airl Pérez Suárez por sus observaciones y sugerencias en la predefensa de la tesis, que ayudaron a mejorar la calidad la misma.*

*A mi mamá y a mi papá, por el amor y el apoyo incondicional que siempre me han dado en todas las etapas y decisiones de mi vida, porque sin ellos nada de esto sería posible, por haberme enseñado siempre a ser perseverante y perseguir mis sueños, por darme alas.*

*A Milton, por ser mi otra parte, por su amor, por su paciencia, por ayudarme a vencer mis miedos, por darme fuerza y tranquilidad, por sacarme de mis rincones más oscuros, por hacerme inmensamente feliz todos los días.*

*A Army y Abuela, porque siempre han estado conmigo, soportando mis malos ratos, dándome amor y apoyo.*

*A Heydi, porque de no ser por ella yo no estaría en este camino, por su amistad de siempre, por su apoyo, por aguantar mis peroratas en los momentos de mayor estrés (y de menos estrés también).*

*A Helen y Emma, porque desde que nos conocemos siempre hemos estado juntas (cerca o lejos), para lo que sea, cuando sea y como sea, incondicionalmente.*

*A Kadir, por el tiempo, por el apoyo, por estar.*

*A Noslen, Yoa, Ana, Chang, Ricardo, Danita, Dina, Yeni, Rainer, Airel, Javier, Gabriel, por el día a día.*

*A Shul, por la oportunidad, por su incansable insistencia en que siempre se puede un poco más, por su ejemplo.*

*A Isneri, por su ayuda, por su preocupación casi maternal, por su resistencia, por su ejemplo.*

*A mis compañeros de CENATAV en general, porque todos han sido parte de mi vida en esta etapa y me han ayudado de una forma u otra.*

*A Voro, por su ayuda y hospitalidad en México, por estar pendiente.*

*A mis compañeros del ICID, Gaby, Mily, Deiny, Diani, Zubi, Otniel, Eugenia, Nimia, Lola, Mario, Elo, Felo, Leandro, Alex, porque me acompañaron en mis primeros años de vida laboral, por su ayuda, por lo que aprendí con ellos.*

*A mi familia toda, porque soy muy feliz de que me hayan tocado ustedes, porque todo lo que soy es un reflejo de la crianza, el amor y el apoyo que siempre he recibido.*

*A la Revolución cubana, por abrirme las puertas de todo lo que me he propuesto alcanzar.*

*A todos los que de una forma u otra han formado parte de mi vida.*

ANNETTE MORALES GONZÁLEZ-QUEVEDO.  
30 de octubre de 2014

# SÍNTESIS

La investigación realizada en esta tesis está relacionada con el reconocimiento automático de objetos en imágenes 2D. Se propone una representación jerárquica de la escena, combinando información visual y espacial, mediante pirámides irregulares con atributos en los vértices y en las aristas. De este modo, en los vértices son almacenados rasgos visuales para representar la apariencia de las regiones de las imágenes segmentadas a diferentes escalas. En las aristas se almacenan descriptores que representan las relaciones espaciales y topológicas que existen entre las regiones adyacentes. A partir de esta representación, se proponen dos métodos de reconocimiento de objetos en escenarios simples y un método de reconocimiento de objetos en escenarios complejos. El primer método plantea el reconocimiento de clases de objetos como un problema de correspondencia de grafos, el segundo enfoca el reconocimiento de objetos específicos como un problema de clasificación utilizando el paradigma de bolsa de palabras y el tercero, como un problema de re-etiquetado relajado mediante Campos Aleatorios de Markov. Los experimentos se realizaron sobre bases de datos de comparación de dominio público para mostrar los beneficios de la representación y los métodos propuestos. Los resultados obtenidos en este trabajo de tesis contribuyen a la creación de sistemas propios para la video-protección, la recuperación de imágenes por contenido, etc.





# ÍNDICE

<b>Introducción</b>	<b>1</b>
<b>1. El reconocimiento de objetos desde los puntos de vista cognitivo y computacional</b>	<b>11</b>
1.1. Estudios de la Psicología Cognitiva sobre el reconocimiento visual de objetos	13
1.2. Enfoques computacionales para el reconocimiento de objetos . . . . .	16
1.2.1. Reconocimiento de objetos basado en ventanas . . . . .	20
1.2.2. Reconocimiento de objetos basado en segmentación . . . . .	21
1.2.3. Enfoques que utilizan relaciones espaciales . . . . .	25
1.2.4. Enfoques jerárquicos . . . . .	29
1.2.5. Conclusiones parciales . . . . .	33
<b>2. Representación jerárquica de imágenes combinando apariencia y relaciones espaciales</b>	<b>35</b>
2.1. Pirámides Irregulares . . . . .	37
2.2. Descripción visual de las regiones . . . . .	42
2.2.1. Descripción de regiones usando color y textura . . . . .	42
2.2.2. Descripción de regiones usando rasgos contextuales . . . . .	43
2.2.3. Similitud visual . . . . .	46
2.3. Descripción espacial entre regiones . . . . .	46
2.3.1. Descriptor espacial . . . . .	47
2.3.2. Similitud espacial . . . . .	48
2.4. Costo computacional de la representación . . . . .	49
2.5. Conclusiones parciales . . . . .	51
<b>3. Reconocimiento de objetos en escenarios simples usando relaciones espaciales</b>	<b>53</b>
3.1. Selección de los niveles de la pirámide basada en los bordes de las particiones	55

3.2.	Reconocimiento de objetos en escenarios simples usando correspondencia de grafos . . . . .	57
3.2.1.	Algoritmo de correspondencia de grafos . . . . .	58
3.2.2.	Representación de la forma . . . . .	60
3.2.3.	Similitud global entre subestructuras . . . . .	63
3.2.4.	Complejidad computacional de MATCH-Pyr . . . . .	64
3.3.	Reconocimiento de objetos simples usando el enfoque de bolsa de palabras	65
3.3.1.	Construcción del vocabulario visual . . . . .	66
3.3.2.	Adaptación de la representación visual para trabajar con algoritmos de minería de FAS . . . . .	67
3.3.3.	Creación de matrices de sustitución para los algoritmos de minería de FAS . . . . .	67
3.3.4.	Esquema de clasificación . . . . .	68
3.3.5.	Complejidad computacional de BoFAS-Pyr . . . . .	69
3.4.	Experimentos . . . . .	70
3.4.1.	Experimentos en la colección COIL-100 . . . . .	71
3.4.2.	Experimentos en la colección ETH-80 . . . . .	72
3.4.3.	Evaluación de la relevancia de las relaciones espaciales . . . . .	75
3.4.4.	Análisis de la selección de umbrales y pesos . . . . .	77
3.5.	Consideraciones generales sobre los métodos propuestos . . . . .	80
3.5.1.	Comparación del costo computacional de MATCH-Pyr y BoFAS-Pyr	81
3.5.2.	Aplicabilidad de MATCH-Pyr y BoFAS-Pyr . . . . .	81
3.6.	Conclusiones parciales . . . . .	82
<b>4.</b>	<b>Reconocimiento de objetos en imágenes con escenarios complejos usando relaciones espaciales y jerárquicas</b>	<b>83</b>
4.1.	Campos Aleatorios de Markov . . . . .	85
4.2.	MRFs aplicados a la Visión por Computadora . . . . .	87
4.3.	Re-etiquetado de imágenes usando MRFs jerárquicos . . . . .	88
4.4.	Re-etiquetado y segmentación simultáneos . . . . .	92
4.5.	Experimentos . . . . .	96
4.6.	Complejidad computacional de HMRF-Pyr . . . . .	102
4.7.	Conclusiones parciales . . . . .	103

Referencias bibliográficas	107
Producción científica de la autora sobre el tema de la tesis	127
Glosario de acrónimos	129
Glosario de términos	131
Anexos	135
Anexo 1: Ejemplo de las ventajas que proporciona el descriptor espacial propuesto	137
Anexo 2: Pruebas del algoritmo MATCH-PYR para el reconocimiento de objetos en escenarios reales de videovigilancia . . . . .	139

# INTRODUCCIÓN

La meta dorada de la Visión por Computadora es poder narrar o describir lo que aparece plasmado en las imágenes, de forma tal que una computadora sea capaz de procesar esta información de la misma forma en que lo haría un humano, deseablemente a velocidades superiores. En este caso entra a jugar el Reconocimiento Automático de Objetos (como una parte crucial en el análisis de las escenas), que consiste en identificar instancias de una categoría de objetos en una imagen. Se está haciendo referencia en este trabajo al reconocimiento de objetos tridimensionales en imágenes de dos dimensiones o 2D.

En [Dickinson 09] se define el objetivo del reconocimiento de objetos como la habilidad de: (1) discriminar eficazmente cada objeto específico (identificación) o conjunto de objetos (categorización) del resto de los objetos, materiales, texturas, etc; y (2) realizar esta operación sobre un rango de transformaciones que preserven la identidad de la imagen retinal del objeto en cuestión (ej. posición, tamaño, pose, etc.).

Este tema ha sido objeto de investigación por más de cuatro décadas [Dickinson 09] y ha sido también estudiado extensivamente en psicología, neurociencia computacional y ciencias cognitivas [Riesenhuber 00, DiCarlo 12]. El reconocimiento visual de objetos en sí mismo tiene una gran variedad de aplicaciones potenciales donde coinciden áreas de la Inteligencia Artificial y Recuperación de Información, incluyendo, por ejemplo, la recuperación de imágenes por contenido, minería de videos o identificación de objetos para robots móviles.

El reconocimiento de objetos es una habilidad que los humanos adquieren fácilmente desde la infancia. Con solo observar un objeto por un instante, una persona puede determinar su identidad sin importar sus variaciones de pose, textura, color o deformaciones, incluso ante oclusión. De hecho, los humanos son capaces de hacer generalizaciones a partir de un conjunto reducido de objetos e incluso, reconocer objetos que nunca han visto. No obstante, crear un sistema de visión que se acerque a las capacidades cognitivas de las personas no solo es un problema abierto actualmente, sino que las tasas de eficacia alcanzadas aún están lejos de ser aceptables para problemas reales.

En esta área de investigación los esfuerzos han estado encaminados a desarrollar representaciones y algoritmos que permitan reconocer objetos específicos o clases de objetos en imágenes ante distintas condiciones. En un ámbito limitado de objetos tales como señales de tránsito, huellas digitales y rostros, los avances han sido sustanciales, pero cuando se sale de los dominios específicos y se busca más generalización, los resultados alcanzados aún son insuficientes.

Uno de los problemas principales en la interpretación de imágenes y en el reconocimiento de objetos es que existe una baja variabilidad inter-clases (objetos de distintas clases son muy similares) y una alta variabilidad intra-clase (objetos pertenecientes a la misma clase son muy diferentes entre sí), causada por varios factores como condiciones de iluminación, pose de los objetos, localización de la cámara, oclusiones parciales y fondos no relacionados con el objeto o ruidosos (Ver Figura 1). De hecho, en muchos casos la apariencia puede resultar ambigua cuando el objeto se observa aisladamente, por lo que se hace necesario modelar no solo la categoría de objetos, sino también sus relaciones con el contexto en ocurrencias usuales. De manera muy general, el reconocimiento automático de objetos debe seguir dos pasos fundamentales: (1) la extracción de rasgos visuales de bajo nivel y (2) la interpretación de estos rasgos como conceptos de alto nivel. En la literatura, la falta de correspondencia entre los rasgos de bajo nivel y los conceptos de alto nivel es conocida como *brecha semántica* [Tousch 12]. El vínculo entre rasgos de bajo nivel (i.e. datos numéricos) y los metadatos semánticos, aunque natural para un ser humano, está lejos de ser obvio para una máquina. Aunque la brecha semántica es un problema que aparece en múltiples dominios científicos, su influencia en la Visión por Computadoras es particularmente significativa.

De manera general, el reconocimiento de objetos se puede dividir en dos grandes grupos en cuanto a la forma de reconocimiento. Uno es el reconocimiento de objetos específicos, que busca identificar un objeto en particular (ej. la Torre Eiffel, el rostro del Che, el carro de Pepe) y la otra es la categorización de objetos, que pretende reconocer instancias de una categoría pertenecientes a una misma clase conceptual (ej. edificios, tazas, carros, tomates) [Grauman 11].

Desde el punto de vista computacional se han desarrollado métodos para dar solución a estas dos formas de reconocimiento. Para el caso del reconocimiento de objetos específicos, los métodos más utilizados han sido los basados en representaciones globales de la imagen y los que usan características locales.

Las representaciones globales son aquellas que codifican la información de la imagen como

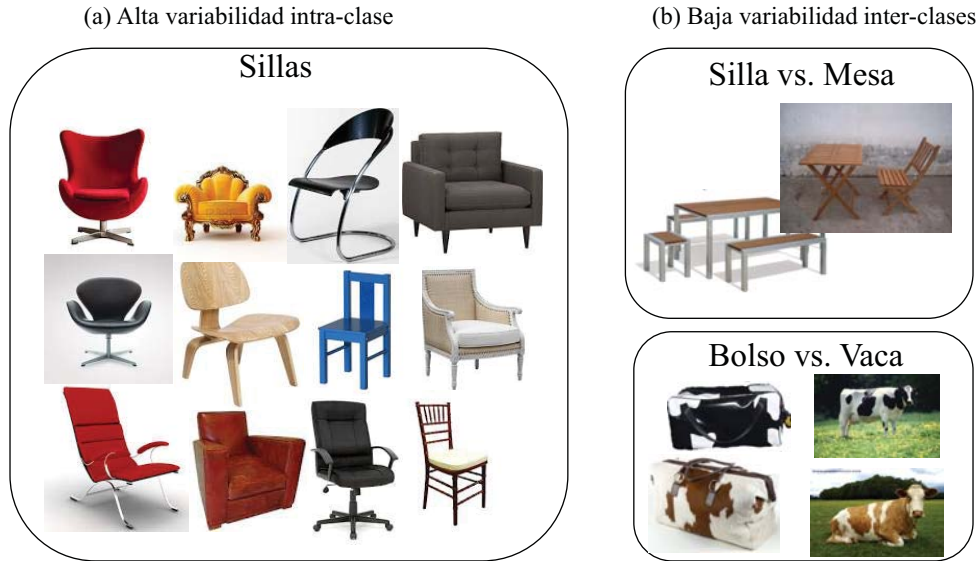


Figura 1: Ejemplos de la dificultad en la clasificación de objetos debido a la alta variabilidad intraclase y la baja variabilidad interclase que se puede encontrar.

un todo (ej. secuencias de todos los píxeles de la imagen, histogramas que capturan la distribución global de características de la imagen, etc.). Este tipo de representación logra capturar estructuras globales de las imágenes, pero falla en la clasificación de conceptos semánticos cuando hay oclusiones, cambios grandes del punto de vista o con objetos deformables. Los enfoques que utilizan características locales incluyen la detección de puntos de interés en la imagen, la descripción de cada punto y finalmente, la búsqueda de coincidencias entre dos conjuntos de puntos que representan a dos imágenes, mediante la correspondencia de estos conjuntos. Estos enfoques se encuentran entre los más exitosos en la actualidad en el reconocimiento de objetos específicos, pero esta representación carece de generalidad para la categorización de objetos.

Para el caso del reconocimiento de clases de objetos, los enfoques se pueden dividir en los basados en ventanas (los cuales hacen una división regular de la imagen, sobre la cual extraen rasgos y realizan operaciones de clasificación) y los basados en partes. Ambos enfoques utilizan descripciones similares a las descritas para el reconocimiento de objetos específicos, pero las aplican de forma local en determinadas áreas de la imagen y utilizan cierta distribución espacial de las mismas. Uno de los enfoques más populares en esta rama es la bolsa de palabras visuales [Csurka 04, Sivic 03], que utiliza un vocabulario visual creado a partir de puntos de interés o de un muestreo denso de parches regulares de la imagen (rejilla regular). Muchos de los métodos más eficaces en reconocimiento de objetos utilizan este tipo de representación [Fei-Fei 04, Everingham 08], no obstante,

la creación de vocabularios visuales óptimos aún es un tema de estudio abierto. Los enfoques basados en partes intentan crear modelos de los objetos basados en sus partes y sus relaciones espaciales, pero usualmente los objetos y sus partes están delimitados por regiones rectangulares (más conocidas por su término en inglés bounding boxes). No obstante, una región rectangular que delimite un objeto no constituye una forma suficientemente efectiva de localización y representación de los mismos para muchas tareas.

Aunque la segmentación de la imagen en regiones irregulares parece ser un proceso más parecido a lo que hacen los humanos al interpretar una escena, la segmentación automática de imágenes basada en apariencia aún presenta problemas para separar eficazmente un objeto del fondo. Es por esto que la mayoría de los enfoques mencionados anteriormente no utilizan segmentaciones irregulares en su proceso de reconocimiento. En los últimos 10 años, desde que Viola y Jones [Viola 04] popularizaron las ventanas deslizantes, la clasificación basada en una representación densa de ventanas ha predominado en las competencias internacionales de detección y clasificación de objetos. Aunque estos métodos son apropiados para el reconocimiento en un contexto amplio (ej. etiquetar una escena en cuanto a si contiene personas o muebles), la ausencia de segmentación no permite la localización de objetos individuales que pudieran apuntar a un contexto más específico. De hecho, los contextos en los cuales los objetos difieren en cuanto a forma y no en apariencia, no pueden ser modelados sin segmentación. Además, cuando la identidad de un objeto es ambigua es importante tener en cuenta las pistas contextuales ofrecidas por objetos cercanos [Dickinson 09, Galleguillos 10].

Por un tiempo, los avances en cuanto al reconocimiento utilizando el paradigma de segmentación disminuyeron, con variaciones menores de las mismas ideas dominantes, y los desempeños se estancaron alrededor del 40 % de eficacia de reconocimiento en la competencia Pascal VOC (*Visual Object Challenge*) del 2011 [Everingham 11] y del 2012 [Everingham 12], una de las competencias internacionales más reconocidas en este ámbito. No obstante, nuevos enfoques e investigaciones han mostrado perspectivas promisorias en esta área, y los algoritmos de detección de objetos más recientes han abandonado las ventanas deslizantes y han incluido la segmentación de la imagen entre los pasos de la detección. De hecho, los tres ganadores de la competencia ILSVRC2013 (ImageNet Large Scale Visual Recognition Challenge 2013) [Russakovsky 14] en la categoría de detección de objetos utilizan segmentación en lugar de ventanas deslizantes para generar posibles localizaciones de los objetos que se buscan.

Con el objetivo de mejorar las representaciones basadas en segmentación, se ha propuesto mayormente la utilización de diferentes segmentaciones para una misma imagen, de

forma que se contrarresten los problemas inherentes a una segmentación única, y realizar las tareas de reconocimiento de objetos sobre o combinando estas representaciones. En [Pantofaru 08] se expone la idea de que la utilización de una jerarquía de segmentaciones proporciona múltiples oportunidades para descubrir los bordes de los objetos y para crear regiones apropiadas para obtener rasgos distintivos.

Otro aspecto a considerar para abordar el problema de la brecha semántica, es cómo tener en cuenta el contexto de cada objeto o región de la imagen. En este sentido, en [Markman 00] se expone que las relaciones estructurales entre los componentes de la imagen juegan un rol fundamental en el proceso de interpretación de las mismas que realizan los humanos. En el área de Visión por Computadora muchos enfoques de reconocimiento de objetos han explotado las relaciones semánticas definidas en [Biederman 72]. Estas son el contexto semántico (la *probabilidad* de co-ocurrencia de los objetos), el contexto espacial (la *posición* relativa entre los objetos) y el contexto de escala (el *tamaño* relativo entre los objetos).

Además de los aspectos antes mencionados, en [Dickinson 09] se resumen algunos principios importantes para el reconocimiento de objetos que emergieron en la década de los '70, de los cuales muchos están siendo descubiertos nuevamente y retomados en la actualidad para abordar el problema de la brecha semántica. Estos son:

1. la importancia de la forma (contornos) en la definición de categorías de objetos;
2. la necesidad de representaciones distribuidas compuestas por partes y sus relaciones, para soportar la articulación de los objetos y para facilitar el reconocimiento de objetos ocultos;
3. la necesidad de representaciones jerárquicas, incluidas las jerarquías parte/todo y las jerarquías de abstracción;
4. la necesidad de estructuras variables (i.e. la cantidad de partes y sus identidades pueden variar entre ejemplares de una misma categoría).

Existen aplicaciones de reconocimiento visual de objetos en las cuales se combinan enfoques de reconocimiento de objetos en escenarios simples (solo hay un objeto de interés en una imagen) y de objetos en escenarios más complejos (como imágenes naturales donde interactúan varios objetos a la vez), para obtener una mayor eficacia en la comprensión de la escena. Por ejemplo, en el seguimiento de objetos en video, una vez aplicado el método de substracción de fondo para la detección de regiones de movimiento, las regiones obtenidas en los primeros fotogramas se consideran objetos simples a los cuales es necesario asignar un concepto semántico. Sin embargo, cuando ocurren problemas de



oclusión entre regiones en movimiento, el algoritmo de detección de movimiento devuelve objetos complejos (formado por subregiones de objetos simples), que es necesario separar en subregiones con sus correspondientes conceptos semánticos. Esto se puede observar en la Figura 2. En este sentido, sería conveniente contar con una representación jerárquica (multiresolución) de las regiones, que permita tanto la clasificación de objetos en escenarios simples como de objetos en escenarios más complejos y que al mismo tiempo permita encontrar la correspondencia entre los objetos reconocidos en el fotograma anterior con los objetos reconocidos en el fotograma actual. Con esto se podría garantizar la consistencia en el seguimiento automático cuando ocurren oclusiones.

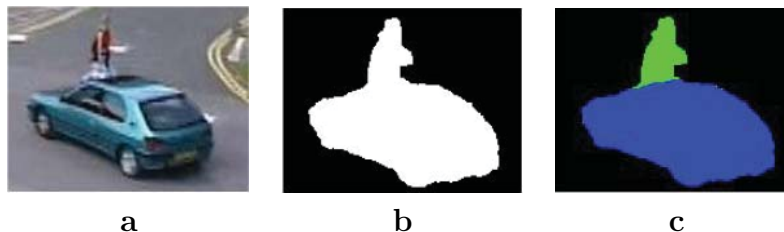


Figura 2: Ejemplo de oclusión en el seguimiento de objetos en video. a) Fotograma original, b) región única en movimiento (luego de detectar regiones en movimiento), c) segmentación y etiquetado de las regiones. (Imágenes tomadas de [Conte 06])

Teniendo en cuenta todos los aspectos descritos anteriormente, se plantea como **problema de investigación** que los métodos de reconocimiento automático de objetos utilizan descripciones de bajo nivel aún insuficientes para realizar con mayor eficacia la vinculación de estas representaciones con sus clases conceptuales.

Para abordar este problema, se establece como **objetivo general** de esta investigación desarrollar algoritmos de reconocimiento de objetos, que incorporen las relaciones espaciales y jerárquicas (multiresolución) presentes en las descripciones de las imágenes, de manera que se logre un aumento de la eficacia.

Para dar cumplimiento a este objetivo se plantean los siguientes **objetivos específicos**:

1. Proponer una representación de las imágenes basada en regiones irregulares, que incorpore información jerárquica y espacial entre las mismas.
2. Desarrollar métodos de reconocimiento de objetos en escenarios simples que utilicen la representación propuesta y que sean más eficaces que los métodos existentes en la literatura.
3. Desarrollar métodos de reconocimiento de objetos en escenarios complejos mediante la representación propuesta y que sean más eficaces que los métodos existentes en la literatura.

Se propone como **hipótesis** de esta investigación que si, sobre la base de una segmentación jerárquica de la imagen, cada región es descrita por sus rasgos visuales y se le incorpora información proveniente de sus relaciones espaciales y jerárquicas con otras regiones, así como el contexto espacial definido por las relaciones padre-hijo en la jerarquía, es posible incrementar la eficacia que se reporta en el estado del arte para los métodos de reconocimiento de objetos.

Para dar cumplimiento a los objetivos y demostrar la hipótesis planteada se definen las siguientes **tareas**:

1. Estudiar el estado actual de los algoritmos computacionales de reconocimiento de objetos, específicamente los basados en segmentaciones irregulares, los que hacen uso de las relaciones espaciales, así como los que utilizan representaciones jerárquicas, para identificar ventajas y desventajas de cada uno al enfrentar el problema de la brecha semántica.
2. Proponer una representación de las imágenes basada en segmentación que permita codificar las relaciones espaciales y jerárquicas de las regiones que la componen.
3. Desarrollar métodos de reconocimiento de objetos en escenarios simples que utilicen la representación propuesta.
4. Desarrollar métodos de reconocimiento de objetos en escenarios complejos que utilicen la representación propuesta.
5. Realizar experimentos que permitan comparar los resultados de los métodos propuestos de reconocimiento de objetos en escenarios simples y objetos en escenarios complejos con otros métodos existentes en la literatura, utilizando bases de datos internacionales.

Los **métodos de investigación** empleados en este trabajo se especifican a continuación. En un primer paso, el método lógico *inductivo-deductivo* fue empleado al realizar un estudio del estado actual de los algoritmos computacionales de reconocimiento de objetos, que permitió obtener un conocimiento general de los problemas que afectan la eficacia de estos y proponer una hipótesis de partida para darle solución a las deficiencias detectadas. El método general *hipotético-deductivo*, guiado por la observación de las problemáticas detectadas, permitió la elaboración de la hipótesis central de la investigación, así como el planteamiento de nuevas líneas de trabajo. Todo esto, apoyado por el método *histórico-lógico* y el *dialéctico*, que se evidencia en el estudio crítico de trabajos anteriores, así como su uso como punto de referencia y comparación de los resultados alcanzados.

Dado que el problema detectado consta de varios sub-problemas a resolver, se utilizó el método *analítico-sintético* para descomponerlo en partes (el problema de la representación de la imagen y el problema de los métodos que vinculen la representación con sus clases conceptuales) de forma que se puedan estudiar dichas partes por separado, buscando soluciones parciales que se sintetizan posteriormente en las soluciones propuestas.

Los métodos matemáticos *algebraicos y aritméticos* permitieron formalizar las expresiones utilizadas en los nuevos métodos de reconocimiento de objetos.

Mediante el método empírico *experimental* se comprueban y fundamentan los estudios comparativos entre distintos métodos de reconocimiento de objetos, apoyado por el método de la *comparación-clasificación* que fue utilizado en el análisis de los resultados experimentales de las propuestas, así como en su comparación con los resultados previos de la literatura. Con la presentación y discusión de los resultados en ámbitos científicos se pone de manifiesto el método empírico *coloquial*.

La principal **novedad científica** de este trabajo radica en una nueva representación para las imágenes, que codifica información visual, espacial y jerárquica, y que fue aplicada en dos escenarios diferentes y para distintas tareas de reconocimiento de objetos, obteniendo resultados superiores a los encontrados en la literatura. La novedad se puede resumir en:

1. Una nueva combinación de atributos visuales y espaciales para una representación de imágenes basada en pirámides irregulares.
2. Un nuevo método de reconocimiento de clases de objetos basado en correspondencia de grafos combinando información visual y espacial, que logra eficacias superiores que otros métodos del estado del arte.
3. Un nuevo método de reconocimiento de objetos específicos basado en bolsa de subgrafos frecuentes, que logra eficacias superiores que otros métodos del estado del arte.
4. Un nuevo método de reconocimiento de objetos basado en re-etiquetado y segmentación simultáneos de imágenes utilizando información espacial y jerárquica, más efectivo que los métodos de etiquetado básicos y que otros métodos del estado del arte.

Los principales resultados teóricos y prácticos que se han obtenido han sido publicados en revistas de impacto referenciadas en la *Web of Science* [Morales-González 13a, Morales-González 14], en conferencias de alto prestigio internacional (referenciadas en Scopus) [Morales-González 10a, Morales-González 12, Acosta-Mendoza 12b, Morales-González 13b], en un taller [Morales-González 10b], en

un reporte técnico [Morales-González 09], y tributaron a los resultados obtenidos en un trabajo de diploma de Licenciatura en Ciencias de la Computación [Hernández-Saura 13].

La **significación práctica** de este trabajo viene dada en primer lugar, por el uso de los métodos propuestos en el desarrollo de sistemas propios que precisen el reconocimiento de objetos en imágenes o videos. Las aplicaciones que utilizan reconocimiento de objetos son muy costosas, además de estar limitadas en muchos casos a dominios específicos. Una de las aplicaciones más interesantes del reconocimiento de objetos para Cuba es en la videovigilancia. Precisamente, por la importancia de este campo para la defensa y el orden interior de cualquier país, gran parte de los sistemas de videovigilancia no aparecen libremente disponibles en el mercado. Para validar la aplicabilidad de los resultados de esta tesis en situaciones reales, se realizaron pruebas en videos provenientes de cámaras de videovigilancia ubicadas en calles de Cuba, lo cual fue el resultado del trabajo de diploma mencionado previamente [Hernández-Saura 13]. Algunos detalles sobre estas pruebas se pueden consultar en el Anexo 2.

Este trabajo se encuentra estructurado en cuatro capítulos. En el primer capítulo se analizan inicialmente algunos estudios de Psicología Cognitiva que muestran descubrimientos o teorías sobre cómo funciona la percepción visual humana. Estas brindan pistas importantes sobre las estrategias a seguir en el desarrollo de métodos computacionales de reconocimiento de objetos, y en especial, de las propuestas hechas en este trabajo. Luego se presenta un estudio de los métodos de reconocimiento de objetos existentes, haciendo especial énfasis en los basados en segmentación, en los que usan relaciones espaciales y los que emplean estructuras jerárquicas, de manera que se puedan comprender las problemáticas existentes y las soluciones que se proponen. En el Capítulo 2 se describe la representación propuesta, se brinda una breve panorámica sobre la teoría de las pirámides irregulares y posteriormente se describe cómo estas serán utilizadas en la representación de las imágenes. En el Capítulo 3 se describen dos nuevos métodos para el reconocimiento de objetos en escenarios simples que utilizan la representación propuesta, explotándola según la tarea que se desea resolver. Ambos métodos se evalúan y comparan con los existentes en la actualidad, mostrando eficacias superiores, y se resalta la utilidad de cada uno para distintas tareas de reconocimiento de objetos. El Capítulo 4 describe un método de reconocimiento de objetos en escenarios complejos, el cual también utiliza la representación propuesta, adaptando su uso para esta nueva situación. Luego el método es extendido para ser tolerante a los problemas de la segmentación inicial. Ambos métodos son evaluados y se pone de manifiesto las ventajas de la combinación de la información visual, espacial y jerárquica al obtener resultados superiores ante otros

métodos del estado del arte. Finalmente se llega a las conclusiones y recomendaciones de la investigación realizada, se listan las referencias bibliográficas utilizadas, se incluyen los glosarios de acrónimos y términos, y los anexos que complementan el trabajo presentado.

# Capítulo 1

El reconocimiento de objetos desde  
los puntos de vista cognitivo y  
computacional



## Capítulo 1

# EL RECONOCIMIENTO DE OBJETOS DESDE LOS PUNTOS DE VISTA COGNITIVO Y COMPUTACIONAL

En este capítulo se ofrece una síntesis sobre los estudios realizados en el campo del reconocimiento visual de objetos, tanto desde el punto de vista cognitivo como el computacional. Los estudios de la Psicología Cognitiva relacionados con la percepción humana han servido en muchos casos como base e inspiración para modelaciones computacionales del problema, y aunque se ha establecido que la máquina no debe necesariamente imitar al humano, algunas investigaciones apoyan que aspectos que se tienen en cuenta en la percepción humana son útiles también en el reconocimiento automático de objetos. Para ilustrar este punto y para proporcionar un marco de referencia que permita interpretar las soluciones dadas a los problemas que se han planteado, se hace una panorámica de los algoritmos de reconocimiento de objetos basados en segmentación, la utilización de relaciones espaciales en los mismos y el empleo de estructuras jerárquicas en el análisis.

### 1.1. Estudios de la Psicología Cognitiva sobre el reconocimiento visual de objetos

Para poder instruir a una computadora sobre cómo reconocer un objeto sería de mucha utilidad lograr comprender cómo funciona el reconocimiento de objetos, es decir, entender cómo un sistema visual (ya sea biológico o sintético) puede tomar una entrada visual y reportar las identidades o categorías de los objetos en la escena.

Aún no se comprende cómo el cerebro construye la representación neuronal que soporta el reconocimiento de objetos y muchas ideas diferentes han sido sugeridas para acercarse a esta cuestión. Los estudios psicológicos han permitido obtener pistas de cómo los humanos



realizan el reconocimiento visual de objetos y la interpretación de las imágenes. Uno de las primeras y más prominentes investigaciones al respecto fue llevada a cabo por Max Wertheimer, dando origen a la teoría *Gestalt* [Wertheimer 23]. *Gestalt* es una palabra alemana que significa forma o silueta y de acuerdo con esta teoría los humanos y los animales imponen cierta organización en sus percepciones sensoriales. En esta teoría se identificaron varios principios organizacionales que aparentemente son usados en el sistema visual humano. A continuación se muestran algunos ejemplos de estos principios de organización perceptual:

1. Proximidad: Patrones formados por grupos de puntos idénticos parecen ser organizados mediante la agrupación de puntos más cercanos.
2. Semejanza: La mente agrupa los elementos similares en una entidad. La semejanza depende de la forma, el tamaño, el color y otros aspectos visuales de los elementos.
3. Cierre: Existe una tendencia a percibir formas cerradas en lugar de contornos abiertos o discontinuos.
4. Continuidad: Los detalles que mantienen un patrón o dirección tienden a agruparse juntos, como parte de un modelo. Es decir, percibir elementos continuos aunque estén interrumpidos entre sí.

Wertheimer defendió la idea de que la organización perceptual es un proceso que ocurre de arriba hacia abajo, y que el agrupamiento es guiado por las características del objeto percibido como un todo.

Gibson presentó una teoría psicológica de la percepción [Gibson 50], basada principalmente en gradientes de textura y flujos visuales y sugirió que la percepción del mundo visual podía ser dividido en la percepción del *mundo espacial* y la percepción de los objetos.

En [Humphreys 89] se realizó un estudio psicológico para analizar el tiempo de respuesta y la efectividad del reconocimiento en imágenes que eran presentadas por corto tiempo a los sujetos investigados. También les eran presentadas pistas visuales o máscaras relacionadas con la imagen mostrada y se encontró que cuando estas pistas eran presentadas antes que la imagen, se interactuaba con el procesamiento visual de bajo nivel, mientras que cuando se presentaban después, se interactuaba con procesos de nivel superior. Posteriormente realizó un estudio neuropsicológico en pacientes que habían sufrido algún tipo de lesión cerebral. Estos experimentos arrojaron conclusiones en el sentido de que los niveles más bajos del sistema visual humano contienen muchos canales separados (o módulos) y que los resultados de estos canales son integrados por los niveles superiores del sistema.

En [Johnson 80] se plantea que para resolver una tarea de reconocimiento, el sujeto debe utilizar cierta representación neuronal interna de la escena visual, para poder tomar una decisión del tipo "¿El objeto A está presente o no en la imagen?". Ellos sugieren que, computacionalmente, el cerebro debe aplicar una función de decisión para dividir un espacio de representación neuronal subyacente en regiones donde el objeto A está presente y en regiones donde no está (una función por cada objeto a reportar potencialmente).

Por otra parte, [Lowe 85] planteó que los humanos pueden detectar inmediatamente relaciones como colinearidad, paralelismo, conectividad y patrones repetitivos entre elementos de una imagen. Aseveró también que la similitud entre dos grupos de objetos no es igual a la suma de las similitudes entre los objetos individuales. En un estudio más reciente que investigó cómo la estructura y la correspondencia influyen la percepción de similitud [Markman 00], concluyeron que las relaciones estructurales entre las componentes de la imagen juegan un rol central en el proceso humano de comparación por similitud.

Las primeras investigaciones en el campo visión por computadora hicieron uso de lo que se conocía sobre óptica geométrica y el proceso de formación de las imágenes, pero no tenían un basamento en estudios psicológicos de la retina y el cerebro. No obstante, en 1963 Roberts desarrolló un sistema que reconocía cubos, cuñas y prismas hexagonales, para lo cual reconoció haberse basado en el trabajo de Gibson (1950) sobre la psicología de la percepción [Roberts 63].

En 1980 Marr y Hildreth propusieron un enfoque que es conocido como el *paradigma de Marr* u *óptica inversa* [Marr 80]. Se realizó un estudio de los procesos matemáticos realizados en los mecanismos biológicos, aunque no se sugirió que un sistema de visión por computadora debiera imitar necesariamente los mecanismos neuronales. En particular, se concluyó que el sistema visual humano detecta bordes y contornos a partir de una versión de la imagen procesada primero con un filtro Gaussiano (suavizada) y luego con un filtro Laplaciano. Marr también planteó que existen distintos niveles de comprensión en la visión y que es necesario entender cada uno de estos niveles.

Algunos estudios más recientes han mostrado que en la visión natural, el proceso preatentivo divide una entrada visual en objetos primitivos, en lugar de en objetos bien definidos [Olson 01]. En [Dickinson 09] se plantean algunos principios basados en un estudio de la literatura especializada, sobre cómo podría el sistema visual ventral crear buenas representaciones. Sugieren que, con respecto al problema global de reconocimiento de objetos, la estrategia computacional no necesita abordar el problema como un todo,

sino ir desenredando la identidad de cada objeto en una serie de pasos sucesivos. De hecho, las neuronas de tipo V1 en una vecindad local solo “ven” el mundo a través de una apertura pequeña (no pueden ver objetos completos), pero pueden realizar operaciones para discriminar sobre la información que reciben como entrada. Las neuronas de tipo V2 pueden hacer lo mismo sobre las entradas que reciben de las V1, y así sucesivamente. Los autores sugieren que los algoritmos de reconocimiento de objetos más fructíferos serán aquellos que puedan ser aplicados local, iterativa y jerárquicamente por un sistema visual (natural o artificial) en cada etapa de procesamiento. En [Tsotsos 88] se muestra que una representación jerárquica interna y un procesamiento jerárquico son los enfoques más confiables para lidiar con las restricciones de espacio y desempeño observados en sistemas humanos. De hecho, Tsotsos concluyó que, además de un procesamiento paralelo del espacio, una organización jerárquica está entre los rasgos más importantes de los sistemas visuales humanos.

## **1.2. Enfoques computacionales para el reconocimiento de objetos**

Desde el punto de vista computacional, son innumerables los intentos realizados por la comunidad científica por dotar a las máquinas de un sistema de percepción parecido al de los humanos. Básicamente, una computadora “percibe” una imagen como una matriz de números y las herramientas matemáticas y algoritmos desarrollados para procesarla no son más que un intento por lograr una interpretación de la imagen como lo haría una persona. Desafortunadamente, es muy poco lo que se conoce sobre cómo el cerebro humano realiza esta tarea y los enfoques desarrollados por la comunidad de Visión por Computadora están lejos de lograr la efectividad deseada.

Uno de los primeros trabajos estrechamente relacionado con la detección y reconocimiento de objetos fue propuesto por [Duygulu 02]. En este trabajo los autores enfocan el problema de reconocimiento como el proceso de asignar palabras a segmentos de la imagen, considerando esta tarea como una traducción de un lenguaje (palabras en inglés) a otro (palabras visuales o regiones).

El reconocimiento de objetos, como rama de investigación, ha tenido un desarrollo vasto, explorándose un sinnúmero de variantes que apuntan a posibles soluciones. La forma en que se aborda esta tarea tiene distintos puntos de vista, como son:

1. La forma de reconocimiento.

2. La forma de analizar la imagen.
3. Los rasgos visuales para describir la imagen.
4. Los rasgos semánticos para describir la imagen.
5. La representación de la imagen.
6. Enfoques de clasificación.

La descomposición de cada uno de estos puntos de vista se muestra en la Figura 1.1 usando una taxonomía para una mejor comprensión. Aquí se pueden observar, de forma muy general, distintos aspectos que se tienen en cuenta a la hora de comenzar a investigar en el problema del reconocimiento de objetos. La secuencia de flechas verticales denota el orden en que cada uno de los pasos debe ser analizado al abordar una tarea de reconocimiento de objetos, y el despliegue de cada paso hacia la derecha indica variantes generales de cada uno, las cuales son a su vez líneas de investigación muy amplias.

Lo primero que se debe determinar es la forma de reconocimiento que se quiere abordar, que puede ser el reconocimiento de instancias de objetos específicos, o categorización de objetos. Posteriormente se debe establecer la forma en que serán analizadas las imágenes. Una variante es el análisis global de las mismas, por ejemplo, la utilización de un vector formado por las intensidades de todos los píxeles de la imagen sobre el que se realizan operaciones de reducción de dimensionalidad y luego se comparan dichos vectores. Otra alternativa de esta modalidad es la utilización de una descripción holística de la imagen basada en la distribución de los valores de los píxeles de la imagen, creando un histograma que puede representar distribuciones de color, de intensidades o de salidas de filtros aplicados a la imagen, entre otras variantes. La mayoría de los métodos que usan estas representaciones se basan en una comparación entre las imágenes enteras o ventanas enteras, lo cual es apropiado para capturar estructuras globales de los objetos pero no funcionan bien ante oclusiones, cambios grandes del punto de vista o con objetos deformables [Grauman 11].

La variante de analizar las imágenes por vecindades locales puede lograrse a través de divisiones regulares del espacio de la imágenes, como son las representaciones basadas en ventanas o parches. Detalles sobre esta variante se brindan en la sección 1.2.1. La opción de analizar la imagen mediante regiones irregulares será abordada en más detalle en la sección 1.2.2.

Una vez que se decida cómo se realizará el análisis de la imagen, se deben seleccionar los rasgos visuales (o de bajo nivel) que serán empleados en la descripción de la imagen

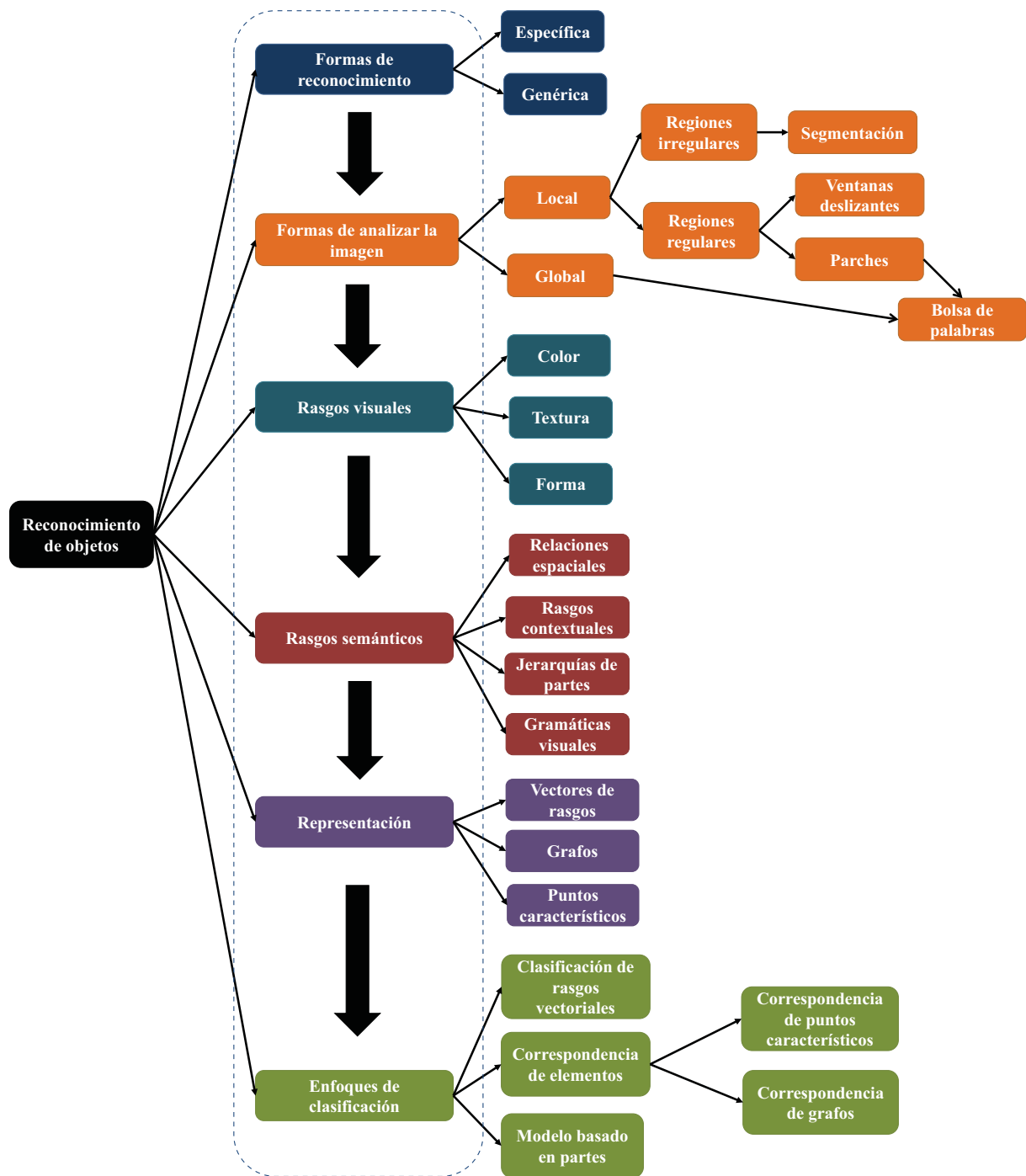


Figura 1.1: Taxonomía de las distintas ramas o atributos que caracterizan el reconocimiento de objetos.

entera o de las regiones. Estos rasgos pueden estar dirigidos a codificar el color, la textura o la forma, según la aplicación y el tipo de objeto con que se trate.

Pudiera decidirse utilizar luego algunos tipos de rasgos de corte más semántico, que

intentan añadir información de más alto nivel a las descripciones visuales básicas. Estos son, las relaciones espaciales entre regiones o entre rasgos visuales y rasgos que incluyan información del contexto, para ayudar a desambiguar la identidad o categoría de los objetos. Las jerarquías de partes y las gramáticas visuales introducen una descomposición de las imágenes en partes y sus relaciones semánticas, como relaciones de herencia, de agregación o descomposición.

Las representaciones de los rasgos y sus relaciones pueden ser variadas, entre ellas se encuentran los vectores de rasgos, la cual es una de las más populares por la posibilidad de disponer de un conjunto muy amplio de herramientas que funcionan en espacios vectoriales. En el caso de los enfoques que utilizan puntos característicos, se pueden generalizar con una secuencia de pasos que incluye la detección de puntos de interés en la imagen, la descripción de cada punto y finalmente la búsqueda de coincidencias entre dos conjuntos de puntos que representan a dos imágenes, mediante la correspondencia de estos conjuntos [Fischler 81, Liao 12]. Algunos enfoques añaden un paso adicional que consiste en la verificación de la consistencia geométrica de los puntos que coincidieron en cada comparación de imágenes. Los descriptores locales más utilizados actualmente son el SIFT [Lowe 04] y el SURF [Bay 08]. Estos enfoques se encuentran entre los más exitosos en la actualidad en el reconocimiento de objetos específicos, pero esta representación carece de generalidad para la categorización de objetos. Las características de las representaciones basadas en grafos se detallan en la sección 1.2.3.

El último paso en un esquema general de reconocimiento de objetos es el enfoque de clasificación que se utilizará. Este puede ser utilizando clasificadores supervisados clásicos (ej. SVM, el clasificador bayesiano simple, *Random Forest*, etc.) [Caruana 08] para las representaciones de rasgos vectoriales. Para el caso que no se usen rasgos vectoriales, se pueden establecer correspondencias entre elementos (ej. correspondencia de grafos o de puntos característicos). Los modelos basados en partes es otro enfoque de clasificación que crea modelos de los objetos basados en sus partes y sus relaciones espaciales. Un ejemplo de esto es el Modelo de Constelación [Fergus 03], que es un modelo conectado completamente y expresa relaciones entre cualquier par de partes. Otras variantes son el Modelo de Estrella y el Modelo de Árbol (entre otros) que exploran otros tipos de conexiones entre las partes. Otra variante en este sentido es el *modelo deformable basado en partes* [Felzenszwalb 13], el cual combina una plantilla global de los objetos, detectada en una escala amplia de la imagen, con un modelo basado en partes en forma de estrella detectado a una escala más pequeña. Estos métodos son mayormente utilizados en la detección de clases de objetos.

Dado el análisis realizado anteriormente de los estudios sobre la percepción humana, en este trabajo se decidió abordar el enfoque de segmentación como forma de analizar la imagen y la utilización de relaciones espaciales y jerárquicas como forma de añadir rasgos semánticos. La selección del enfoque de segmentación viene apoyada por los estudios cognitivos realizados, por ejemplo, la teoría de *Gestalt*, que hace especial énfasis en la forma y contorno de los objetos. Los principios de continuidad, proximidad, semejanza y cierre son prácticamente inherentes al funcionamiento de los algoritmos de segmentación (Ver Sección 1.1). El empleo de la información espacial en una escena ha sido valorado en los estudios de psicología cognitiva mencionados anteriormente, sobre todo para proporcionar estructura y contexto a los objetos, así como desambiguar su identidad. Los enfoques jerárquicos también son adecuados para los principios *Gestalt* tales como proximidad y continuidad [Pizlo 01], además de que su relación con la forma de percepción del sistema visual humano también fue apoyada por otros trabajos descritos en la Sección 1.1, por ejemplo [Humphreys 89, Tsotsos 88, Dickinson 09].

Es por esto que, a falta de espacio y con el ánimo de que el análisis gire en torno a las propuestas presentadas en este trabajo de tesis, se brindará una breve explicación sobre los métodos basados en ventanas, pero la discusión fundamental se centrará en el estudio de trabajos relacionados con enfoques que utilizan segmentación, relaciones espaciales y estructuras jerárquicas.

### **1.2.1. Reconocimiento de objetos basado en ventanas**

Los enfoques basados en ventanas crean ventanas locales de las cuales extraen histogramas de color, intensidades y gradientes de intensidad, entre otros. Una opción es concatenar dichos histogramas, con lo que se conserva cierta distribución espacial de rasgos localmente desordenados [Lazebnik 06, Shahiduzzaman 10]. Otra variante es la detección de objetos mediante ventanas deslizantes, con las cuales se realiza un recorrido exhaustivo en el espacio de la imagen, buscando la porción de la misma que maximice la respuesta del clasificador que se emplee [Viola 04, Feng 11, Zhang 12].

Un enfoque diferente es la extracción de características locales (como SIFT [Lowe 04]), pero en lugar de hacerlo sobre puntos de interés detectados, se hace un muestreo denso de la imagen, que implica crear una rejilla regular donde se extrae un descriptor por cada parche de la misma. Usualmente se utilizan rejillas a distintas escalas. Uno de los enfoques más populares en esta rama es la bolsa de palabras visuales [Csurka 04, Zagoris 11], que utiliza un vocabulario visual creado a partir de la cuantización del espacio de rasgos

locales, y se establece una correspondencia de cada descriptor a un token discreto (palabra visual) que representa el grupo al que pertenece dicho descriptor. El paso siguiente es almacenar la frecuencia de ocurrencia de las palabras visuales que existen en cada imagen utilizando un histograma, con lo cual se traduce un conjunto (usualmente muy grande) de descriptores de alta dimensionalidad a un solo vector disperso de dimensionalidad fija para todas las imágenes. Esto permite el uso de algoritmos de clasificación que asumen por defecto que el espacio de entrada es vectorial. Muchos de los métodos más eficaces en reconocimiento de objetos utilizan este tipo de representación [Fei-Fei 04, Everingham 08], no obstante, la creación de vocabularios visuales óptimos aún es un tema de estudio abierto.

Los modelos basados en partes [Felzenszwalb 13] también utilizan ventanas deslizantes como forma de analizar la imagen, pero imponen restricciones de vecindad a las posibles localizaciones de las partes (representadas por ventanas) de los objetos modelados, por lo que no se busca exhaustivamente en toda la imagen, sino solo en las vecindades en las que las partes pueden aparecer.

Cómo ya se ha mencionado anteriormente, estos métodos basados en regiones regulares son los más populares actualmente y en muchos casos presentan muy buenos resultados, pero el éxito viene dado en gran medida para el reconocimiento de clases de objetos que pueden ser enmarcados en formas rectangulares, (i.e, rostros, carros, peatones). Los resultados para tipos de objetos más generales aún están lejos de ser buenos. Para el caso en que los bounding boxes no cubren bien un objeto, los enfoques basados en ventanas deslizantes tienen más problemas para distinguir entre los objetos en primer plano y los de fondo. Un ejemplo de esto puede ser observado en la Figura 1.2. Además, en estos enfoques quedan prácticamente descartados rasgos discriminantes, como la forma de los objetos, y las relaciones espaciales que se establecen entre ellos también están limitadas a los contornos de los bounding boxes. Por otro lado, la búsqueda exhaustiva en el espacio de la imagen impone restricciones sobre el tipo de procesamiento que puede ser ejecutado en cada ubicación.

### 1.2.2. Reconocimiento de objetos basado en segmentación

El objetivo fundamental de los enfoques de segmentación para el reconocimiento de objetos es la agrupación de información visual, es decir, agrupar píxeles de la imagen en entidades de tamaño creciente y de un significado semántico.



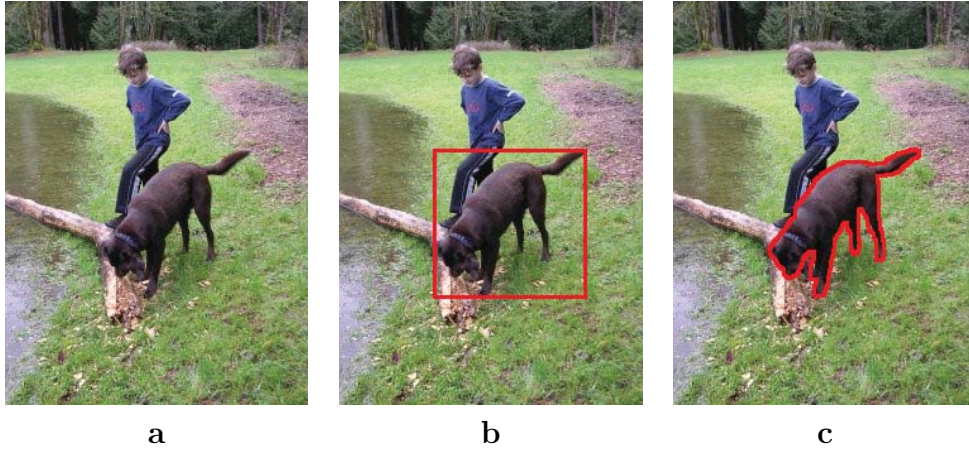


Figura 1.2: Formas de delimitar un objeto en la imagen. a) Imagen original, b) el perro es delimitado por su bounding box, donde se puede notar que se incluyen regiones de fondo, c) el perro es delimitado por un posible resultado de segmentación de la imagen.

La segmentación de una imagen es definida en [González 03] como el proceso que particiona la región total de la imagen  $R$  en  $n$  subregiones,  $R_1, R_2, \dots, R_n$ , tal que:

1.  $\bigcup_{i=1}^n R_i = R$ .
2.  $R_i$  es una región conectada,  $i = 1, 2, \dots, n$ .
3.  $R_i \cap R_j = \emptyset$  para toda  $i$  y  $j$ ,  $i \neq j$ .
4.  $P(R_i) = TRUE$  para  $i = 1, 2, \dots, n$ .
5.  $P(R_i \cup R_j) = FALSE$  para cualquier par de regiones adyacentes  $R_i$  y  $R_j$ .

En esta definición  $P(R_i)$  es un predicado lógico definido sobre los puntos del conjunto  $R_i$  y  $\emptyset$  es el conjunto vacío.  $P(R_i)$  representa las propiedades que deben ser satisfechas por los píxeles de una región segmentada (ej.  $P(R_i) = TRUE$  si todos los píxeles en  $R_i$  tienen el mismo nivel de gris).

En la concepción inicial del reconocimiento de objetos, la segmentación de la imagen era vista como un paso de preprocesamiento necesario [Marr 82]. Con el éxito posterior de los métodos basados en apariencia, los cuales obtenían buenos resultados sin necesidad de una segmentación previa, los métodos basados en segmentación y en apariencia se dividieron en dos áreas. En la actualidad, no obstante, estas áreas están convergiendo nuevamente, debido al creciente conocimiento de que el reconocimiento y la segmentación son procesos entrelazados en el cerebro humano [Peterson 94, Vecera 98] y que los resultados intermedios del reconocimiento pueden ser utilizados para guiar un proceso de segmentación.

Los métodos basados en regiones no son predominantes en el campo de Visión por Computadora. Aunque el reconocimiento de objetos en imágenes basado en regiones puede parecer más similar a como lo hacen los humanos, este enfoque depende en gran medida de la eficacia de los algoritmos de segmentación usados para obtener las regiones. Sin la habilidad de identificar eficazmente las segmentaciones, la eficacia resultante del algoritmo de reconocimiento de objetos será peor que los algoritmos de búsqueda a fuerza bruta (i.e. los algoritmos que escanean exhaustivamente la imagen utilizando ventanas deslizantes). Se ha planteado que la segmentación de nivel medio-bajo no puede determinar dónde un objeto termina y otro comienza, ya que sin el reconocimiento de objetos, es imposible que dicho proceso “sepa” de qué objetos se trata. Esta es la razón por la cual en estos momentos, los métodos basados en bolsas de palabras o en ventanas deslizantes son los paradigmas dominantes en el campo de reconocimiento de objetos.

En [Malisiewicz 07] se plantearon algunas ventajas potenciales de la utilización de enfoques basados en segmentación (asumiendo una segmentación ideal) para el reconocimiento de objetos. Estas son:

1. Se podría reducir la cantidad de ruido en los datos, pues se emplearían solo rasgos de un objeto en cuestión (delimitado por una región irregular) en cualquier método de aprendizaje estadístico. Los rasgos relevantes no estarían contaminados por otros irrelevantes o conflictivos.
2. La forma del objeto también puede ser usada para el reconocimiento.
3. En cuanto a la interpretación de una imagen, la información de los bordes entre objetos es muy útil, no solo para resolver la ambigüedad de un objeto a través de su contexto, sino también para la modelación más eficaz de la estructura de la escena.

Este trabajo ([Malisiewicz 07]) intenta responder dos preguntas básicas en este ámbito: (1) ¿El soporte espacial es importante? O sea, ¿es necesario segmentar los objetos, o los bounding boxes son suficientes para el reconocimiento? (2) ¿Puede la segmentación proporcionar mejor soporte espacial para los objetos? Es decir, incluso si el soporte espacial es importante, ¿puede la segmentación proporcionarlo?

Para responder la primera pregunta, los autores realizan un experimento usando las máscaras de segmentación proporcionadas como ground truth en una colección de imágenes y los bounding boxes de los objetos. Utilizando los mismos rasgos y el mismo esquema de clasificación, los resultados del enfoque usando segmentación mejoraron en 18 de las 21 clases de objetos presentes. De hecho, para objetos en que el cubrimiento de los bounding boxes es pobre (ej, ovejas, bicicletas, aviones) es donde se perciben las mejoras

más notables. Para la segunda pregunta también obtienen una respuesta afirmativa, mostrando que un enfoque que involucre varias segmentaciones produce mejores resultados que la utilización de una única segmentación.

De hecho, el empleo de varias segmentaciones o de una jerarquía de segmentaciones es una de las apuestas actuales para mitigar la desventaja de depender de los resultados de los algoritmos de segmentación que aún son poco eficaces [Pantofaru 08].

Existen varios métodos que emplean la segmentación de la imagen como base para un proceso de reconocimiento de objetos [Duygulu 02, Roth 06, Malisiewicz 07, Su 11], pero en este proceso ignoran la configuración espacial y el contexto de las regiones en la imagen. Notando que este es un factor que puede contribuir a crear métodos más robustos, varios trabajos han explorado el contexto espacial en el proceso de reconocimiento. Estos serán analizados en la Sección 1.2.3.

Otro aspecto importante a tener en cuenta para mitigar las desventajas de los algoritmos de segmentación subyacentes, son los recientes enfoques que intentan entrelazar los procesos de segmentación y reconocimiento, bajo el precepto de que la segmentación sea guiada por pistas semánticas (se puede obtener una segmentación mejor si se sabe qué es lo que se busca [Tegen 14]) y que el reconocimiento se beneficie de una mejor segmentación.

Se han encontrado trabajos en la literatura que emplean de alguna manera este enfoque. Tal es el caso de [Ullman 07], el cual combina el reconocimiento con una segmentación de arriba hacia abajo en una jerarquía. En este trabajo el proceso de segmentación se beneficia del reconocimiento, pero este último no se beneficia de una mejor segmentación. En [Vieux 12] se combinan cuatro algoritmos de segmentación para obtener una mejor partición de la imagen y se refina la clasificación en esta mejor segmentación, empleando información de la clasificación de las segmentaciones iniciales. Algo similar ocurre en [Ion 14], donde no segmentan una imagen completa, sino que generan una bolsa de segmentos independientes que luego son combinados para obtener distintas configuraciones de regiones sobre la imagen. Partiendo de estas se busca la configuración de regiones que maximice la probabilidad del reconocimiento. En estos dos últimos trabajos el reconocimiento se beneficia de una mejor segmentación, pero no se emplea información semántica del reconocimiento para mejorar la segmentación subyacente. En [Torrent 13] se realiza detección y segmentación de objetos, intercambiando información entre ambos procesos durante las distintas rondas de un clasificador boosting, pero siendo un enfoque de detección, solo puede ser aplicado al reconocimiento de un objeto por imagen.

### 1.2.3. Enfoques que utilizan relaciones espaciales

Muchas representaciones utilizadas para el reconocimiento de objetos consisten en un conjunto desordenado de rasgos. Este es el caso del enfoque de bolsa de palabras [Zhang 07] y el núcleo para correspondencia entre pirámides (pyramid match kernel, en inglés) [Grauman 07], los cuales, en términos generales, extraen rasgos de parches de la imagen y se utiliza para la clasificación de la imagen la frecuencia de la aparición de los rasgos modelados en dicha imagen. En estos enfoques, los objetos son usualmente caracterizados en cuanto a apariencia, texturas, formas o partes, desatendiendo la configuración espacial que existe entre ellos.

Los objetos que son creados y utilizados por los humanos en distintos ambientes no ocurren de manera aleatoria, sino que las personas diseñan y organizan espacios de forma que sirvan para distintos propósitos. Algo parecido ocurre en la naturaleza, dentro del aparente desorden, surgen patrones espaciales específicos. Esta organización se expresa en términos de relaciones espaciales. Estas últimas son abstracciones de la configuración de los objetos en el espacio, tales como distancias, direcciones o relaciones topológicas. Esta organización ayuda a los humanos a estructurar y recordar aspectos de su entorno, y de la misma manera, representa un gran potencial a la hora de reconocer objetos automáticamente.

Las relaciones espaciales entre los objetos de una escena han recibido mucha atención en los campos de análisis y recuperación de imágenes, ya que estas pueden revelar información importante sobre la escena que se analiza. Además, se ha afirmado que las relaciones estructurales entre los componentes de la imagen son fundamentales en el proceso humano de comparación por similitud [Markman 00].

De manera general, las relaciones espaciales pueden ser clasificadas en cuatro grandes categorías [Guting 94, Hernández-Gracidas 07]:

1. **Relaciones topológicas.** Son las relaciones que permanecen invariantes ante transformaciones como traslación, escalado y rotación. Ejemplos de estas relaciones son *adyacencia*, *solapamiento*, *inclusión*. Estas relaciones son apropiadas para describir objetos deformables, i.e. objetos articulados en los cuales sus partes pueden variar de posición entre sí, pero mantendrán invarianza topológica en cuánto a su adyacencia (ej. una persona, en la que se puede dar el caso que las manos aparezcan por encima o por debajo del brazo, pero siempre adyacente a él).
2. **Relaciones de dirección (orientación).** Son las relaciones que especifican la ubicación espacial absoluta o relativa de los objetos. Ejemplos de estas relaciones son *arriba de*, *a la derecha de*. Estas relaciones son invariantes al escalado y a

la traslación, pero no a la rotación. Estas relaciones son más útiles para objetos rígidos, donde siempre se encontrarán las partes distribuidas con la misma relación de dirección (ej. las ruedas de un carro siempre estarán debajo de la carrocería).

3. **Relaciones métricas.** Son relaciones que tratan con el tamaño de los objetos o la distancia entre ellos. No son invariantes al escalado, pero sí a la rotación y traslación. Ejemplos de estas relaciones son: *a 2 kms de o 30 metros a la redonda*. Estas relaciones son menos apropiadas para las imágenes 2D, ya que es muy difícil tener métricas de este tipo teniendo en cuenta que no se conoce a priori la perspectiva de la imagen, la escala o si los objetos están en primer o segundo planos.
4. **Relaciones difusas.** Estas relaciones son medidas en términos vagos, y consecuentemente, son difíciles de cuantificar. Ejemplos de ellas son *cerca y lejos*. Este tipo de relaciones tienen también inconvenientes para el uso en 2D, por las mismas razones que las relaciones métricas.

Dentro de este contexto, han surgido muchos enfoques que intentan añadir información espacial a la descripción de las imágenes mediante los rasgos visuales [Lazebnik 06, Hurtut 08] y los puntos de interés. Ellos intentan capturar la distribución espacial de los rasgos visuales en la imagen, pero no llegan a identificar regiones ni las relaciones explícitas entre ellas. Para tratar de resolver esto, se han presentados métodos basados en regiones que tienen en cuenta las relaciones espaciales entre distintas regiones de la imagen.

Existen muchos trabajos de representaciones basadas en regiones que no utilizan la información espacial entre las mismas [Su 11], o lo hacen pobremente [Vieux 10, Yao 10]. También existen métodos que solo utilizan relaciones de dirección [Punitha 06, Morioka 08], o relaciones topológicas [Lin 03, Sjöo 12]. En [Fouquier 12] se combinan relaciones direccionales con relaciones difusas, [Weiss 12, Hedau 12, Hu 13] combinan relaciones topológicas y métricas, [Noma 12, Choi 12] combina relaciones direccionales y métricas, mientras que un número mayor de trabajos combinan relaciones de dirección y topológicas [Hodé 07, Tsapatsoulis 07, Hernández-Gracidas 07, Galleguillos 08, Aydemir 11]. La mayoría de estas representaciones consideran que los objetos están idealmente identificados o trabajan con su bounding box para obtener las relaciones espaciales. No obstante, esto no puede ser aplicado cuando se usa segmentación automática de imágenes, donde los objetos usualmente están sobsegmentados o subsegmentados, o en los casos en que los bounding boxes se solapan.

En [Yao 10] se emplean relaciones difusas, quedando solamente la configuración espacial descrita por la proximidad de los rasgos visuales. La propuesta de [Hedau 12] para escenas

de interiores es calcular valores relacionados con oclusión, inclusión y traslape entre objetos, así como las distancias de los objetos a las paredes, los cuales son empleados como rasgos contextuales en forma de vectores en el proceso de detección de dichos objetos. [Hu 13] propone el empleo de las relaciones *adyacente* y *disjunta* con varias métricas de los objetos que se relacionan (distancias entre bounding boxes, longitud de los mismos, etc) para calcular una medida de similitud espacial que será empleada en un grafo de asociación para hallar correspondencias entre escenas. Para el caso de [Tsapatsoulis 07], las relaciones espaciales están definidas para el contexto de escenas de deporte, donde modelan varias relaciones de orientación, pero solo una relación topológica (*solapamiento*). En [Hernández-Gracidas 07] se utilizan relaciones espaciales para mejorar el etiquetado automático de imágenes, y aunque se utilizan varias relaciones de orientación, de las topológicas solo se utiliza la *adyacencia*. En [Hodé 07] se intenta expresar relaciones espaciales complejas mediante relaciones espaciales elementales, y esta información es usada para validar la consistencia semántica de tareas como la segmentación. Las relaciones son expresadas de forma cualitativa y resulta complicado establecer similitudes entre ellas. En el trabajo presentado en [Galleguillos 08], se cuantizan las relaciones espaciales en 4 relaciones prototipo: *arriba*, *abajo*, *adentro* y *alrededor* y se crean matrices de ocurrencias de estas relaciones entre cada par de categorías. En el contexto de la creación de robots de servicios, en [Aydemir 11] se crea un algoritmo de búsqueda visual de objetos donde se emplean solo dos relaciones: *dentro de* y *arriba de*. En [Felzenszwalb 13], se propone un modelo deformable basado en partes para la detección de objetos, donde utilizan un modelo espacial que refleja el costo de ubicar el centro de una parte en distintos lugares relativos al objeto total. Aunque este enfoque es muy interesante y ampliamente utilizado y extendido, las relaciones espaciales representadas de esta forma son limitadas, además de que restringe el tipo de objeto a ser detectado (objetos articulados). En [Fouquier 12] definen la adyacencia y la proximidad entre objetos empleando una representación difusa, y con esto modelan también relaciones direccionales, que son empleadas en un modelo deformable basado en partes. Su principal desventaja es su especificidad (segmentación de estructuras del cerebro en imágenes 3D de resonancias magnéticas), pues emplean un modelo genérico del cerebro para restringir la búsqueda. En [Weiss 12] se combinan relaciones métricas con relaciones topológicas entre partes de los objetos pero la forma de definir estas relaciones y los objetos es muy limitada al contexto utilizado. Otra combinación es explorada en [Choi 12], donde se utilizan relaciones métricas y direccionales en la creación de un modelo a priori de relaciones espaciales entre objetos. En este caso se utilizan relaciones de escala (distancia entre

los centros de los objetos normalizadas con respecto al tamaño de los objetos) y de las direccionales se utiliza solo la relación vertical entre objetos.

## Representaciones basadas en grafos

Dentro del conjunto de rasgos de bajo nivel desarrollados hasta la fecha, los grafos constituyen una de las representaciones que pueden proporcionar cierta información de alto nivel implícitamente, siendo por tanto una representación promisoría para los investigadores con el fin de encontrar nuevas soluciones. Muchos trabajos han representado las imágenes como grafos con este propósito y han desarrollado métodos para clasificar usando este tipo de estructura de datos. Una preocupación en esta área es que, aunque los grafos son poderosas herramientas de representación, son complejos de utilizar, dando lugar generalmente a algoritmos con alto costo computacional o la simplificación de la estructura de datos, perdiendo de esta forma parte de la información embebida.

Una representación explícita de las relaciones espaciales entre regiones es el grafo de adyacencia de regiones (RAG, por sus siglas en inglés) [Brun 06]. Este define un grafo simple para una partición dada de la imagen, asociando un vértice a cada región y creando una arista entre dos vértices si las regiones que representan son adyacentes. No obstante, la noción de adyacencia por si sola es muy pobre para describir organizaciones espaciales complejas entre las distintas partes de un objeto, y no proporcionan información suficiente para diferenciar una relación de *adyacencia* de una relación de *inclusión* [Brun 06]. En [Pham 10] emplean un grafo dirigido que puede ser visto como un RAG, donde los vértices representan conceptos visuales obtenidos de regiones regulares de la imagen y las aristas representan solo dos relaciones espaciales: *a la derecha de* y *arriba de*. La propuesta de [Noma 12] emplea grafos en el plano de la imagen, con atributos en los vértices y las aristas. Las aristas son vistas como vectores en un espacio 2D y los atributos que recibe cada arista son la longitud y la orientación de dicho vector.

En [Yoon 11] se emplea un grafo para representar una imagen, donde los vértices corresponden a rasgos locales y cada vértice tiene un número fijo  $k$  de aristas enlazadas a sus  $k$  vecinos más cercanos en el espacio de la imagen. De esta forma, intentan incorporar la configuración espacial al modelo de bolsa de palabras. Otra intento en este sentido es la propuesta de [Ren 14], donde crean un grafo a partir de rasgos locales de la imagen. Luego este grafo es particionado en  $K$  grafos y se construyen  $K$  histogramas de palabras visuales (uno por subgrafo) que son concatenados para representar la imagen. En estos modelos, la información espacial que se utiliza es la que viene implícitamente representada

en los grafos, sin distinguir entre los distintos tipos de relaciones espaciales existentes.

Se han desarrollado distintos enfoques para utilizar grafos en tareas de clasificación, por ejemplo, algoritmos de correspondencia de grafos [Duchenne 11, Glantz 04, Saux 05, Yoon 11, Noma 12], los cuales utilizan distancias (por ejemplo, la distancia de edición de grafos), técnicas de correspondencia glotona o núcleos de correspondencia, con el fin de comparar grafos. Otra forma de realizar la clasificación es mediante el uso de métodos de *graph embedding* [Bunke 08, He 09], los cuales, en términos generales, proyectan un grafo en un espacio vectorial y luego emplean un clasificador que trabaje en ese espacio para clasificar los vectores resultantes.

#### 1.2.4. Enfoques jerárquicos

Luego de analizar el avance de los algoritmos de reconocimiento de objetos basados en segmentación, se puede observar lo siguiente:

1. Los objetos pueden aparecer a cualquier escala dentro de una imagen. Además, algunos objetos son contenidos dentro de otros objetos, por lo cual es necesario tener una representación a distintas escalas.
2. No existe una única mejor estrategia para agrupar regiones. Un borde puede representar el contorno de un objeto en una imagen, mientras que en otra puede ser el resultado de un degradado de la intensidad o color debido a la iluminación.

Por tanto, teniendo estos dos aspectos en cuenta, se puede notar que, en lugar de apostar por una única mejor segmentación de la imagen, es importante combinar varias segmentaciones complementarias, es decir, diversificar el conjunto de segmentaciones utilizadas. La forma más natural de generar segmentaciones a distintas escalas es emplear un enfoque de segmentación jerárquico.

En las propuestas que se han hecho de enfoques jerárquicos, están los que utilizan jerarquías de partes, como es el caso de [Ullman 07, Maire 13] y los que usan jerarquías de segmentaciones. Algunos enfoques que generan segmentaciones en forma jerárquica ([Akçay 07, van de Sande 11, Arbelaez 12, Zhang 13]) utilizan las regiones en cada nivel de forma independiente, es decir, no tienen en cuenta las relaciones entre regiones. Es posible encontrar estructuras jerárquicas en forma de árboles [Ullman 07, Todorovic 08, Maire 13], donde se aprovechan las relaciones jerárquicas, pero se descartan las relaciones espaciales entre regiones en el plano de la imagen. Ejemplos donde se emplean tanto las relaciones jerárquicas como las espaciales se pueden ver en [Fischer 04, Yang 11, Russell 14].



Abundando más sobre estas propuestas, se tiene que en [Akçay 07] se emplea un enfoque de segmentación jerárquica para obtener regiones a distintos niveles de segmentación. Estas regiones son agrupadas para crear un algoritmo no supervisado de detección de objetos. Por otro lado, en [van de Sande 11] proponen utilizar una jerarquía de particiones para generar posibles ubicaciones (bounding boxes) de objetos en una imagen a distintas escalas. Los autores proponen un algoritmo de segmentación jerárquica en el cual comienzan a partir de una sobre-segmentación de la imagen producida por un algoritmo del estado del arte y luego proceden a unir de forma glotona las regiones más similares en cada nivel. Las regiones obtenidas les proporcionan bounding boxes de objetos que serán utilizados para entrenar y emplear un detector de objetos en imágenes. En este caso, contrario a lo deseado en la segmentación, se generalizan los límites de cada objeto con un bounding box. Algo similar ocurre en [Arbelaez 12] y en [Zhang 13], donde emplean la salida de detectores de partes para clasificar las regiones segmentadas en la jerarquía. El problema en este caso es que introducen el costo de los detectores basados en ventanas deslizantes para clasificar cada región. En ninguno de estos enfoques se utilizan las relaciones espaciales y de jerarquía.

En [Ullman 07] el autor propone un esquema de reconocimiento de objetos que involucra una jerarquía (árbol) de partes o fragmentos de objetos de clases específicas. La descomposición repetida de un objeto en fragmentos (regiones regulares) es lo que genera la jerarquía, que es utilizada luego en el proceso de reconocimiento de objetos y sus partes. Una representación muy parecida es empleada en [Maire 13], pero aplicada a la creación de una herramienta de anotación visual de imágenes. En [Todorovic 08] se genera una jerarquía de segmentaciones, representada como un árbol, donde lo que se codifica es la relación padre-hijo entre regiones, siendo la cima de esta estructura la imagen completa. La jerarquía en su totalidad se representa como un grafo con atributos, conectado, dirigido y sin ciclos, con lo cual se pierde la noción de niveles de la jerarquía. Se establece una correspondencia entre árboles para encontrar semejanzas entre imágenes, buscando el isomorfismo máximo entre sub-árboles. Siendo estos enfoques basados en árboles, las relaciones espaciales entre las regiones en el plano de la imagen quedan descartadas.

Dentro del contexto de la recuperación de imágenes médicas, se propuso utilizar una jerarquía de RAGs en [Fischer 04]. Ellos emplean una jerarquía de segmentaciones para construir un grafo que represente regiones de la imagen. Este grafo codifica la adyacencia de regiones entre vértices y relaciones jerárquicas entre distintos niveles. No obstante, el método propuesto para construir la jerarquía, y por tanto el grafo subyacente, es demasiado costoso computacionalmente para ser aplicado en imágenes

de propósito general. Otro aspecto que se debe notar es que la contribución de las relaciones espaciales en el esquema de correspondencia de grafos empleado se reduce a relaciones muy simples (adyacencia y jerarquía). En [Yang 11] se utiliza una jerarquía de segmentaciones obtenida a partir del algoritmo mean-shift a distintas escalas. Se crea un Campo Condicional Aleatorio (CRF, por sus siglas en inglés) jerárquico, con el cual se etiqueta cada región segmentada de la jerarquía. El CRF jerárquico se resuelve completo para toda la jerarquía, por lo que no se analiza cada nivel individualmente. El caso de [Russell 14] emplea relaciones espaciales y jerárquicas, pero estas últimas no se establecen sobre una jerarquía de segmentaciones, sino en un Campo Aleatorio de Markov, donde las relaciones jerárquicas se emplean para crear potenciales de mayor orden, con lo cual la optimización del problema se hace más compleja.

### **Enfoques que utilizan pirámides irregulares**

Las pirámides irregulares de grafos, representadas como pirámides combinatorias, han sido desarrolladas por el grupo PRIP (Pattern Recognition and Image Processing) (TU Viena) [Brun 01, Haxhimusa 04, Kropatsch 04]. Una pirámide combinatoria recibe como entrada una imagen, y construye una jerarquía de particiones de la imagen usando el algoritmo *Minimum Spanning Tree* (MST) y las diferencias internas y externas de las regiones.

Las pirámides irregulares de grafos pueden solucionar las limitaciones de los RAGs (comentadas en el epígrafe 1.2.3) haciendo uso de los grafos duales para determinar las aristas importantes en la construcción de la pirámide. En este caso, cada nivel es un RAG extendido, donde las aristas paralelas y los lazos codifican relaciones importantes entre dos regiones (las aristas paralelas relevantes representan varios bordes en común y los lazos representan relaciones de inclusión). Además de la representación implícita de las relaciones espaciales, la jerarquía de particiones que proporciona la pirámide irregular es una importante fuente de información a distintos niveles de resolución. Este tipo de descripción puede ser muy provechosa en tareas como reconocimiento de objetos, etiquetado y recuperación de imágenes. Además, el uso de algoritmos de segmentación jerárquicos reduce la influencia de la sobresegmentación y la subsegmentación en dichas tareas.

En [Skurikhin 09] se utiliza un enfoque de pirámides irregulares para la segmentación de imágenes, aunque la diferencia con respecto a [Haxhimusa 04] es que, en lugar de comenzar la construcción de la pirámide a partir de una rejilla de píxeles, ellos comienzan la jerarquía a partir de un teselado triangular y poligonal de la imagen. Cada nivel representa un

teselado poligonal de la imagen y las aristas representan adyacencia entre los polígonos. La triangulación inicial es guiada por los bordes extraídos con un detector de bordes de Canny, lo cual puede provocar que algunos bordes se pierdan desde el mismo comienzo de la construcción de la pirámide.

Las pirámides irregulares de grafos, o pirámides combinatorias, han sido utilizadas mayormente con el fin de crear jerarquías de segmentaciones, es decir, en tareas donde la segmentación de la imagen es el objetivo principal [Haxhimusa 04, Haxhimusa 06, Torres 10, Gerstmayer 11].

Con objetivos ya relacionados con tareas de reconocimiento se han desarrollado menor cantidad de trabajos utilizando este enfoque. Entre ellos, se puede encontrar una propuesta de correspondencia de jerarquías [Brun 08], donde las estructuras subadyacentes son pirámides irregulares. El método va estableciendo la correspondencia entre regiones a distintos niveles de la jerarquía, guiado por los bordes codificados en el mapa combinatorio de cada nivel. Este último hecho hace que el algoritmo esté limitado al reconocimiento de objetos específicos en una misma escena (i.e. cuadros cercanos en una secuencia de video), y los objetos deben poseer un contorno bien definido y discriminativo. En [Antunez 12] se utilizan las pirámides combinatorias en el enfoque de atención visual artificial, donde se busca establecer el foco de atención en una escena, y luego se evalúa si el foco está correctamente posicionado sobre el objeto que se busca. Con la pirámide combinatoria se logra un algoritmo ascendente que descompone la imagen en regiones mediante la segmentación. Luego, un algoritmo descendente busca un objetivo específico en la jerarquía en función de la similitud de las regiones con el objeto que se busca. Se utiliza una estrategia de correspondencia de regiones para el reconocimiento, la cual se resuelve encontrando el clique máximo de un grafo de asociación. El mismo algoritmo de localizar las posibles ubicaciones de un objeto en la imagen es empleado en [Antúnez 13]. En este caso, un mapa combinatorio se codifica como una secuencia de símbolos, conteniendo cada fragmento de borde y su color, y luego se propone un algoritmo para establecer la correspondencia entre sub-mapas combinatorios a través de estas secuencias. Con esto se halla la similitud entre dos contornos (o formas) correspondientes al objeto buscado, lo que hace que el algoritmo sea eficaz para objetos con formas muy discriminantes (ej. señales de tránsito), pero menos apropiado para objetos más generales y en ambientes poco controlados. La propuesta de [Zankl 12] trata del etiquetado semántico de imágenes. Ellos utilizan el enfoque de las pirámides combinatorias, junto con un Campo Aleatorio Condicional (CRF) en el cual se modelan las relaciones jerárquicas entre regiones únicamente. La clasificación inicial de las regiones se va mejorando a partir

de la información proporcionada por los usuarios, los cuales pueden interactuar con el sistema corrigiendo regiones etiquetadas incorrectamente.

### **1.2.5. Conclusiones parciales**

Luego de analizar algunas teorías desarrolladas en el campo de la Psicología Cognitiva sobre cómo podría llevarse a cabo en el cerebro humano la percepción visual, se puede ver que muchas enfocan el problema hacia el trabajo con los contornos y formas de los objetos, al análisis de los objetos y escenas por partes y su descomposición en distintos niveles de resolución.

Desde el punto de vista computacional, aunque existen propuestas que se inspiran en los estudios sobre percepción humana, no son las líneas predominantes en el campo de la Visión por Computadora. Los métodos basados en segmentación son, en comparación, menos desarrollados e investigados, ya que usualmente han proporcionado menores tasas de eficacia debido al problema subyacente de la segmentación automática. No ha sido hasta el 2013 que han surgido nuevas propuestas basadas en segmentación que pueden competir con los métodos basados en regiones regulares. La utilización de la segmentación para el reconocimiento de objetos y otros propósitos de clasificación es útil, ya que brinda una localización precisa de las regiones y los objetos en una escena, con lo que se puede reducir considerablemente el espacio de búsqueda con respecto a los algoritmos de ventanas deslizantes multiescala y permite definir una mejor separación entre el fondo y los objetos en primer plano. También permite tener en cuenta el contexto de cada objeto o región a través de las relaciones espaciales que se establecen entre ellos en el plano de la imagen. No obstante, incluso para imágenes donde se muestra un mismo objeto, las regiones segmentadas pueden ser significativamente diferentes debido a distintas condiciones visuales como la iluminación, la pose del objeto, oclusiones, etc. Se ha analizado que la utilización de una jerarquía de particiones puede ayudar a mejorar estos problemas, pero además, el uso de las relaciones espaciales puede jugar un papel muy importante.

En los algoritmos mostrados, que resumen lo más relevante en esta área, se puede apreciar que son pocos los trabajos que combinan representaciones basadas en regiones irregulares, jerarquías de segmentaciones y relaciones espaciales entre las partes de la imagen. Algunos no explotan del todo las posibles configuraciones espaciales entre regiones, o no aprovechan la información explícita que proporciona la jerarquía. Esto da lugar a la necesidad de explorar de forma más profunda la utilización de las relaciones espaciales y jerárquicas

en representaciones basadas en segmentación, para la tarea particular del reconocimiento automático de objetos. El empleo de atributos de este tipo que puedan añadir información contextual, podría lograr que los algoritmos de reconocimiento basados en segmentación alcancen una mayor robustez, incluso ante segmentaciones automáticas no ideales. Es por esto que se considera que el análisis basado en regiones tendrá más importancia en esfuerzos futuros relacionados con el reconocimiento de objetos.

## Capítulo 2

# Representación jerárquica de imágenes combinando apariencia y relaciones espaciales



## Capítulo 1

# REPRESENTACIÓN JERÁRQUICA DE IMÁGENES COMBINANDO APARIENCIA Y RELACIONES ESPACIALES

En este capítulo se propone una representación jerárquica de la imagen que codifica la apariencia de las regiones y las relaciones espaciales entre ellas. Para crear la estructura jerárquica básica se utilizan las pirámides irregulares o pirámides de grafos, las cuales son introducidas al inicio del capítulo. A continuación se detallan las representaciones visuales o de apariencia escogidas para describir las regiones y posteriormente, se propone un nuevo descriptor espacial que permite caracterizar las aristas de los grafos y establecer similitudes entre las configuraciones espaciales de dos pares de regiones. Esta representación es común para los métodos de reconocimiento que se proponen en los capítulos siguientes.

## 2.1. Pirámides Irregulares

Las pirámides irregulares de grafos están formadas por un Grafo de Adyacencia de Regiones (*RAG*, por sus siglas en inglés) por nivel [Haxhimusa 04]. En estos grafos  $G = (V, E)$ , los vértices  $v \in V$  representan regiones en la imagen y las aristas  $e \in E$  representan relaciones de adyacencia entre cada par de regiones.

Cuando se construye una pirámide irregular de una imagen, cada nivel representa una partición del conjunto de píxeles en celdas o regiones, es decir, subconjuntos de píxeles conectados. En el primer nivel (nivel 0) de la pirámide, cada celda es un píxel y la vecindad de las celdas está definida por la 4-conectividad de los píxeles. Una celda en el nivel  $k + 1$  (nivel padre) es la unión de celdas vecinas del nivel  $k$  (nivel hijo) [Haxhimusa 04]. A los vértices y aristas de cada grafo se le pueden añadir atributos representando descripciones de las regiones correspondientes en las imágenes (ej. color, tamaño, valores de gris de los



píxeles) o descripciones de relaciones entre ellas (ej. valores de diferencia entre las regiones de los extremos), respectivamente.

Una pirámide irregular es, por tanto, un conjunto de grafos dispuestos en forma de pila, y reducidos sucesivamente a partir del nivel anterior, siendo el nivel base la imagen de entrada completa. Cada grafo es construido a partir del grafo del nivel inferior, seleccionando un conjunto de vértices (llamados vértices sobrevivientes) y uniendo cada vértice no sobreviviente a uno sobreviviente. De esta forma, cada vértice sobreviviente representa a todos los vértices no sobrevivientes que se unieron a él y se convierte en su padre [Kropatsch 04]. Esta relación padre-hijo puede ser recorrida hasta el nivel base y el conjunto de píxeles correspondiente a un vértice en el nivel base es llamado su *campo receptivo* (CR). Los pasos para construir una pirámide irregular son mostrados en rasgos generales en la Figura 2.1.

La selección de los vértices sobrevivientes (los vértices blancos en la Figura 2.1c) se puede realizar de diferentes formas. Una de ellas es buscar un *conjunto maximal independiente* (MIS, por sus siglas en inglés) que satisfaga las condiciones de que cada vértice no sobreviviente debe ser adyacente al menos a un vértice sobreviviente y dos vértices adyacentes no pueden sobrevivir. Más información sobre la selección de los vértices sobrevivientes puede ser encontrada en [Brun 01, Kropatsch 05].

Si los grafos utilizados en cada nivel de la pirámide son grafos simples (grafos sin lazos ni aristas paralelas), la representación de la estructura espacial de una imagen pudiera no ser del todo completa ya que solo quedaría representada implícitamente la relación de adyacencia simple. Usando solo estos tipos de grafos no se puede tener información sobre relaciones de inclusión entre 2 regiones, ni se puede representar el hecho de que dos regiones tengan más de un borde común (múltiple adyacencia).

Las pirámides de grafos duales son introducidas precisamente para dar solución a estos problemas. Un RAG es un grafo planar (i.e. un grafo en el cual las aristas pueden ser dibujadas en el plano sin cortarse), y para los grafos planares, siempre existe su correspondiente grafo dual. En el grafo dual  $\overline{G} = (\overline{V}, \overline{E})$ , cada vértice describe una cara (o área) en el grafo original  $G$ , y las aristas que conectan los vértices en  $\overline{G}$  corresponden 1:1 a las aristas del grafo  $G$  (ver Figura 2.2c). Ambos grafos son duales entre sí. En la pirámide de grafos duales, por cada nivel se almacenan ambos grafos  $(G, \overline{G})$ . Utilizando esta información sobre las caras del grafo  $G$ , es posible discriminar en el proceso de construcción de la pirámide cuáles son las aristas paralelas y los lazos relevantes que deben ser preservados porque representan relaciones de inclusión o múltiples adyacencias (ver

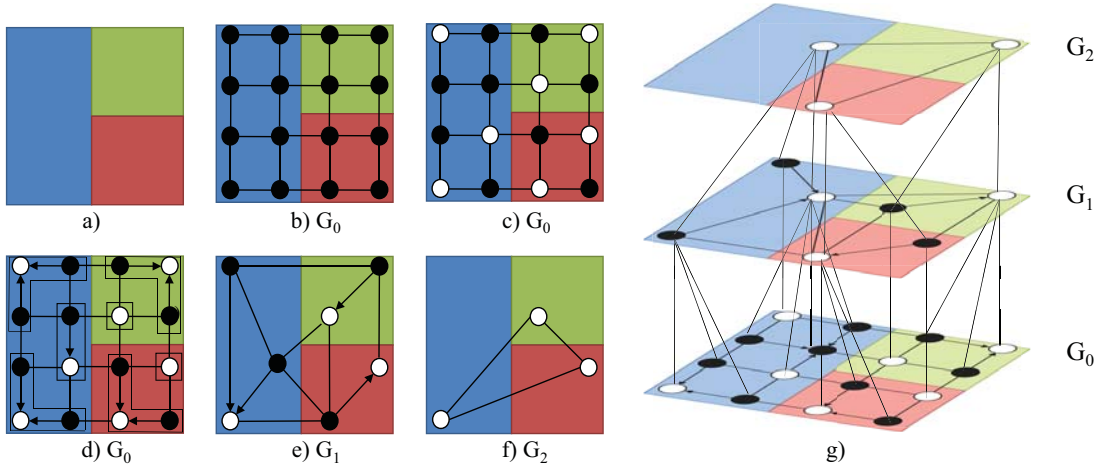


Figura 2.1: Construcción de una pirámide irregular a partir de una imagen. a) Imagen original, b) grafo del nivel base ( $G_0$ ), c) los vértices blancos son los que sobrevivirán en el nivel siguiente, d) núcleos de contracción (CK) para cada vértice sobreviviente (las flechas indican el vértice sobreviviente que corresponde a cada vértice no sobreviviente), e) nivel  $G_1$  construido a partir de  $G_0$ , f) nivel  $G_2$  construido a partir de  $G_1$ . En g) se muestra la jerarquía de niveles (los niveles  $G_0$ ,  $G_1$  y  $G_2$  se representan de abajo hacia arriba)

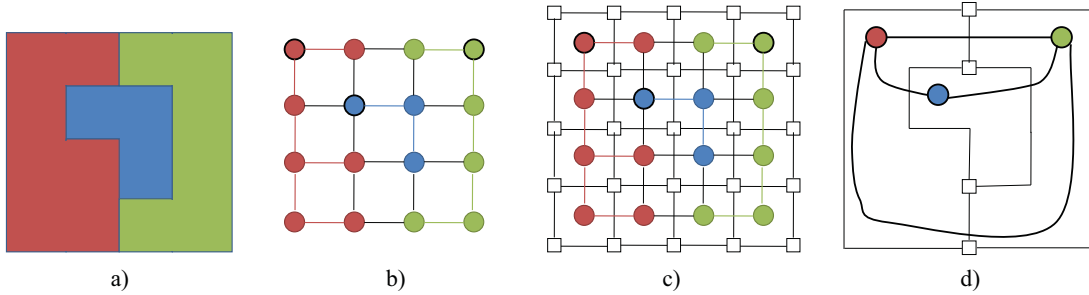


Figura 2.2: Grafo dual en la construcción de la pirámide. a) Imagen original, b) grafo del nivel base ( $G_0$ ), c) los cuadrados blancos corresponden a los vértices del grafo dual  $\overline{G}$ , representando las caras del grafo original  $G_0$ , d) último nivel de la pirámide donde se observa que entre los vértices rojo y verde existen dos aristas paralelas, las cuales representan dos adyacencias entre las regiones correspondientes (las aristas del grafo dual permiten determinar cuáles aristas del grafo original son relevantes).

Figura 2.2d). De esta forma, el RAG es sustituido por un RAG+ (grafo de adyacencia de regiones mejorado), el cual es un RAG que incluye lazos y aristas paralelas no redundantes [Kropatsch 04].

En la pirámide de grafos duales el proceso de reducción se lleva a cabo a través de un conjunto de contracciones de aristas. La contracción de una arista colapsa dos vértices adyacentes en un solo vértice y elimina la arista que los relacionaba. Este conjunto es llamado núcleo de contracción (CK de ahora en adelante, por sus siglas en inglés)

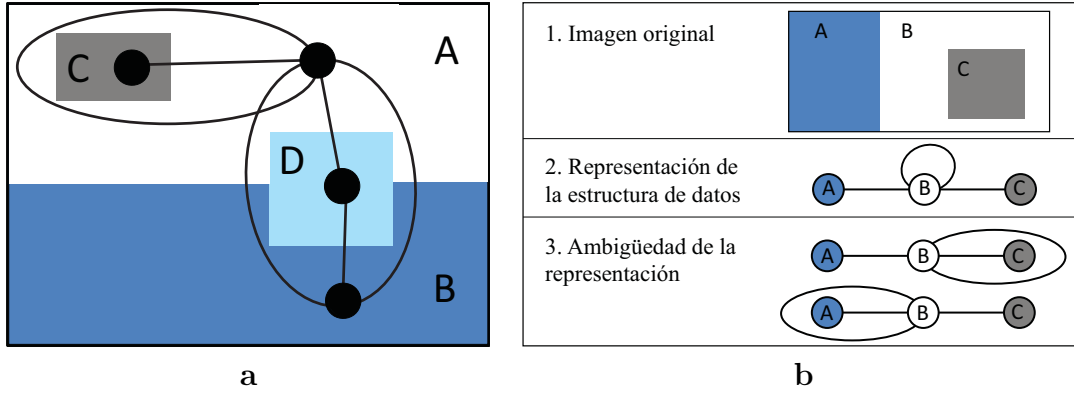


Figura 2.3: Representación de relaciones espaciales en la pirámide irregular. a) Representación de la adyacencia simple, múltiple y la relación de inclusión. b) Caso de ejemplo en el cual la representación de la inclusión no logra distinguir cuál región está adentro y cual está afuera.

[Brun 06]. La Figura 2.1d muestra un ejemplo de los CKs en un nivel de la pirámide. La contracción del grafo reduce el número de vértices, manteniendo la conexión a otros vértices. Como consecuencia, la reducción de un grafo mediante los CKs puede inducir la creación de aristas redundantes. El proceso de contracción debe seguir dos pasos [Brun 01]:

1. Realizar las contracciones de aristas en el grafo  $G_K$  correspondientes a los CK. El grafo dual del grafo reducido  $G_{K+1}$  se obtiene a partir del grafo dual  $\overline{G_K}$  eliminando las aristas duales de las aristas contenidas en el CK.
2. Eliminar las aristas redundantes luego de aplicar el CK al grafo dual. La contracción de aristas realizada en el grafo dual tiene que ser seguida por la eliminación de las aristas correspondientes en el grafo inicial, de forma que se preserve la dualidad entre los grafos reducidos.

Las relaciones espaciales representadas por la pirámide se muestran en la Figura 2.3. En un nivel dado de la pirámide, las aristas representan relaciones de adyacencia entre dos vértices (regiones). En la Figura 2.3a esto está representado por las aristas que conectan a la región A con la región D, y a la región D con la B. Además, es posible tener aristas paralelas entre dos vértices, las cuales representan múltiples adyacencias, como se muestra en el caso de la región A con la B. La relación de inclusión se representa con una arista simple denotando la adyacencia entre las regiones y un lazo, el cual rodea a la región que está adentro. Esta configuración se puede ver entre las regiones A y C.

Uno de los problemas de la codificación de las pirámides irregulares es que puede representar la presencia de una relación de inclusión, pero usando grafos no es posible

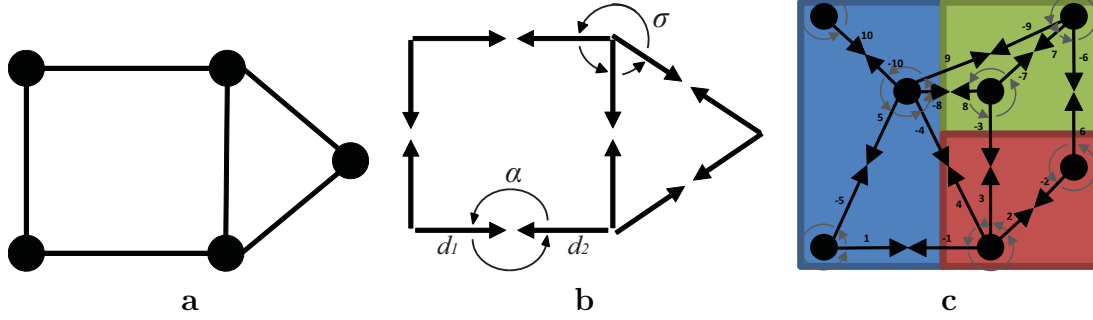


Figura 2.4: Equivalencia entre un grafo y un CM. a) Grafo de ejemplo, b) CM equivalente al grafo mostrado en a, c) CM equivalente al grafo de la Figura 2.1e. En c, los dardos son representados por segmentos numerados,  $\sigma$  por flechas grises (ejemplo  $\sigma(1)=-5$ ) y dos dardos enlazados por  $\alpha$  son dibujados uno a continuación del otro y comparten el mismo número, pero con distinto signo (ejemplo  $\alpha(1)=-1$ ).

saber cuál región está adentro y cuál está afuera teniendo únicamente un lazo [Brun 06]. Esto se puede observar en la Figura 2.3b, donde se observa que en la representación de la estructura de datos en forma de grafo, la única información que se tiene con respecto a la relación de inclusión es el lazo del nodo B, pero desde el punto de vista de la estructura de datos, queda la ambigüedad de si dicho lazo está rodeando a la región A o a la C, es decir, no se sabe cuál es la región incluida dentro de B.

Las pirámides combinatorias [Brun 06] fueron introducidas con el fin de caracterizar de una forma más adecuada la relación de inclusión, para lo cual se necesita la orientación de las aristas alrededor de un vértice. Un *mapa combinatorio* (CM, por sus siglas en inglés) se puede ver como un grafo planar que codifica explícitamente la orientación de las aristas, llamadas dardos, donde cada dardo tiene su origen en el vértice al cual está unido. Un CM se define como  $G = (D, \sigma, \alpha)$ , donde  $D$  es un conjunto de dardos (una arista que conecta dos vértices está compuesta por dos dardos  $d_1$  y  $d_2$ , perteneciendo cada dardo a solo un vértice),  $\alpha$  es la permutación inversa que relaciona a  $d_1$  con  $d_2$  y viceversa, y  $\sigma$  es la permutación de sucesión que codifica la secuencia de dardos que se encuentra al moverse alrededor de un vértice. Para mayor claridad, esto puede verse gráficamente en la Figura 2.4

El mapa dual de un mapa combinatorio es definido por  $G = (D, \varphi, \alpha)$ , con  $\varphi = \sigma \circ \alpha$ . Los ciclos de la permutación  $\varphi$  codifican el conjunto de dardos encontrados cuando se recorre una cara de  $G$  [Brun 06].

Una pirámide combinatoria se define entonces como un conjunto de mapas combinatorios

reducidos sucesivamente, teniendo la ventaja de que cada CM codifica explícitamente la orientación de los dardos alrededor de cada vértice, y el mapa dual se define sobre el mismo conjunto de dardos usando la permutación  $\varphi$ , por tanto, solo es necesario almacenar y actualizar una estructura en cada nivel de la pirámide [Brun 03].

La representación desarrollada en este trabajo se basa en pirámides combinatorias, por lo que de ahora en adelante se utilizarán los términos *pirámides irregulares* y *pirámides combinatorias* indistintamente, considerándose equivalentes.

## 2.2. Descripción visual de las regiones

Más allá de utilizar pirámides combinatorias para obtener un conjunto de segmentaciones de una imagen a distintos niveles, se puede ver la pirámide como una estructura que puede ser utilizada como un esqueleto para representar variados tipos de información y que puede ser usada en distintos procesos. La cuestión sería seleccionar para cada tarea específica, la información con que se llenará dicha estructura para que pueda ser explotada adecuadamente en cada dominio.

En el presente caso, en el reconocimiento de objetos, la pirámide se construye a partir de una imagen, como fue explicado en la Sección 2.1 y cada vértice corresponde con una región en una partición de dicha imagen, siendo las aristas la representación de adyacencia entre estas regiones. Cada región de las imágenes puede ser caracterizada utilizando rasgos de bajo nivel, de forma que estos representen partes distintivas de los objetos.

Existe un gran número de rasgos visuales propuestos en la literatura para esta tarea. En este trabajo se propone utilizar dos variantes de representación visual como atributos de los vértices de la pirámide.

### 2.2.1. Descripción de regiones usando color y textura

La primera variante, más sencilla y fácil de extraer de las imágenes, consiste en la representación del color y la textura de cada región. Los rasgos de color son extensamente utilizados en tareas de reconocimiento de objetos y de visión por computadora de forma general. Es por esto que se seleccionó como rasgo de bajo nivel el histograma de color en el espacio RGB. Los valores tridimensionales de espacio de color RGB hacen que el poder discriminativo de esta representación sea superior a los valores unidimensionales de las

imágenes en escala de grises. Por esta razón se decidió construir un histograma con 16 elementos por canal, quedando finalmente en un vector de dimensión 48.

Para la representación de la textura se seleccionó el histograma de patrones binarios locales (LBP, por sus siglas en inglés) [Ojala 96]. El operador LBP codifica el patrón de una vecindad local a partir de una región de textura, y su histograma es usualmente utilizado como rasgo de textura en problemas de clasificación. Entre las ventajas de este tipo de rasgo se encuentran su invarianza ante cambios monotónicos de niveles de gris y su simplicidad computacional. Además, algunos estudios han mostrado que los patrones binarios locales logran una buena discriminación entre texturas [Heikkilä 09, Takala 05].

En este caso se extraen los patrones binarios locales usando una ventana circular local para representar el pixel central y la distribución de los LBPs de una región es aproximada por un histograma LBP de dimensión 256.

La estructura de la pirámide combinatoria es adecuada para el cálculo de rasgos estadísticos tales como histogramas. El cálculo del histograma de cada región puede ser realizado durante la construcción de la pirámide de forma expedita, actualizando cada nivel con la información calculada en el nivel anterior. Dada una imagen obtenida a partir del cálculo de los LBPs de la imagen original, es posible actualizar el histograma de cada región en cada nivel usando la Ecuación 2.1, donde  $n$  es la cantidad de regiones que se unieron para formar la región analizada  $R$ , y  $j$  es el nivel de la pirámide. El mismo procedimiento se puede aplicar para el cálculo de los histogramas de color.

$$H(R)_j = \sum_{i=1}^n H(i)_{j-1} \quad (2.1)$$

Cuando se utiliza esta variante como descripción visual de las regiones, a cada vértice de la pirámide en cada nivel se le añade como atributo la concatenación de un histograma de color y un histograma LBP extraídos de la región que él representa. Esta concatenación proporciona un vector con 304 valores.

### 2.2.2. Descripción de regiones usando rasgos contextuales

La segunda variante de representación visual es más compleja de extraer de las imágenes y su utilización posterior resulta más costosa que la primera variante, pero su poder discriminativo es mayor, dado que incorpora el contexto de cada región en su representación individual.

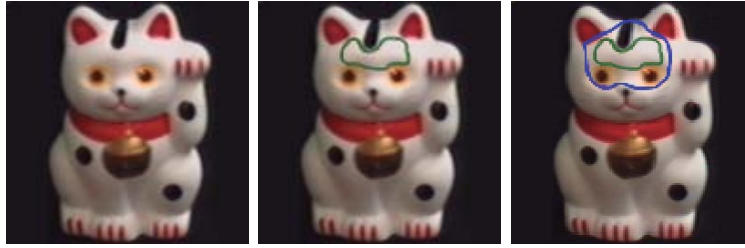


Figura 2.5: Ejemplo de problemas de las representaciones de bajo nivel cuando se utilizan algoritmos de segmentación irregular de imágenes. Primero se muestra la imagen original; a continuación, una región homogénea que podría ser obtenida mediante un método de segmentación; y finalmente, un ejemplo de cómo la utilización de una región mayor puede añadir información de contexto a la región analizada.

Como fue planteado en [Galleguillos 10], el contexto juega un papel crucial en muchos casos, especialmente cuando se utilizan segmentaciones irregulares de una imagen. Usualmente, los algoritmos de segmentación tienden a crear regiones mayormente homogéneas que brindan poca información cuando se extraen rasgos de bajo nivel de ellas. Es por esto que incluir un área mayor que la región que se analiza en el paso de extracción de rasgos, puede proporcionar pistas sobre las particularidades de dicha región que doten a las representaciones de mayor poder discriminativo. Para ilustrar esto se puede observar la Figura 2.5. Se puede ver la imagen original primero y en la segunda imagen se muestra un ejemplo de región irregular (delimitada por un borde verde) que podría ser resultado de un algoritmo de segmentación aplicado a la imagen. Como se puede observar en este caso, esta región es homogénea, de color blanco y no presenta una textura discriminativa. Por tanto, utilizar solamente la información de esta región para clasificarla puede introducir errores en el proceso, ya que una región con estas características puede aparecer en un gran número de objetos variados. Incluir un área mayor en la representación de la región (como la que se muestra en la tercera imagen delimitada por el borde azul), puede proporcionar información de contexto que ayude a desambiguar la clasificación de la misma.

Siguiendo esta idea, se decidió utilizar los rasgos de contexto basados en regiones (RCF, por sus siglas en inglés) propuestos en [Pantofaru 08], que combinan regiones irregulares con parches regulares para obtener una representación más discriminativa. Se escogió la representación usada en [Pantofaru 08], que consiste en un histograma de dimensión 100 que describe el color de la región, usando una cuantización del canal H en el espacio de color HSV y un histograma de rasgos RCF de dimensión 300.

Para calcular los rasgos RCF se extrajeron rasgos SIFT [Lowe 04]) de la división de la imagen mediante una rejilla regular para dos escalas distintas. Luego se creó un

vocabulario visual agrupando este conjunto de rasgos, quedando cada parche de la malla densa representado por una palabra de dicho vocabulario. Los primeros 150 elementos del histograma resultante corresponden a las ocurrencias de cada palabra visual dentro de la región analizada, solo teniendo en cuenta la primera escala. En el enfoque original de estos rasgos, los siguientes 150 elementos corresponden a las ocurrencias de las palabras visuales de la segunda escala que aparecen lo suficientemente cerca (utilizando una distancia) de la región en cuestión, añadiendo información de contexto de esta forma.

En esta tesis, para este último paso, se introdujo una modificación de forma que se aproveche la información brindada por la pirámide irregular. En lugar de utilizar una distancia para definir el contexto de una región, se utiliza como contexto de una región determinada en un nivel  $k$ , su región padre en el nivel  $k + 1$  de la pirámide. Esto significa que si se analiza una región  $r_k$  en el nivel  $k$  de la pirámide, los primeros 150 elementos del histograma RCF contarán las ocurrencias de las palabras visuales correspondientes a la primera escala que aparecen dentro de  $r_k$ . Los segundos 150 elementos del histograma contarán las ocurrencias de las palabras visuales correspondientes a la segunda escala que aparecen dentro de  $r_{k+1}$  en el nivel  $k + 1$ , donde  $r_k$  pertenece al núcleo de contracción de  $r_{k+1}$ .

Luego de este paso, se obtiene un vector de dimensión 400 concatenando el histograma de color con el histograma RCF para cada región. Además de esto, a la representación de cada región se le añade otro histograma de dimensión 400 conformado por un histograma de color y uno de RCF de la imagen completa, para un vector final de dimensión 800. Con este paso se le añade a cada región información de contexto de toda la imagen.

Con esta variante de representación visual, cada vértice de la pirámide recibe como atributo el vector de dimensión 800 antes mencionado. Para mayor claridad, en la Figura 2.6 se ilustra la formación de dicho vector.

1	...	100	101	...	400	401	...	500	501	...	800
Histograma del canal H de la región			Histograma de rasgos RCF de la región			Histograma del canal H de la imagen			Histograma de rasgos RCF de la imagen		

Figura 2.6: Concatenación del histogramas del canal H (del espacio de color HSV) y el histograma de rasgos RCF de la región analizada con los correspondientes histogramas de la imagen completa. Cada región de la imagen está caracterizada por este vector de rasgos.



### 2.2.3. Similitud visual

Una vez seleccionados los rasgos visuales que serán utilizados, un paso importante es la selección de las medidas de similitud entre ellos. Ya que la contribución de este trabajo no es en el campo de similitud visual, se decidió escoger una medida de similitud visual conocida en la literatura. Se emplea en este caso la distancia Euclideana, invirtiéndola para ser usada como medida de similitud.

Para ambas representaciones (la basada en color y textura o la que emplea rasgos contextuales) los atributos de apariencia son normalizados entre 0 y 1, pues los tamaños variables de las regiones pueden producir histogramas en rangos diferentes. Posteriormente se utiliza la distancia Euclideana entre los vectores correspondientes a las regiones que se comparen.

## 2.3. Descripción espacial entre regiones

En la literatura se han propuesto varios modelos para representar las relaciones espaciales entre regiones. Para el caso de las relaciones topológicas los modelos 4IM y 9IM [Egenhofer 93] son muy conocidos. En estos modelos, para el caso de las imágenes 2D, se describen 8 relaciones topológicas: *disjuntas*, *contiene a*, *adentro de*, *iguales*, *se tocan*, *cubierta por*, *cubre a* y *solapadas*. El principal problema de estos modelos es que no pueden representar relaciones topológicas más complejas, como cuando dos regiones tienen más de un borde en común.

Para el caso de las imágenes 2D, 8 relaciones topológicas son innecesarias ya que algunas de ellas nunca estarán presentes (ejemplo: solapamiento). En imágenes 2D puede darse el caso de la oclusión, que sería cuando dos objetos están solapados en el espacio, pero en el momento de efectuar la segmentación no será posible establecer la diferencia entre una oclusión y una adyacencia simple, ya que solo se tendrá un borde común entre los objetos. Es por esto que se decidió en este trabajo utilizar solamente, de las 8 relaciones topológicas mencionadas, las 3 más representativas para las imágenes 2D. Estas relaciones se muestran en la Figura 2.7. Para el caso de la relación *disjunta de*, se decidió no utilizarla porque es la que menos información brinda, es decir, si dos regiones no se tocan, probablemente la relación entre ellas sea menor. Además, si se incluyera en una representación basada en grafos, en lugar de tener un grafo de adyacencia de regiones, se tendría un grafo completo, lo cual sería computacionalmente mucho más costoso. La relación *igual a* tampoco se utiliza porque en una imagen 2D es imposible distinguirla.









							
A disjunta de B	A toca a B	A solapa a B	A igual a B	A contiene a B	A cubre a B	A adentro de B	A cubierta por B
A adyacente a B			A contiene a B			A está adentro de B	

Figura 2.7: Relaciones topológicas entre regiones 2D y la selección para utilizar en imágenes 2D.

Aunque estas relaciones están representadas implícitamente en la vecindad local de las pirámides combinatorias, recuperarlas explícitamente no siempre es una tarea sencilla y en algunos casos puede requerir varios pasos [Brun 06]. Los grafos en la pirámide irregular no son simples (contienen lazos y aristas múltiples), por lo tanto, realizar operaciones de correspondencia entre ellos no es trivial. Por esta razón se decidió crear un descriptor espacial que codifique explícitamente la configuración espacial entre dos regiones.

Fueron consideradas además las relaciones de orientación entre las regiones, pues las mismas pueden proporcionar información importante también. Por tanto, el nuevo descriptor espacial debe tomar en cuenta los dos tipos de relaciones. De acuerdo con esto, se decidió utilizar las relaciones de orden *a la izquierda de*, *a la derecha de*, *arriba de*, *debajo de*, *alineadas horizontalmente* y *alineadas verticalmente*, de alguna forma similares a la propuesta de [Hernández-Gracidas 07]. Estas relaciones serán calculadas a partir de la disposición espacial de los centroides de cada par de regiones.

### 2.3.1. Descriptor espacial

El descriptor espacial propuesto en esta tesis (previamente publicado en [Morales-González 10a, Morales-González 13a]) consiste en un vector binario que codificará las relaciones topológicas y de orden entre cada par de regiones. La ventaja de utilizar esta representación, es que se tiene una forma compacta y explícita de codificar las relaciones espaciales en forma de atributos en las aristas de un grafo, haciendo más fácil de manejar los grafos de la pirámide. Aunque la presencia de los lazos y aristas múltiples en los RAG+ es útil, a la hora de realizar operaciones con los grafos, por ejemplo, correspondencia de grafos, esta característica complejiza dichas operaciones. Esto significa que al reducir el número de aristas y caracterizar las que quedan con este atributo, se contribuye a la disminución del costo computacional. Otra ventaja que aporta utilizar este descriptor y no la representación del RAG+, es que se puede establecer una medida de similitud entre las aristas (aproximaciones entre configuraciones espaciales), mientras que la comparación únicamente entre estructuras proporcionaría

H	V	L	R	T	B	A	C	I
H – Alineados horizontalmente		L – A la izquierda de		T – Arriba de		A – Adyacente a		
V – Alineados verticalmente		R – A la derecha de		B – Debajo de		C – Contiene a		
						I – Adentro de		

Figura 2.8: Descriptor espacial que combina relaciones topológicas y de orientación.

una correspondencia exacta, lo cual es menos apropiado dada la variabilidad y ruido que pueden presentar los objetos y sus partes en las imágenes. Hasta este momento, no se ha encontrado en la literatura una representación de relaciones espaciales en esta forma.

El vector tendrá 9 elementos, cada uno representando una relación básica, como se muestra en la Figura 2.8. En cada posición se coloca un 1 si la relación espacial que representa existe entre el par de regiones que se analiza y 0 en caso contrario. Estas relaciones básicas se dividen en tres categorías: (1) relaciones topológicas - *adyacentes*, *contiene a* y *adentro de*, (2) relaciones de alineación - *alineadas horizontalmente* y *alineadas verticalmente*, (3) relaciones de orientación - *a la izquierda de*, *a la derecha de*, *arriba de* y *debajo de*. De esta manera, este descriptor es capaz de representar una relación espacial compleja a partir de relaciones espaciales básicas. La configuración espacial codificada de esta manera se puede analizar relación a relación o como un todo (Ver Anexo 1 para mayor claridad).

Para fines computacionales, cada valor del descriptor espacial será almacenado como bits, lo cual conduce a una representación de 9 bits (2 bytes con 7 bits sin utilizar), la cual es simple y fácil de utilizar.

Debido a que la múltiple adyacencia no se puede codificar dentro del vector binario propuesto, se almacena aparte, para cada par de regiones relacionadas, la cantidad de segmentos de bordes comunes que poseen (aristas paralelas entre los vértices), lo cual será un descriptor de la adyacencia entre ellas.

### 2.3.2. Similitud espacial

Para obtener la similitud entre dos configuraciones espaciales es necesario saber cuantas relaciones básicas comparten entre ellas. Por esta razón, se propone en esta tesis una medida de similitud que pueda ser utilizada para vectores binarios. Se decidió utilizar la medida de Sokal-Michener [Sokal 58] ya que en esta medida se tienen en cuenta tanto las correspondencias positivas como las negativas por igual. Además, ha demostrado un buen desempeño en comparación con otras medidas reportadas en [Zhang 03] y es fácil

de calcular.

Sean  $X$  y  $Y$  vectores binarios de la misma longitud  $d$  y  $S(X, Y)_{ij}$  ( $i, j \in \{0, 1\}$ ) denotará la cantidad de veces que el valor  $i$  en  $X$  y el valor  $j$  en  $Y$  aparecen en la misma posición en sus vectores respectivos. La medida Sokal-Michener para dos descriptores espaciales  $X$  y  $Y$ , que serán atributos de las aristas  $e_1$  y  $e_2$  respectivamente, puede ser calculada como se expresa en la Ecuación 2.2.

$$S_{SD}(X, Y) = \frac{S(X, Y)_{11} + S(X, Y)_{00}}{d} \quad (2.2)$$

El término  $S(X, Y)_{11}$  denota las correspondencias positivas (i.e. la cantidad de bits en 1 que correspondieron entre  $X$  y  $Y$ ) y el término  $S(X, Y)_{00}$  denota las correspondencias negativas (i.e. la cantidad de bits en 0 que correspondieron entre  $X$  y  $Y$ ).

En el cálculo de la similitud espacial entre dos pares de regiones se tuvo en cuenta además la consideración de que no todas las relaciones básicas deben contribuir con igual peso en el resultado final. Se consideró que las relaciones topológicas son más confiables que las otras ya que estas son invariantes a la rotación, escalado y traslación. Por esta razón, estas relaciones deben tener un peso mayor en la decisión de si dos configuraciones espaciales son similares o no. De la misma forma, se consideró que las relaciones de alineación deben ser más importantes que las relaciones de orientación. Es por esto que se utilizaron tres pesos  $\omega_T$ ,  $\omega_A$  y  $\omega_O$  para las relaciones topológicas, de alineación y de orientación, respectivamente, siguiendo el siguiente criterio:  $\omega_T > \omega_A > \omega_O$ . Estos pesos se aplican en el cálculo de la medida Sokal-Michener, a la correspondencia o no correspondencia de cada elemento de los vectores binarios, utilizando el peso correspondiente a la relación básica representada por cada elemento. En el Anexo 1 se ponen dos ejemplos de comparación entre descriptores espaciales y sus valores de similitud empleando esta medida.

## 2.4. Costo computacional de la representación

En el presente trabajo el objetivo fundamental es mejorar las tasas de eficacia encontradas en la literatura, por lo que el costo en memoria y en tiempo fueron factores secundarios en el desarrollo del mismo. No obstante, es importante hacer un análisis de cómo se comporta el costo en tiempo y memoria de la representación propuesta para brindar elementos que tributen a sus posibles aplicaciones.

De acuerdo con [González-Díaz 09], el costo computacional en tiempo de calcular la

pirámide irregular está dado por  $n$  que es la altura de la pirámide (número de niveles) y  $v_0$  que es el número de vértices en el nivel base (o número de píxeles en la imagen), tal que  $n \approx \log v_0$  (altura logarítmica). La cota superior de la complejidad computacional de contruir la pirámide es  $O(v_0 n)$ .

Existe un compromiso entre el costo en memoria de la representación y la eficiencia de los algoritmos que la utilizan. En cada nivel, cada vértice y arista recibe un atributo, que aumenta el espacio en memoria. Si se almacenan todos los atributos de todos los niveles, el costo en memoria será alto pero los algoritmos serán más eficientes en tiempo. Por el contrario, pudiera no almacenarse los atributos y calcularlos cada vez que se necesiten, con lo cual se disminuirá el costo en espacio, pero aumentará grandemente el costo en tiempo. En el trabajo de diploma “Recuperación de imágenes y videos por contenido utilizando MatchPyr” [Hernández-Saura 13], se propuso almacenar solamente los rasgos de los niveles que se utilizarán en cada método. En la Tabla 2.1 se muestra la cantidad de vértices y aristas de una pirámide irregular para una imagen de 128 x 128 píxeles. Se puede observar esta cantidad por cada nivel, el total para toda la pirámide y finalmente el total cuando se utilizarán solo los 3 niveles marcados con un asterisco (\*).

Tabla 2.1: Cantidad de vértices y aristas de una pirámide irregular para una imagen de 128 x 128 píxeles

Nivel	Cantidad de vértices	Cantidad de aristas
0	16385	33020
1	8410	21073
2	4458	12109
3	2395	6752
4	1281	3704
5	710	2080
6	393	1165
7	222	659
8*	128	375
9*	77	225
10*	51	149
11	33	93
12	21	59
13	14	36
14	6	14
15	3	5
16	2	2
Total	34589	81520
3 niveles (*)	256	749

Suponiendo que se utilice la representación visual de color y textura (que se denominará cLBP), se tiene un vector de 304 dimensiones por cada vértice. Asumiendo que sea un vector de tipo float (4 bytes), esta representación requerirá 1216 bytes por cada vértice. El descriptor espacial consta de 9 bits, por tanto se representa con 2 bytes

de memoria.

**Si se almacenan los rasgos de la pirámide completa:**

Vértices:  $34589 * 1216 \text{ bytes} = 42060224 \text{ bytes} = 40.11 \text{ Mb}$

Aristas:  $81520 * 2 \text{ bytes} = 163040 \text{ bytes} = 0.15 \text{ Mb}$

Total = 40.26 Mb

**Si se almacenan los rasgos de los 3 niveles que se utilizarán:**

Vértices:  $256 * 1216 \text{ bytes} = 311296 \text{ bytes} = 0.29 \text{ Mb}$

Aristas:  $749 * 2 \text{ bytes} = 1498 \text{ bytes} = 0.0014 \text{ Mb}$

Total: 0.29 Mb

Como se puede ver, es notable la reducción de espacio en memoria cuando se emplean solo 3 niveles de la pirámide con respecto a cuando se utiliza completa (de 40 Mb a 0.29 Mb). Se puede notar además que el costo en espacio de la representación de las aristas es despreciable con respecto a la representación visual. Esto es una de las características del descriptor espacial propuesto, que permite una representación muy compacta.

## 2.5. Conclusiones parciales

Mediante la utilización de las pirámides irregulares de grafos como esqueleto para la representación de las imágenes de forma jerárquica, fue posible combinar descripciones visuales que caracterizan la apariencia de las regiones en cada nivel con las relaciones espaciales que se establecen en el mismo entre dichas regiones.

Al realizar un análisis de las relaciones espaciales más importantes entre regiones bidimensionales, fue posible proponer un nuevo descriptor espacial que logra codificar distintas configuraciones espaciales teniendo en cuenta aspectos topológicos y direccionales. Este descriptor es utilizado como atributo en las aristas de los grafos de adyacencia de la pirámide para reducir la complejidad de los RAG+ (que poseen múltiples aristas paralelas y lazos) en procesos de comparación.

La propuesta de una medida para comparar vectores binarios permite establecer valores de similitud entre dos descriptores espaciales, lo cual es provechoso en algoritmos que busquen similitudes o aproximaciones entre subestructuras espaciales.



## Capítulo 3

# Reconocimiento de objetos en escenarios simples usando relaciones espaciales





## Capítulo 1

# RECONOCIMIENTO DE OBJETOS EN ESCENARIOS SIMPLES USANDO RELACIONES ESPACIALES

En este capítulo se presentarán dos nuevos métodos para el reconocimiento de objetos en escenarios simples (i.e. un objeto por imagen en condiciones controladas) utilizando la representación propuesta. El primer método fue pensado para el reconocimiento de clases de objetos y utiliza el enfoque de correspondencia de grafos, mientras que el segundo fue desarrollado para el reconocimiento de objetos específicos, utilizando el paradigma de bolsa de palabras. En ambos se hace explícita la utilización de las relaciones espaciales dentro de la representación jerárquica. Dado que la utilización de todos los niveles de la jerarquía puede resultar muy costoso computacionalmente, se propuso un criterio para seleccionar los niveles de segmentación que mejor preserven los bordes de la imagen. Estos algoritmos fueron evaluados en bases de datos de competencia internacionales, mostrando las ventajas y desventajas de ambos en tareas de reconocimiento de objetos específicos y de clases de objetos.

### 3.1. Selección de los niveles de la pirámide basada en los bordes de las particiones

Una de las ventajas de utilizar pirámides irregulares es que estas brindan una jerarquía de particiones para una imagen. Tener varios niveles de segmentación en lugar de uno solo puede ser muy útil en el proceso de reconocimiento, ya que diferentes particiones de una misma imagen proporcionarán información diversa. No obstante, no todos los niveles de la pirámide brindarán información relevante, pues es posible encontrar niveles demasiado sobre-segmentados o sub-segmentados, en los cuales la representación, lejos de ayudar, podría ser un lastre en cuanto a eficacia, pero especialmente en cuanto a eficiencia. Estos

niveles serán una carga adicional en los procesos de reconocimiento, haciéndolos más lentos sin añadir datos importantes.

Teniendo la jerarquía completa de particiones que brinda la pirámide, sería relativamente sencillo para una persona seleccionar manualmente un nivel o varios niveles que ella considere mejor segmentados, de acuerdo con algún criterio. No obstante, cuando se intenta hacer esto mismo de manera automática, surgen problemas tales como ¿cuáles niveles de la pirámide deberían ser utilizados para una tarea específica? o, si se fuera a seleccionar solo un nivel, ¿podría asegurarse que las partes de los objetos o los objetos mismos estén correctamente representados en esa partición?

Por estas razones, en este trabajo de tesis se decidió evaluar los niveles de la pirámide con el propósito de decidir cuáles niveles son los mejores (de acuerdo con algún criterio definido). En [Song 10] se propone un método para simplificar jerarquías de segmentaciones. La selección de los niveles más semánticos de la jerarquía se basa en teoría espectral de grafos, lo cual no es aplicable en este caso, ya que los grafos de la pirámide irregular contienen aristas múltiples para representar la adyacencia múltiple y la inclusión.

Se considera que los bordes de las imágenes pueden ser un criterio útil para evaluar las particiones. Cuando una partición no preserva los bordes relevantes de la imagen, esto usualmente significa que varias regiones de distintos objetos o del fondo fueron fusionadas en una sola, con lo cual se pierde información sobre las partes específicas que se unieron. Incluso una partición que segmente un objeto como una silueta completa podría no ser conveniente para el proceso de reconocimiento, ya que es de mayor interés encontrar las partes de un objeto y sus relaciones, de forma que se obtenga información discriminativa sobre el mismo.

Ya que no es posible contar con conocimiento a priori sobre los bordes de la imagen, se utilizó el filtro de Canny [Canny 86] para determinar los bordes más relevantes. La máscara de bordes resultante de este detector puede ser utilizada como referencia para evaluar la segmentación en cada nivel. De forma general, este detector muestra buen desempeño y es muy rápido, aunque presenta algunas desventajas, tales como los tres parámetros que es necesario ajustar y el problema de las uniones Y. Detectores de bordes más sofisticados podrían ser utilizados en el futuro para mejorar este proceso de selección.

Antes de aplicar el detector de Canny, las imágenes son suavizadas para disminuir la influencia del ruido en las mismas. Para evaluar cada partición de la pirámide, se proponen las medidas mostradas en las Ecuaciones 3.1 y 3.2,

$$B_{OK} = \frac{|P \cap R|}{|R|} \quad (3.1)$$

$$B_{NOK} = 1 - \frac{|P \setminus R|}{n} \quad (3.2)$$

donde  $P$  es el conjunto de todos los píxeles de borde de la partición que está siendo evaluada,  $R$  es el conjunto de píxeles de borde en la máscara de Canny y  $n$  es la cantidad total de píxeles en la imagen.  $|\cdot|$  denota la cardinalidad de un conjunto. La medida  $B_{OK}$  (ecuación 3.1) evalúa cuán bien los bordes de la partición se corresponden con los de la máscara de Canny, o sea, mide la presencia de bordes correctos en la partición; mientras que  $B_{NOK}$  (ecuación 3.2) evalúa cuántos píxeles de borde de la partición no están presentes en la máscara de Canny, es decir, cuantifica la presencia de bordes incorrectos. Por tanto,  $B_{OK}$  tiende a favorecer particiones sobre-segmentadas mientras que  $B_{NOK}$  hace lo opuesto y penaliza particiones con más bordes que los presentes en la máscara de referencia. Por esta razón se busca un balance entre estas dos medidas, combinándolas en una sola medida global, como se muestra en la Ecuación 3.3 en una suma pesada, utilizando los pesos  $\omega_G$  y  $\omega_B$ .

$$B = \omega_{OK} * B_{OK} + \omega_{NOK} * B_{NOK} \quad (3.3)$$

En la Figura 3.1 se puede observar un ejemplo de la evaluación de los niveles de la pirámide utilizando la medida  $B$ . En este ejemplo, el nivel 9 de la pirámide obtuvo la mejor evaluación. Se puede observar en el nivel siguiente (el 10) que algunos bordes de la cara de Lena se han perdido, mezclándose regiones del rostro con regiones de fondo. Además, algunos bordes de los objetos de fondo han desaparecido también. Es por esta razón que el valor de la medida  $B$  comienza a decrecer a partir del nivel 10 en adelante. Este proceso de evaluación de los niveles de la pirámide irregular propuesto por la autora de la tesis fue previamente publicado en [Morales-González 10b, Morales-González 13a].

## 3.2. Reconocimiento de objetos en escenarios simples usando correspondencia de grafos

Una tarea importante en el reconocimiento de objetos es la detección de objetos en las imágenes. Para esto es necesario contar con un método que delimite con suficiente precisión



Figura 3.1: Evaluación de los niveles de la pirámide utilizando la imagen de Lena.

dónde se localiza el objeto dentro de la imagen. Con este propósito se decidió desarrollar un método de correspondencia de grafos de manera que se explote la representación propuesta teniendo en cuenta las relaciones espaciales entre las partes de los objetos. En este caso, no se trata de encontrar la similitud entre dos imágenes completas, sino entre objetos presentes en las imágenes, por tanto se trata de un problema de correspondencia de subgrafos. El objetivo fundamental de este método es el reconocimiento de clases de objetos, lo cual es posible lograr relajando la correspondencia visual y espacial en busca de subestructuras comunes aproximadas, es decir, que preserven aspectos generales de apariencia y configuración espacial, y que no sean demasiado específicas de instancias de objetos.

### 3.2.1. Algoritmo de correspondencia de grafos

Se desarrolló en este trabajo de tesis, como parte del método de reconocimiento de objetos simples, un algoritmo glotón para encontrar la correspondencia entre subestructuras. Para podar el espacio de búsqueda (en el proceso de correspondencia) se utiliza la medida de similitud visual y la medida de similitud espacial (Ecuación 2.2) propuestas previamente para discriminar vértices y aristas que sean muy distintos entre sí. Inicialmente se obtienen todas las combinaciones de  $N \times M$  donde  $N$  y  $M$  son la cantidad de vertices de los grafos que se comparan. Con esto se reduce significativamente el espacio de posibles soluciones. Para hacer el algoritmo más explorativo, cuando no se puede avanzar más

en la correspondencia en una solución candidata debido a las restricciones visuales y espaciales impuestas, se comienza por otra de las posibles soluciones del espacio de soluciones  $N \times M$  obtenido previamente. Con esta estrategia de múltiples comienzos (multi-start) se intenta evitar caer en mínimos locales. Además, con el ánimo de reducir el tiempo de procesamiento, se utilizan para la comparación los mejores tres niveles de cada pirámide, evaluados por la medida  $B$  (Ecuación 3.3). Este algoritmo fue publicado previamente en [Morales-González 10a, Morales-González 13a].

A grandes rasgos, el algoritmo de correspondencia recibe como entrada un grafo  $G = (V, E)$  (que en este caso es el mejor nivel evaluado en la pirámide), que pertenece a la imagen donde se desea reconocer el objeto. Este será comparado con cierta cantidad de niveles seleccionados de alguna pirámide previamente almacenada. Para cada grafo (nivel seleccionado) en la pirámide se hallan todas las subestructuras que son similares al grafo de entrada. En cada solución candidata se toma cada vértice del grafo de entrada y se compara con cada vértice de un nivel de la pirámide, y si son similares visualmente, entonces se intenta expandir la estructura -de forma glotona- comparando las aristas de los vértices usando la medida  $S_{SD}$  pesada como se mostró en la ecuación 2.2. Si son similares espacialmente, se repite el proceso para cada vértice conectado por estas aristas. Esta estrategia de correspondencia de subgrafos está basada en el algoritmo propuesto en [Iglesias-Ham 07]. La representación visual de los vértices de las pirámides utilizada en este enfoque es la que emplea color y textura según lo descrito en la Sección 2.2.1.

Una vez que se cuenta con una subestructura  $T$  que correspondió con el grafo de entrada  $G$ , los vértices que correspondieron entre sí  $\{t_1, t_2 \dots t_n\}$  y  $\{g_1, g_2 \dots g_n\}$ , y las aristas que correspondieron entre sí  $\{r_1, r_2 \dots r_m\}$  y  $\{e_1, e_2 \dots e_m\}$ , se calcula la similitud visual global entre las subestructuras como se muestra en la Ecuación 3.4, siendo  $S_V(t_i, g_i)$  la similitud visual entre los vértices  $t_i$  y  $g_i$

$$VS(T, G) = \sum_{i=1}^n S_V(t_i, g_i) \quad (3.4)$$

y la similitud espacial global como se expresa en la Ecuación 3.5.

$$SS(T, G) = \sum_{i=1}^m S_{SD}(r_i, e_i) \quad (3.5)$$

El Algoritmo 1 muestra con más detalle lo que ha sido explicado previamente. En el Algoritmo 2 se muestra la función recursiva **Expandir()** empleada en el Algoritmo 1.

Para describir la vecindad de un v rtice  $v \in V$  se emple  la notaci n  $N(v) = (V_v, E_v)$  donde  $V_v \in V$  es el conjunto de v rtices adyacentes a  $v$  y  $E_v \in E$  es el conjunto de aristas incidente a  $v$ . La funci n `CalcularSimilitudVisual` calcula la similitud entre descriptores visuales, seg n lo explicado en la Secci n 2.2.3 y la funci n `CalcularSimilitudEspacial` calcula la similitud entre descriptores espaciales seg n la Ecuaci n 2.2 descrita en la Secci n 2.3.2.

---

**Algorithm 1:** Algoritmo de correspondencia de MATCH-Pyr.

---

```

input  : Grafo  $D = (V_D, E_D)$  de una imagen (un nivel de la pir mide irregular);
          Subgrafo  $G = (V_G, E_G)$  correspondiente a un objeto de una imagen
output: Similitud visual-espacial  $S$ 

1   $S_{max} = 0$ ;
2  foreach  $vd \in V_D$  do
3      foreach  $vg \in V_G$  do
4          if  $vd$  y  $vg$  no han sido marcados como correspondidos then
5               $v_{sim} \leftarrow \text{CalcularSimilitudVisual}(vd, vg)$ ;
6              if  $v_{sim} > 0$  then
7                  Marcar como ya correspondidos a  $vd, vg$ ;
8                   $[SS, VS] \leftarrow \text{Expandir}(vd, vg)$ ;
9                   $VS = VS + v_{sim}$ ;
10                  $S = VS * SS$ ;
11                  $S_{max} = \text{m x}(S, S_{max})$ ;
12             end
13         end
14     end
15 end
16 return  $S_{max}$ ;

```

---

Una vez que se encontr  la subestructura que mejor correspondi  con el grafo de entrada, se procede a analizar la forma de las regiones subyacentes para descartar objetos diferentes que pueden tener apariencias visuales similares.

### 3.2.2. Representaci n de la forma

La forma es un rasgo muy importante en el proceso de reconocimiento de objetos. Dos objetos pueden tener colores y texturas similares y ser solo diferenciables por su forma, por ejemplo, esto se puede ver en el caso de manzanas y peras, donde la diferencia principal entre estas categor as est  dada por su forma.

---

**Algorithm 2:** Función Expandir().

---

**input** : Vértices  $vd \in V_D$ ,  $vg \in V_G$  a ser expandidos

**output**: Similitud espacial  $SS$ ;

Similitud visual  $VS$

```
1 foreach  $ed \in E_{ed}$  do
2   foreach  $eg \in E_{eg}$  do
3     if  $ed$  y  $eg$  no han sido marcados como correspondidos then
4        $s_{sim} \leftarrow \text{CalcularSimilitudEspacial}(ed, eg)$ ;
5       if  $s_{sim} > 0$  then
6          $ug \leftarrow$  vértice adyacente a  $vg$  a través de  $eg$ ;
7          $ud \leftarrow$  vértice adyacente a  $vd$  a través de  $ed$ ;
8         if  $ud$  y  $ug$  no han sido marcados como correspondidos then
9            $v_{sim} \leftarrow \text{CalcularSimilitudVisual}(ud, ug)$ ;
10          if  $v_{sim} > 0$  then
11            Marcar como ya correspondidos a  $ed, eg, ud, ug$ ;
12             $SS_i, VS_i \leftarrow \text{Expandir}(ud, ug)$ ;
13            return  $[SS_i + s_{sim}, VS_i + V_{sim}]$ ;
14          else
15            return  $[0, 0]$ ;
16          end
17        end
18      else
19        return  $[0, 0]$ ;
20      end
21    end
22  end
23 end
```

---

En la actualidad se desarrollan muchos métodos de reconocimiento de objetos basados en la forma de los mismos, pero para esto se utilizan máscaras de segmentación ideales de la forma de cada objeto. En la práctica, cuando la imagen debe ser sometida a un proceso de segmentación o de extracción de bordes para obtener la silueta de un objeto determinado, los resultados todavía no son buenos, siendo este aún un problema abierto en Visión por Computadora.

Una de las ventajas de utilizar un método de correspondencia de grafos que vaya expandiendo las subestructuras que correspondieron entre sí, es que se puede tener la forma del conjunto de regiones subyacentes representadas por los vértices de los grafos. En este trabajo se decidió utilizar rasgos de forma en el análisis de los objetos, pues su



extracción utilizando la representación de las pirámides irregulares es expedita, y es un atributo importante en la comparación.

Como descripción de la forma de los objetos se escogió utilizar los momentos de Legendre [Cheriet 07], pues de acuerdo a la comparación realizada por [Arif 09], estos muestran muy buen desempeño para reconocer clases de objetos. Los momentos de Legendre también han sido utilizados exitosamente en aplicaciones relacionadas, como reconocimiento de caracteres escritos a mano [Duval 10] y recuperación de imágenes [Rao 10]. Los momentos de Legendre para una imagen digital ( $N \times N$ ) están dados por la Ecuación 3.6,

$$L_{pq} = \sum_{m=0}^{N-1} \sum_{n=0}^{N-1} P_p(m_N) P_q(n_N) f(m, n), \quad (3.6)$$

donde  $p$  y  $q$  son enteros entre  $(0, \infty)$ , siendo  $(p + q)$  el orden del momento de Legendre calculado,  $P_p$  y  $P_q$  son los polinomiales de Legendre, y  $m_N$  está definido por la Ecuación 3.7.

$$m_N = \frac{2m - N + 1}{N - 1} \quad (3.7)$$

El descriptor final de forma es la concatenación de los momentos de Legendre en un único vector. En este trabajo se decidió emplear hasta los momentos de orden 10, según el desempeño mostrado en [Duval 10] para esta configuración, con lo cual el descriptor de forma constará de 66 dimensiones.

El contorno asociado a la subestructura  $T$ ,  $C(T)$ , se puede hallar usando la Ecuación 3.8,

$$C(T) = C\left(\bigcup_{i=0}^n RF(t_i)\right) \quad (3.8)$$

donde  $C(\cdot)$  es el conjunto de píxeles que pertenecen al borde de la región correspondiente en la imagen y  $RF(t_i)$  es el campo receptivo del nodo  $t_i$  (ver Sección 2.1).

Una vez obtenida la forma de la subestructura  $T$  (delimitada por el contorno hallado) y calculado los momentos de Legendre de la misma, la comparación entre los vectores que la representan se realiza utilizando la distancia Euclideana, que brinda una medida de disimilitud entre estos rasgos. De esta manera, el descriptor de forma de la subestructura  $T$  es comparado con la forma del grafo de entrada  $G$ , resultando en una disimilitud de forma  $Sh_D(T, G)$ .

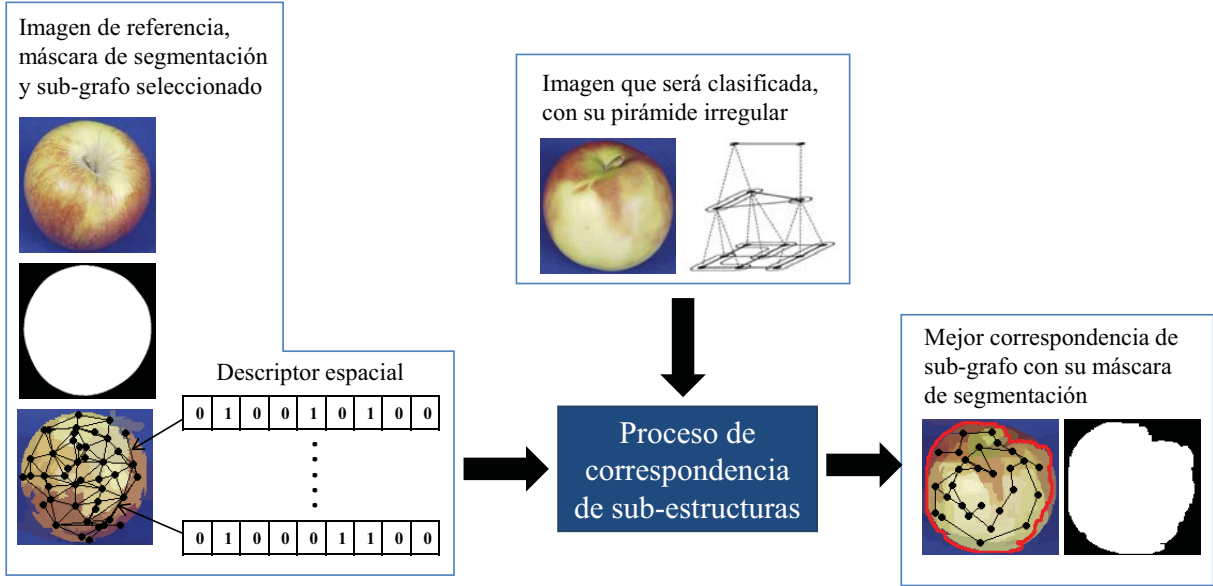


Figura 3.2: Proceso de correspondencia de subestructuras. Este proceso recibe como entrada la imagen que se desea clasificar junto con su representación como pirámide irregular (centro arriba). Esta imagen será comparada con las imágenes que se encuentran en la base de datos, para las cuales se tiene su máscara de segmentación ideal, así como el subgrafo que representa al objeto de interés en la imagen (izquierda). Después de ejecutado el proceso de correspondencia entre la pirámide a clasificar y el subgrafo del objeto de interés, el algoritmo devuelve el subgrafo de la pirámide que mejor correspondió con el grafo del objeto, así como la forma de la region total que representa dicho subgrafo (derecha).

### 3.2.3. Similitud global entre subestructuras

Una vez que se tienen las medidas de similitud visual  $VS(T, G)$  (3.4), similitud espacial  $SS(T, G)$  (3.5) y disimilitud de forma  $ShD(T, G)$ , la similitud global entre las subestructuras  $T$  y  $G$  está dada por la Ecuación 3.9.

$$S(T, G) = \frac{VS(T, G) * SS(T, G)}{ShD(T, G)} \quad (3.9)$$

La subestructura con el valor más alto de  $S$  será la mejor correspondencia para el grafo de entrada. Esta estrategia de correspondencia es mostrada a grandes rasgos en la Figura 3.2.

A este método de correspondencia de grafos utilizando la representación basada en pirámides irregulares se le nombrará MATCH-PYR y será con este nombre que se le referenciará en la sección de los resultados experimentales.

Para clasificar una imagen de entrada teniendo un conjunto de objetos de los cuales se conoce sus identidades, se emplea el clasificador del vecino más cercano (1-NN). Se compara esta imagen con cada uno de los objetos conocidos, aplicando la estrategia de correspondencia y obteniendo una similitud global en cada caso. La imagen de entrada será clasificada con la identidad del objeto que resulte más cercano a ella después de calculadas todas las similitudes globales.

### 3.2.4. Complejidad computacional de MATCH-Pyr

Haciendo un análisis de la complejidad computacional del algoritmo de correspondencia de MATCH-PYR, se tiene que para probar cada vértice de un grafo contra todos los vértices en el otro grafo se requieren  $O(n^2)$  operaciones, donde  $n$  es el número de vértices en un grafo. Una vez que se selecciona un par de vértices determinado, la expansión de la subestructura común entre ellos se realiza en  $O(n^2)$  operaciones, ya que el máximo grado de un vértice es  $n - 1$  y no se visita 2 veces un vértice o arista que ya se haya hecho corresponder previamente. Las similitudes entre aristas se calculan solo una vez y son almacenadas para futuras iteraciones. Esto significa que la complejidad computacional total es  $O(n^4)$ .

No obstante, esta complejidad computacional corresponde al peor caso, el cual es muy improbable que ocurra. De hecho, las podas que se realizan mediante la similitud visual y espacial juegan un papel crucial en el desempeño del algoritmo.

Se ejecutó una prueba para evaluar cuánto estas medidas son capaces de podar el espacio de búsqueda y los resultados obtenidos se muestran en la Tabla 3.1. La poda espacial se realiza sobre los resultados de aplicar la poda visual. La prueba fue realizada para el caso en que se comparan imágenes aleatorias (primera columna), que viene a representar el caso promedio, pues se trata de cualquier par de imágenes que se tome al azar. La segunda columna muestra el peor caso, en el cual se comparan imágenes similares por lo que se espera que la subestructura común se pueda expandir mucho y por tanto, la poda visual y espacial tendrán menor efecto. La tercera columna muestra el mejor caso, comparando imágenes diferentes, en las cuales deben existir pocas correspondencias y por tanto las podas deben tener un papel más activo. Lógicamente la poda para imágenes similares es menor que la poda para imágenes diferentes. Este análisis muestra que aunque la complejidad teórica del algoritmo es alta, es posible podar aproximadamente el 97% de las ramas involucradas en el proceso de expansión.

Tabla 3.1: Evaluación de cuánto podan el espacio de búsqueda las medidas de similitud visual y espacial.

	Imágenes aleatorias (caso promedio)	Imágenes similares (peor caso)	Imágenes diferentes (mejor caso)
Poda espacial	16.3 %	9.4 %	52.4 %
Poda visual	97.7 %	93.5 %	99.8 %

### 3.3. Reconocimiento de objetos simples usando el enfoque de bolsa de palabras

Para las aplicaciones de recuperación de imágenes por contenido no es necesario tener un método que delimite con precisión la ubicación del objeto dentro de la imagen (aunque esto no se descarta). Sería suficiente clasificar la imagen en cuanto a la presencia o ausencia del objeto en la misma y muchas veces esto ayuda a mejorar la eficiencia del método.

En este sentido, un enfoque de reconocimiento de objetos que difiere en gran medida de la correspondencia de grafos es el enfoque de bolsa de palabras. Los métodos de correspondencia de grafos usualmente tienen el problema de la alta complejidad computacional a la hora de realizar la clasificación. Los métodos basados en bolsa de palabras tienden a reducir la complejidad computacional y de forma general son más eficientes, pues agrupan los rasgos visuales de bajo nivel creando un vocabulario visual, donde cada palabra visual es el centro de cada grupo creado. Los rasgos visuales son luego sustituidos por estas palabras visuales y se crea un histograma de ocurrencia de palabras visuales en cada imagen.

Una de las mayores desventajas de los métodos de bolsa de palabras es que se pierde la relación espacial entre los rasgos visuales con los que se trata, teniéndolos todos precisamente en una especie de bolsa, sin orden de ningún tipo. Se han propuesto varias opciones para resolver este problema, la mayoría encaminadas a buscar relaciones entre las palabras visuales. Algunos ejemplos que siguen esta dirección empleando grafos fueron analizados en la Sección 1.2.3.

En este caso la propuesta que se hace en este trabajo (previamente publicada en [Acosta-Mendoza 12b, Morales-González 14]) es tener en cuenta las relaciones espaciales en la creación del vocabulario visual. Concretamente, las palabras visuales, en lugar de ser creadas a partir de rasgos visuales individuales, son creadas a partir de subgrafos frecuentes en las imágenes. Esta característica hace que el método sea más apropiado

para el reconocimiento de objetos específicos. Contrario a los enfoques clásicos de bolsa de palabras, que tienden a generalizar en cada palabra visual un conjunto grande de rasgos, al utilizar subgrafos frecuentes para crear el vocabulario se está promoviendo la aparición de subestructuras específicas de cada objeto.

### 3.3.1. Construcción del vocabulario visual

En los enfoques de bolsa de palabras se utilizan métodos de agrupamiento (usualmente k-means) para reunir rasgos visuales similares en un grupo cuyo centro será una palabra visual del vocabulario, y todos sus miembros serán etiquetados con esta palabra visual.

Teniendo una representación basada en grafos de una imagen, se pueden tener las relaciones espaciales entre los rasgos visuales representados por el grafo y se puede pensar en encontrar palabras visuales que sean subgrafos en sí. Es decir, una palabra visual, en lugar de representar un solo rasgo de bajo nivel, viene a representar una configuración espacial de rasgos que puede resultar significativa en el reconocimiento de los objetos específicos.

Para encontrar un vocabulario de subgrafos se decidió utilizar técnicas de minería de datos, que minan grandes colecciones de grafos buscando los subgrafos que resultan frecuentes en las mismas. Previamente se han utilizado métodos de minería de grafos exactos en el campo de las imágenes, teniendo como desventaja principal que estos métodos buscan una correspondencia exacta entre los grafos, y al trabajar con imágenes reales no tienen en cuenta la similitud que puede existir entre regiones parecidas. En este caso, la minería de Subgrafos Frecuentes Aproximados (FAS, por sus siglas en inglés) es una mejor opción, pues tiene en cuenta posibles distorsiones de los datos y no hace una correspondencia exacta, sino aproximada de los mismos a la hora del descubrimiento de los subgrafos frecuentes. Esta variante ha sido utilizada con dos algoritmos de minería de FAS: APGM [Jia 11] y VEAM [Acosta-Mendoza 12a], donde el primero trabaja únicamente con aproximación por similitud visual y el segundo utiliza aproximación por similitud visual y espacial. Estos métodos han sido probados en bases de imágenes creadas sintéticamente, o sea, que no se ha evaluado la complejidad de utilizarlos en imágenes reales donde la variabilidad en las condiciones de las imágenes es alta.

Un algoritmo de minería de subgrafos devuelve un conjunto de subgrafos  $S$  que aparecen frecuentemente en la colección para un umbral de soporte dado. El umbral de soporte indica la frecuencia con que debe ocurrir un subgrafo para ser considerado frecuente en la

colección. El conjunto  $S$  será el nuevo vocabulario visual empleado, donde cada palabra visual es un subgrafo.

### **3.3.2. Adaptación de la representación visual para trabajar con algoritmos de minería de FAS**

Para la utilización del algoritmo de minería de FAS es necesario hacer algunas modificaciones a la representación de las imágenes en el grafo, pues este tipo de algoritmos trabaja con grafos etiquetados.

Para este caso, se construye una pirámide irregular para cada imagen y se utiliza el método de evaluación de los niveles presentado en la Sección 3.1 para escoger el nivel (grafo) que representará a la imagen. El mejor nivel evaluado por la medida  $B$  es el que será utilizado.

Como representación visual de las regiones de la imagen (vértices del grafo) se utiliza la descripción mediante rasgos contextuales presentada en la Sección 2.2.2.

Para crear etiquetas para los vértices de cada grafo se utiliza un algoritmo de agrupamiento (k-means [Wang 13] en este caso) para agrupar los rasgos que representan todas las regiones de toda la colección. Los centros de cada grupo son tomados como etiquetas de los vértices. Para el caso de las aristas, mantienen la etiqueta asociada al descriptor visual presentado en la Sección 2.3.1, tomando como etiqueta el número decimal asociado al vector binario creado.

### **3.3.3. Creación de matrices de sustitución para los algoritmos de minería de FAS**

Para poder usar algoritmos de minería de subgrafos aproximados es necesario construir matrices de sustitución de vértices y de aristas, que, intuitivamente, representan la probabilidad de sustituir una etiqueta por otra, usando un criterio determinado según la aplicación. En este caso, teniendo una representación de las imágenes basadas en grafos, es de interés saber qué vértices pueden ser sustituidos por otros en términos de similitud visual de las regiones subyacentes. Además, es necesario saber cuáles aristas pueden ser consideradas equivalentes en términos de la similitud espacial que representan. Estas matrices de sustitución son la base que utilizan los algoritmos de minería de FAS para lograr la aproximación de los subgrafos.

En el presente caso, la matriz de sustitución de vértices para una colección dada será una matriz de  $n \times n$ , donde  $n$  es la cantidad de etiquetas creadas en el proceso de agrupamiento descrito en la Sección 3.3.2. Cada elemento de esta matriz almacenará la similitud entre dos etiquetas, la cual está dada por la similitud entre los centroides de cada grupo al que pertenecen. Esta similitud visual se calcula como se describió en la Sección 2.2.3, según la variante de representación visual usada para las regiones. Esto significa que cada elemento de esta matriz puede ser interpretado como la confianza de sustituir un vértice con etiqueta  $x$  con otro vértice con etiqueta  $y$  en un esquema de correspondencia.

La matriz de sustitución de las aristas es más sencilla de construir, ya que utilizando el descriptor espacial se tienen 27 configuraciones posibles de relaciones espaciales. La similitud entre aristas (i.e. los valores que serán almacenados en cada elemento de la matriz de sustitución) puede ser calculada usando la medida de similitud espacial descrita en la Sección 2.3.2.

### 3.3.4. Esquema de clasificación

El método de clasificación general puede resumirse en los siguientes pasos:

1. Se obtienen las representación basada en grafos de las imágenes de una colección utilizando la estructura de la pirámide irregular.
2. Se selecciona, para cada imagen, el nivel mejor evaluado con la medida  $B$  (Ecuación 3.3), el cual será el grafo que representará a la imagen en los pasos siguientes.
3. Se aplica K-means al conjunto de rasgos extraídos de las regiones de las imágenes de la colección y se etiqueta cada vértice usando el identificador del grupo al que corresponde su rasgo visual.
4. Se etiquetan las aristas de los grafos usando la representación decimal del vector binario que se utiliza como descriptor espacial (descrito en la Sección 2.3.1).
5. Se construye la matriz de sustitución de los vértices calculando la similitud visual entre todos los centroides de los grupos creados en el paso 3.
6. Se construye la matriz de sustitución de las aristas calculando la similitud espacial entre todas las posibles configuraciones espaciales que permite el descriptor espacial propuesto.
7. Se aplica el algoritmo de minería de FAS dándole como entrada la colección de grafos etiquetados y las matrices de sustitución de vértices y aristas y este devuelve un conjunto de subgrafos frecuentes en la colección que serán las palabras de un vocabulario visual  $V$ .

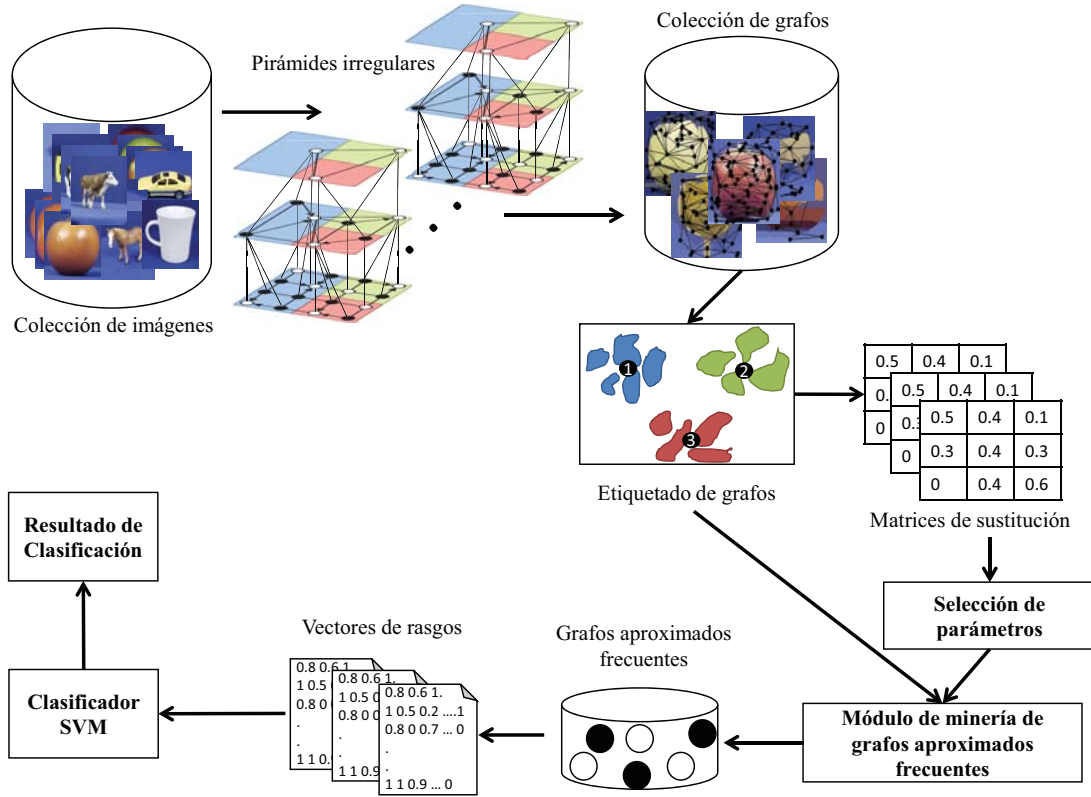


Figura 3.3: Método de clasificación usando un enfoque de bolsa de subgrafos.

8. Se crea un histograma por imagen, con dimensión del tamaño del vocabulario de subgrafos  $V$ , contando las ocurrencias de cada palabra del vocabulario en esa imagen.
9. Se utiliza un clasificador (ej. SVM [Perronnin 12]) para entrenar y clasificar las imágenes utilizando estos rasgos.

Para una mejor comprensión del flujo, estos pasos son ilustrados en la Figura 3.3. A este método se le llamará BoFAS-PYR y será referido con este nombre en lo adelante.

### 3.3.5. Complejidad computacional de BoFAS-Pyr

La complejidad computacional de minar los subgrafos frecuentes en una colección es exponencial de acuerdo al conjunto de patrones detectados como frecuentes. Uno de sus pasos es encontrar la forma canónica de cada grafo, que se realiza en  $O(n!)$ , donde  $n$  es la cantidad de vértices del grafo. No obstante, la operación de minado solo se realiza para obtener el vocabulario visual en el entrenamiento. Una vez que se tiene el vocabulario



se construyen histogramas de ocurrencias de los subgrafos frecuentes en cada imagen, y se realiza el entrenamiento y la clasificación con estos histogramas en tiempo lineal con respecto al número de imágenes a clasificar.

### 3.4. Experimentos

Se escogieron dos bases de datos reconocidas internacionalmente para el reconocimiento de objetos en escenarios simples: la COIL-100 [Nene 96] y la ETH-80 [Leibe 03]. Ambas colecciones contienen imágenes de objetos tomados desde distintos puntos de vista. Además se cuenta con máscaras de segmentación ground truth que delimitan con precisión la posición y forma de los objetos en cada imagen.

Para ambas colecciones de imágenes la configuración de los experimentos es la misma según el método que se pruebe. Para el caso del método MATCH-PYR, teniendo un conjunto de entrenamiento y un conjunto de prueba, las imágenes del conjunto de entrenamiento son representadas por un grafo que está dado por el mejor nivel evaluado por la medida  $B$  en cada pirámide. Se selecciona el subgrafo que representará a cada objeto usando la máscara de segmentación y se utilizan todas las regiones segmentadas que quedan totalmente dentro de esta máscara. Las imágenes del conjunto de prueba se representan con la pirámide completa, pero para el proceso de comparación se utilizan solamente los 3 mejores niveles evaluados por la medida  $B$ . En ambas colecciones, los resultados que se muestran de MATCH-PYR y BoFAS-PYR son el promedio de un proceso de validación cruzada con 10 corridas, donde en cada una se seleccionó aleatoriamente el conjunto de entrenamiento y el conjunto de prueba.

En el caso del método BoFAS-PYR, dado que cada imagen está representada por un vector, el proceso de comparación es mucho más rápido y eficiente, por lo que es factible en los experimentos la utilización de un clasificador (SVM en este caso).

Las pirámides irregulares construidas para estas imágenes tienen un promedio de 16 niveles. El nivel base contiene 16385 vértices y 33020 aristas, mientras que el nivel tope usualmente está formado por 2 vértices y 1 arista. En la mayoría de los casos, el mejor nivel seleccionado en el proceso de evaluación tiene entre 40 y 50 vértices y alrededor de 130 aristas.



Figura 3.4: Imágenes de ejemplo de la colección COIL-100 en la fila superior y de la colección ETH-80 en la fila inferior.

### 3.4.1. Experimentos en la colección COIL-100

Se realizaron experimentos en la colección de imágenes COIL-100 (Columbia Object Image Library) [Nene 96], la cual está conformada por imágenes en colores de 100 objetos. Estas imágenes fueron tomadas a intervalos de pose de 5 grados alrededor de un eje de rotación, quedando un total de 72 poses por objeto. En la fila superior de la Figura 3.4 se muestran algunos ejemplos.

Para los experimentos realizados se tomaron 25 objetos seleccionados aleatoriamente. Se seleccionó el 11 % de las imágenes para el conjunto de entrenamiento y el resto de las imágenes fueron usadas como prueba.

En esta colección de imágenes el objetivo del experimento es la identificación de objetos y los resultados alcanzados con los algoritmos presentados se comparan con otros métodos encontrados en la literatura. Esta comparación puede verse en la Tabla 3.2, donde en la primera columna se muestra el nombre del algoritmo (o algún alias), la segunda columna indica si dicho algoritmo utiliza o no relaciones espaciales y la tercera columna muestra la eficacia global de reconocimiento para esta colección.

Como puede observarse con estos resultados, los métodos propuestos mejoran todos los

Tabla 3.2: Eficacia global de reconocimiento en la colección COIL-100

Algoritmo	Relaciones Espaciales	Eficacia Global
DTROD-AdaBoost [Wang 06]		84.5 %
RSW+Boosting [Marée 05]		89.2 %
Patrones Secuenciales [Morioka 08]	X	89.8 %
MATCH-PYR	<b>X</b>	<b>91.6 %</b>
BoFAS-PYR	<b>X</b>	<b>99.4 %</b>
LAF [Obdržálek 02]		99.4 %

resultados mostrados, excepto para el algoritmo LAF (Local Affine Frames), el cual es un método específicamente diseñado para manejar cambios severos en la pose de los objetos, lo cual lo hace un buen algoritmo para la identificación de objetos, pero probablemente no lo sea para generalizar, como es el caso de la categorización de objetos. No obstante, BoFAS-PYR logra empatar este resultado, mostrando además desempeño notablemente mejor que MATCH-PYR para la tarea de reconocimiento de objetos específicos, como era de esperar. No obstante, se puede interpretar como buen resultado el hecho de que MATCH-PYR, incluso utilizando rasgos visuales bastante simples para las regiones, logre obtener mejores resultados que los otros algoritmos. Esto se pone de manifiesto si se analiza especialmente la comparación con el algoritmo Patrones Secuenciales [Morioka 08]. Este algoritmo utiliza relaciones espaciales en forma de patrones secuenciales, donde FOIs (rasgos visuales relevantes que ellos denominan foci-of-interest) presentes en las imágenes son concatenados a través de caminos extraídos de un grafo completo que comprende todos los FOIs. En los patrones secuenciales que representan las imágenes se introducen relaciones direccionales entre los FOIs. Este algoritmo, además de hacer uso de las relaciones espaciales entre partes de la imagen, utiliza rasgos visuales bastante más sofisticados que los empleados en MATCH-PYR, y aún así su resultado de eficacia global es menor que el de MATCH-PYR.

### 3.4.2. Experimentos en la colección ETH-80

La segunda colección utilizada es la ETH-80 [Leibe 03], la cual contiene 80 objetos pertenecientes a 8 categorías (*manzanas, carros, vacas, tazas, perros, caballos, peras y tomates*). Cada objeto es representado por 41 vistas distintas para un total de 3280 imágenes. Esta colección es más desafiante que la COIL-100 en el sentido de la diversidad de puntos de vista. En la fila inferior de la Figura 3.4 se muestran imágenes de ejemplo de esta colección.

Para el caso del método MATCH-PYR se dividió la colección, utilizando el 24 % como entrenamiento y el 76 % como imágenes de prueba a ser clasificadas. Para el método BoFAS-PYR se utilizaron 6 de las 8 clases para el experimento.

El objetivo de este experimento es la categorización de los objetos, lo cual es diferente a la tarea que se debía realizar en COIL-100, donde no era necesario especificar la categoría del objeto, sino simplemente identificarlos individualmente.

Los resultados en esta colección fueron comparados con otros métodos del estado del arte en términos de eficacia global de reconocimiento, y pueden ser observados en la Tabla 3.3.

Tabla 3.3: Eficacia global de reconocimiento en la colección ETH-80

LAF [Liu 12] <sup>1</sup>	68.4 %
DTROD-AdaBoost [Wang 06]	76.0 %
RSW+Boosting [Marée 05]	79.6 %
Pyramid match kernel [Grauman 05]	82.0 %
BoFAS-PYR	<b>84.4 %</b>
Discriminative Parts (DP) [Liu 12]	86.6 %
L5 ensemble classifier [Nomiya 09]	87.6 %
MATCH-PYR	<b>90.2 %</b>

En esta comparación se puede observar que el método MATCH-PYR obtuvo mejores resultados que los otros algoritmos. El método BoFAS-PYR no logró superar los resultados de MATCH-PYR ni del método *L5 ensemble classifier*, pero es importante resaltar que la colección ETH-80 es mayormente utilizada para probar algoritmos basados en la forma de los objetos, lo cual de cierta manera indica que la forma es una característica de gran peso en esta colección <sup>2</sup>. Los métodos MATCH-PYR y *L5 ensemble classifier* (de ahora en adelante L5) utilizan rasgos de forma en su proceso de clasificación, mientras que BoFAS-PYR no utiliza la forma en lo absoluto, con lo cual se encuentra en desventaja con respecto a estos dos métodos. No obstante, por la naturaleza de ambos, era de esperar que MATCH-PYR alcanzara mejores resultados que BoFAS-PYR en tareas de reconocimiento de clases de objetos.

Como se puede observar en la Tabla 3.3, la diferencia de los resultados entre MATCH-PYR y L5 es de 2.6%. El método L5 propone un modelo de aprendizaje combinado colaborativo, para el cual construyen cuatro tipos de combinaciones de clasificadores (L2, L3, L4 y L5) integrando dos, tres, cuatro y cinco clasificadores básicos respectivamente. Debe tenerse en cuenta en este caso que la colección ETH-80 cuenta con un conjunto de máscaras de segmentación ground truth de todas las imágenes, con lo cual cada objeto en las imágenes está idealmente segmentado. En el método L5, los cuatro primeros clasificadores básicos son basados en la apariencia de los objetos, mientras que el quinto clasificador básico usa únicamente la forma de los objetos, para lo cual se utiliza la máscara ideal del objeto. Con esto logran mejorar mucho sus resultados con respecto

---

<sup>2</sup>Los rasgos de forma, como se ha indicado con anterioridad, tienen de manera general un gran peso en el reconocimiento de los objetos, no obstante, desde el punto de vista del contenido de las colecciones y del propósito para el cual fueron creadas, no siempre la forma es un factor esencial para el reconocimiento en algunas de ellas.

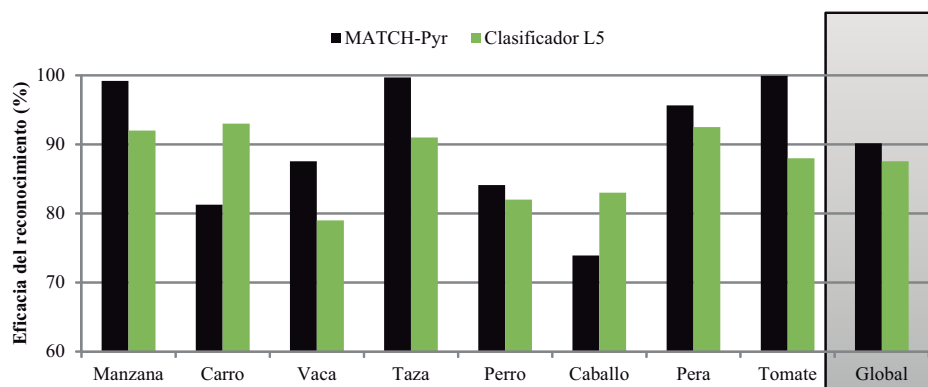


Figura 3.5: Eficacia del reconocimiento por categoría entre los métodos MATCH-PYR y L5 en la colección ETH-80.

a las otras combinaciones de clasificadores que no usan forma. Por otro lado, en el método MATCH-PYR se utilizan descripciones de forma en el proceso de reconocimiento, pero no se comparan las siluetas ideales de los objetos, sino que se va descubriendo la forma de los subgrafos que van correspondiendo entre sí, por lo que se pudiera considerar que este método se encuentra en desventaja con respecto al L5 en cuanto al tratamiento de la forma, ya que no utiliza la silueta ideal de los objetos en el proceso de correspondencia.

Para hacer un análisis más profundo de lo que sucede entre estos dos métodos, se puede observar en la Figura 3.5 una comparación de la eficacia en el reconocimiento por cada categoría de objetos.

De acuerdo con estos resultados, MATCH-PYR mejora los resultados de L5 en el reconocimiento de manzanas, tazas, tomates y vacas. El resto de las categorías no mostraron mejora con respecto a L5, aunque las peras quedaron bastante cerca. Esto muestra que MATCH-PYR es mejor para discriminar objetos con diferencias obvias en cuanto a apariencia y forma, pero es peor en la situación opuesta, siendo el caso de perros y caballos, donde la forma y apariencia son muy similares.

La matriz de confusión para MATCH-PYR se puede observar en la Tabla 3.4. Los valores que se muestran están en % y son un promedio de las 10 corridas. Estos han sido redondeados a sus enteros más cercanos para mayor claridad. Por ejemplo, en la primera fila de valores se muestra que, del total de manzanas a clasificar, el 99 % fue clasificado como tal y el 1 % fue clasificado como peras. Aquí se puede observar que la mayor confusión a la hora de clasificar ocurre entre objetos del mismo tipo, por ejemplo, animales (perros, vacas y caballos), los cuales en muchos casos presentan texturas, colores y formas similares. MATCH-PYR hace una mejor diferenciación entre manzanas y tomates, que comparten

Tabla 3.4: Matriz de confusión del método MATCH-PYR en la colección ETH-80.

	manzana	carro	vaca	taza	perro	caballo	pera	tomate
manzana	99	0	0	0	0	0	1	0
carro	5	81	3	0	5	4	2	0
vaca	1	1	87	2	3	5	1	0
taza	0	0	0	100	0	0	0	0
perro	3	3	5	0	84	3	1	0
caballo	4	3	8	1	6	76	2	0
pera	4	0	0	0	0	0	96	0
tomate	0	0	0	0	0	0	0	100

formas y colores similares, pero distintas texturas. También logra una mejor división entre manzanas y peras, que comparten colores y texturas similares, pero diferentes formas.

Es de notar además que el método LAFs, el cual obtuvo mejores resultados en la colección COIL-100, en esta colección obtiene los peores resultados, con un 68.4 % de eficacia global en la clasificación.

### 3.4.3. Evaluación de la relevancia de las relaciones espaciales

Con el fin de determinar la relevancia del uso de las relaciones espaciales en los métodos propuestos, se decidieron hacer experimentos adicionales donde se utilice el mismo enfoque, la misma representación propuesta, pero que no se haga uso de las relaciones espaciales.

Para el caso de MATCH-PYR, dadas dos imágenes y sus representaciones piramidales (según lo propuesto en este trabajo), se trata de hallar las similitudes entre cada región de una de las imágenes contra todas las regiones de la otra imagen utilizando solamente la medida de similitud visual. Se crea una matriz  $N \times M$ , donde  $N$  es la cantidad de regiones en la primera imagen y  $M$  es la cantidad de regiones en la segunda imagen. En esta matriz se almacena la similitud entre cada par de regiones. Luego de este paso, se utiliza el algoritmo Húngaro para encontrar la mejor configuración de coincidencia entre las regiones, desechando de esta manera toda la información espacial que existe entre ellas. Luego de encontrada la mejor configuración de coincidencia entre los grafos, se suman todas las similitudes visuales de las regiones correspondientes, según lo expresado en la ecuación 3.5, obteniendo el valor global  $VS(T, G)$  para todas las imágenes. La imagen

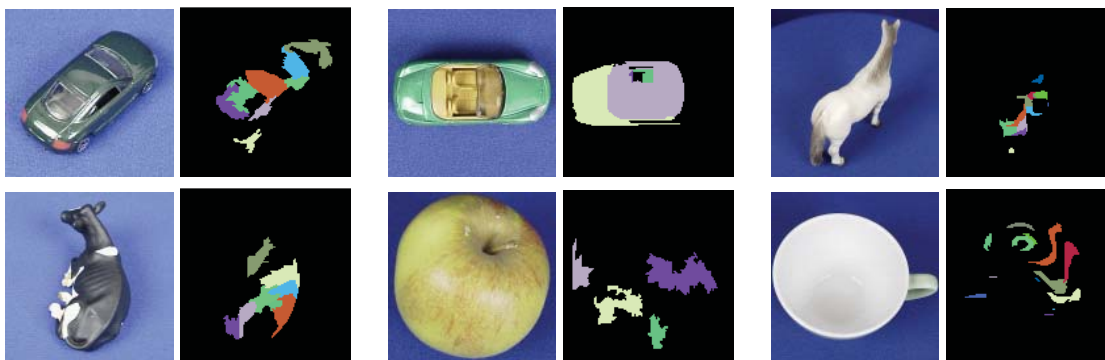


Figura 3.6: Ejemplos de imágenes clasificadas erróneamente utilizando solo la similitud visual. En la fila superior aparecen las imágenes de entrada, y la fila inferior muestra la mejor coincidencia encontrada en la colección en cada caso. Las columnas pares muestran las regiones que coincidieron entre las imágenes, donde estas coincidencias pueden ser identificadas por su color (se recomienda ver estas imágenes en colores).

con mayor valor es considerada la mejor coincidencia para la imagen de entrada.

Utilizando este enfoque, lo primero que se pudo notar fue que las regiones coincidentes entre dos imágenes están usualmente desconectadas, como se muestra en la Figura 3.6, por tanto, extraer el contorno utilizando la ecuación 3.8 y utilizar la distancia de forma  $ShD(T, G)$  no parece tener sentido en este caso. No obstante, en un intento por ser consistentes con el enfoque original de MATCH-PYR, se llevaron a cabo experimentos adicionales utilizando la información de forma de las regiones que coincidieron entre sí.

Los resultados obtenidos con este enfoque en la colección ETH-80 se pueden observar en la Tabla 3.5. Se puede notar que los resultados que se obtienen cuando no se utilizan las relaciones espaciales son bastante inferiores que los obtenidos cuando sí se usan. Cuando la medida de forma es introducida en el experimento, los resultados mejoran ligeramente, pero aún así están muy distantes de ser buenos o cercanos a los del enfoque original de MATCH-PYR. Esto se puede observar gráficamente en la Figura 3.7, donde las ventajas de utilizar la información espacial se pueden ver con mayor claridad.

Estos resultados muestran que integrar la estructura espacial para la correspondencia de diferentes partes de los objetos es importante, pues aunque exista una coincidencia entre los rasgos visuales, es necesario revisar su consistencia espacial. Esta consistencia espacial también es muy importante a la hora de extraer el contorno de la forma de cada subestructura, considerándola como un componente conexo, pues de otra manera el objeto puede ser extraído como una forma dividida en varios pedazos.

Tabla 3.5: Comparación de los resultados obtenidos usando relaciones espaciales y no usándolas, en términos de eficacia de reconocimiento. El método 1 es el enfoque original de MATCH-PYR, que utiliza información visual, espacial y de forma. El método 2 solo utiliza la similitud visual y el método 3 utiliza información visual y de forma (no utiliza relaciones espaciales).

Método	manzana	carro	vaca	taza	perro	caballo	pera	tomate	global
1	<b>99.2 %</b>	<b>81.3 %</b>	<b>87.6 %</b>	<b>99.7 %</b>	<b>84.1 %</b>	<b>73.9 %</b>	<b>95.7 %</b>	<b>99.9 %</b>	<b>90.2 %</b>
2	92.9 %	54.3 %	17.1 %	90.0 %	17.1 %	0 %	0 %	95.7 %	45.9 %
3	97.1 %	67.1 %	21.4 %	94.3 %	15.7 %	4.3 %	48.6 %	97.1 %	55.7 %

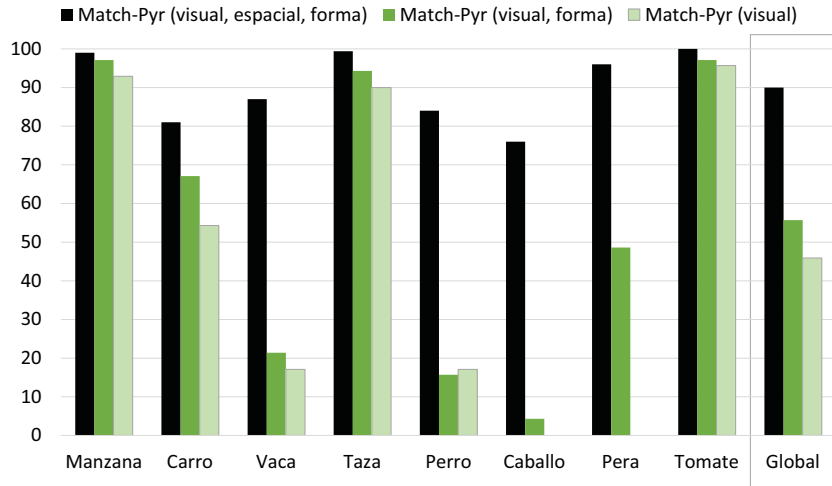


Figura 3.7: Eficacia del reconocimiento por categoría en la colección ETH-80 utilizando la variante original de MATCH-PYR con información visual, espacial y de forma, otra variante usando solo información visual y de forma y por último utilizando solo información visual.

#### 3.4.4. Análisis de la selección de umbrales y pesos

En este epígrafe se brindarán algunas consideraciones sobre la selección e influencia de los pesos y umbrales más importantes empleados en las propuestas.

##### Pesos $\omega_T$ , $\omega_A$ y $\omega_O$ para la comparación de descriptores espaciales

Para determinar los valores de  $\omega_T$ ,  $\omega_A$  y  $\omega_O$  introducidos en la Sección 2.3.2, se trató de modelar cómo los observadores humanos evalúan la similitud entre dos relaciones espaciales. Se preparó un conjunto de configuraciones visuales de relaciones espaciales



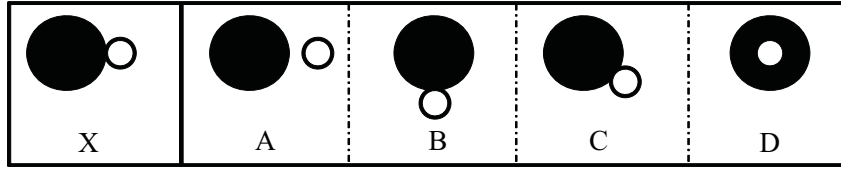


Figura 3.8: Relaciones de ejemplo usadas para probar cómo los humanos perciben la similitud espacial. De acuerdo al modelo propuesto, la relación de la región negra con respecto a la blanca en cada caso es descrita como: Relación X - alineadas horizontalmente, a la izquierda de, adyacentes; Relación A - alineadas horizontalmente, a la izquierda de; Relación B - alineadas verticalmente, arriba de, adyacentes; Relación C - a la izquierda de, adyacentes; Relación D - contiene a.

binarias. Ejemplos de estas pueden ser observados en la Figura 3.8.

Se les pidió a los observadores humanos que ordenaran las relaciones A, B, C y D según su similitud con la relación X. De forma general se encontró que las personas escogieron las relaciones C y B como las más similares a X. Se debe notar que para el caso de B, la relación de alineación es completamente distinta a la de X. No obstante, la relación topológica común (adyacencia) tiene un mayor efecto al ser evaluada la similitud espacial por los observadores humanos. La mayoría de los humanos colocaron a la relación D en la última posición porque la relación topológica es diferente. La relación A fue colocada frecuentemente en penúltima posición aún cuando A y X comparten relaciones de orientación y de alineación similares. Varias configuraciones de pesos fueron probadas, hasta que el orden resultante de la medida de similitud espacial fuera consistente con la percepción humana. Finalmente los valores asignados a los tres pesos fueron  $\omega_T = 4$ ,  $\omega_A = 2$  y  $\omega_O = 1$ .

### Pesos $\omega_{OK}$ y $\omega_{NOK}$ para la evaluación de los niveles de la pirámide

Para la selección de los valores de los pesos  $\omega_{OK}$  y  $\omega_{NOK}$  introducidos en la sección 3.1, se realizaron varias pruebas utilizando un conjunto de imágenes que contiene máscaras de segmentación realizadas por humanos manualmente como ground truth. Fue probado cuan bien el mejor nivel evaluado por la medida  $B$  correspondía con los bordes del ground truth, y se seleccionó la configuración de pesos que obtuvo mejores resultados en este sentido. Los valores resultantes para los pesos fueron  $\omega_{OK} = 0.4$  y  $\omega_{NOK} = 0.6$ , lo cual indica que se tratan de evitar mayormente sobresegmentaciones.

De manera general, la adecuación de esos pesos dependerá de la tarea que se desee resolver y el tipo de objetos que se desea reconocer. Por ejemplo, para el caso de reconocer personas,

se deben preferir niveles sobresegmentados. Para reconocer escenas naturales, se deben preferir los niveles subsegmentados.

En el ejemplo presentado en la Figura 3.9 se puede ver que, con los pesos  $\omega_{OK} = 0.4$  y  $\omega_{NOK} = 0.6$ , se obtiene un buen nivel de segmentación para la manzana (nivel 11), pero para la escena de los niños jugando fútbol, el mejor nivel evaluado (nivel 13) ya no contiene regiones que representen a los niños. Si en dicha escena lo más importante fuera reconocer el terreno y los arbustos, esta evaluación sería correcta, pero si lo que se desea es reconocer a las personas, se deberán modificar los pesos incrementando el valor de  $\omega_{OK}$  para que aumente la importancia de preservar bordes correctos aunque existan aún muchos bordes incorrectos. Por ejemplo, con los pesos  $\omega_{OK} = 0.5$  y  $\omega_{NOK} = 0.5$ , se evalúa como mejor el nivel 11 donde aún existen regiones que preservan los bordes de los niños.











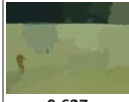






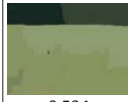
Niveles	3	6	8	11	13	15
$\omega_{OK} = 0.4$ $\omega_{NOK} = 0.6$	 0.548	 0.634	 0.650	 <b>0.657</b>	 0.620	 0.613
	 0.474	 0.561	 0.598	 0.624	 <b>0.627</b>	 0.617
$\omega_{OK} = 0.5$ $\omega_{NOK} = 0.5$	 0.501	 0.532	 0.547	 <b>0.553</b>	 0.543	 0.534

Figura 3.9: Ejemplo de cómo deben variar los pesos  $\omega_{OK}$  y  $\omega_{NOK}$  según la tarea que se desee realizar.

### Comportamiento del umbral de soporte para BoFAS-Pyr

Como se mencionó en la sección 3.3.1, el umbral de soporte indica la frecuencia con que debe ocurrir un subgrafo para ser considerado frecuente en la colección. Ejemplificando, si se habla de un umbral de soporte del 80 %, esto significa que para que un subgrafo sea considerado frecuente en una colección, debe aparecer en el 80 % de la misma. De aquí se puede deducir que cuando la frecuencia de ocurrencia de un subgrafo es muy alta, este no será muy discriminativo, pues aparecerá en la mayoría de las clases de la colección. Por el

contrario, si la frecuencia es demasiado baja, podría tratarse de un patrón ruidoso. Para analizar cómo se comportaron los valores de eficacia con respecto al umbral de soporte seleccionado, se muestran las Tablas 3.6 y 3.7.

Tabla 3.6: Eficacia según el umbral de soporte en la colección COIL-100.

Umbral de soporte (%)	80	70	60	50	40	30	<b>20</b>
Eficacia (%)	72.8	84.1	86.8	90.4	94.9	98.3	<b>99.4</b>

Tabla 3.7: Eficacia según el umbral de soporte en la colección ETH-80.

Umbral de soporte (%)	80	70	60	<b>50</b>	40
Eficacia (%)	77.3	81.0	82.1	<b>84.4</b>	76.2

Como se puede observar, en la colección COIL-100 el mejor resultado se obtuvo para el umbral de soporte 20 %, lo cual significa que se consideran frecuentes subgrafos que aparecen en el 20 % de la colección. Al ser el umbral bajo, esto es un indicador de que los grafos con menor frecuencia en el conjunto (dentro de los más frecuentes), son los que están aportando una mayor discriminación. Esto es consistente con la idea de que el método emplea subestructuras más específicas a la hora de identificar objetos.

Para el caso de ETH-80, el umbral de soporte con el que se obtuvieron mejores resultados fue 50 %. Recordando que este experimento es de categorización, un umbral intermedio es más adecuado a la hora de generalizar.

### 3.5. Consideraciones generales sobre los métodos propuestos

En esta sección se brindan aspectos generales de los métodos desarrollados, de forma que se pueda comprender en qué contextos pueden ser utilizados.

### 3.5.1. Comparación del costo computacional de MATCH-Pyr y BoFAS-Pyr

Tomando en cuenta los aspectos de complejidad computacional mencionados en las secciones 3.2.4 y 3.3.5, se brinda la Tabla 3.8 como resumen comparativo de los métodos MATCH-PYR y BoFAS-PYR en cuanto a eficiencia.

Tabla 3.8: Comparación del costo en tiempo de MATCH-PYR y BoFAS-PYR. Los tiempos fueron obtenidos en una PC Intel Core i5 Quad-Core a 3.20 GHz y 8 GB de RAM.

	MATCH-PYR	BoFAS-PYR
Complejidad computacional	$O(n^4)$	Orden exponencial para minar los subgrafos frecuentes y $O(n!)$ para encontrar la forma canónica de cada grafo
Operación más costosa	Clasificación 1-NN	Creación del vocabulario visual
Estructura utilizada para clasificar	Grafos	Vectores
Tiempo de entrenamiento	-	desde 4 horas hasta 4 días para umbrales de soporte de 80 % - 20 % respectivamente
Tiempo de clasificación de una imagen	40 segundos	< 1 segundo

Como se puede observar, MATCH-PYR no precisa de tiempo de entrenamiento, pero la clasificación es muy costosa, mientras que BoFAS-PYR requiere mucho tiempo para entrenar, pero la clasificación es muy rápida.

### 3.5.2. Aplicabilidad de MATCH-Pyr y BoFAS-Pyr

Como se ha mencionado con anterioridad, el MATCH-PYR tiene más valor en aplicaciones relacionadas con la detección de objetos en las imágenes, pues brinda la ubicación del objeto a partir de la correspondencia entre las regiones segmentadas. También es útil en aplicaciones donde se cuente con pocas imágenes de entrenamiento, pues al buscar generalidad en la categorización, permite encontrar objetos que difieran en el punto de vista, iluminación o que se encuentren ocluidos, contando con pocos ejemplos de referencia. Esto se pudo constatar en la aplicación desarrollada en el trabajo de diploma

[Hernández-Saura 13], donde se utilizó MATCH-PYR para detectar objetos en video, contando con muy pocas muestras de referencia.

BoFAS-PYR, por su parte, es más útil para aplicaciones de recuperación de imágenes donde se busque un objeto por imagen. Requiere más muestras de entrenamiento y este paso es muy costoso, pero la clasificación se realiza muy rápido, lo cual es importante en la recuperación. Este método no permite determinar la ubicación del objeto en la imagen, solo permite saber si está presente o no.

### 3.6. Conclusiones parciales

Mediante el uso de la representación jerárquica y espacial propuesta, los métodos MATCH-PYR y BoFAS-PYR mostraron resultados superiores a otros métodos existentes en la literatura para el reconocimiento de clases de objetos y de objetos específicos. En la colección COIL-100 obtuvieron 91.6 % y 99.4 % de eficacia respectivamente, mientras que en la colección ETH-80 mostraron un 90.2 % y 84.4 % de eficacia. Esto indica que mientras que MATCH-PYR es mejor en la categorización de objetos, BoFAS-PYR se desempeña mejor en la identificación de los mismos. Esto tiene sentido si se analiza que la forma en que trabaja BoFAS-PYR es buscando subestructuras frecuentes, lo cual puede ser más útil a la hora de describir objetos que se distingan más por ciertos patrones característicos (como los presentes en COIL-100). Este no es el caso de las manzanas, peras y tomates presentes en ETH-80, donde no se caracterizan por patrones distintivos, sino más por la apariencia y forma de los objetos presentes. En esto se especializa MATCH-PYR, en generalizar los objetos en cuanto a apariencia y forma, mediante la coherencia espacial que se logra a través de la correspondencia.

Realizando una evaluación de la importancia de las relaciones espaciales en el enfoque propuesto, se mostró una notable diferencia cuando son usadas las relaciones espaciales y cuando no lo son. El caso en el que se utilizan las relaciones espaciales, la apariencia y la forma superó en un 44.3 % a la variante de utilizar solo apariencia, y en un 34.5 % al utilizar apariencia y forma. Esto muestra que el papel que juegan las relaciones espaciales en este enfoque es fundamental.

No obstante estos resultados, estos métodos, por su propio diseño y complejidad, no son adecuados para reconocer objetos en imágenes más complejas, por ejemplo, escenas de paisajes naturales, ciudades, interiores de casas, etc. Una alternativa para reconocer objetos en escenarios complejos será descrita en el próximo capítulo.

## Capítulo 4

Reconocimiento de objetos en  
imágenes con escenarios complejos  
usando relaciones espaciales y  
jerárquicas



## Capítulo 1

# RECONOCIMIENTO DE OBJETOS EN IMÁGENES CON ESCENARIOS COMPLEJOS USANDO RELACIONES ESPACIALES Y JERÁRQUICAS

En el capítulo anterior se presentaron algoritmos para el reconocimiento de objetos en escenarios simples, donde existe solo un objeto por imagen en condiciones controladas. Cuando se analiza un escenario más complejo, como pueden ser paisajes naturales, escenas de ciudades o fotografías personales, la situación es diferente y los algoritmos presentados anteriormente presentan limitantes. En estos casos interactúan varios objetos por imagen, sus formas pueden no ser distintivas (ej. cielo, hierba), y están sujetos a condiciones muy variables en cuanto a iluminación, puntos de vista, oclusiones, etc. Además, la presencia de determinados objetos en la imagen puede influir positivamente para desambiguar la identidad o categoría de otros objetos. Se debe notar que la forma de analizar las relaciones espaciales para los objetos simples difiere para el caso de escenarios complejos, ya que, para objetos simples, se estaría describiendo la configuración espacial entre partes de un objeto, mientras que en escenarios complejos, las relaciones espaciales también podrían describir configuraciones espaciales entre objetos diferentes. Por esto se decidió buscar una alternativa de algoritmo de reconocimiento en el cual pueda ser utilizada la representación propuesta a la hora de encontrar varios objetos en una escena. Esta propuesta incluye además, la modificación de la construcción de la pirámide a medida que se va realizando el proceso de reconocimiento, para mejorar la segmentación subyacente.

### 4.1. Campos Aleatorios de Markov

Los modelos gráficos probabilísticos son una alternativa promisoría utilizada para modelar (en una forma más real) las relaciones entre datos dependientes del contexto. En



particular, los Campos Aleatorios de Markov (MRF, por sus siglas en inglés) [Spitzer 71], han sido empleados en el campo de Visión por Computadora debido a la posibilidad que brindan de modelar las relaciones de vecindad espacial en las imágenes. De manera intuitiva y en una línea de pensamiento basada en la semántica, se podría pensar que existen dependencias entre un píxel y sus píxeles vecinos, o entre una región y sus regiones adyacentes. Por tanto, estas relaciones se deberían tener en cuenta si se pretende acortar la brecha semántica.

Informalmente, los MRF son modelos gráficos no dirigidos que combinan la información de un conjunto de observaciones con la información de las interacciones de dichas observaciones con sus observaciones vecinas. Formalmente se puede decir que  $Y = \{Y_1, Y_2, \dots, Y_n\}$  se llama un campo aleatorio de Markov, siendo  $Y_i$  variables aleatorias sobre un conjunto de sitios  $S$ , que pueden tomar valores  $y_i$  de un conjunto de etiquetas  $L$ . Esto puede ser visto como un grafo no dirigido, donde cada vértice  $i$  representa a la variable aleatoria  $Y_i$  y las aristas representan relaciones de dependencia directa entre variables. De aquí en adelante los términos “vértice” y “variable” se utilizarán indistintamente.

Un MRF es un campo aleatorio que obedece la propiedad de Markov  $P(y_i | y_{i-1}, y_{i-2}, \dots, y_1) = P(y_i | N(y_i))$ , donde  $N(y_i)$  es el conjunto de vecinos de  $y_i$ . Esto significa que un vértice  $i$  solo es dependiente del conjunto de sus vecinos  $N(y_i)$ , y es independiente del resto de los vértices del grafo. Esto se ilustra en la Figura 4.1.

La configuración más probable de etiquetas  $Y^*$  para un MRF es la que maximiza la probabilidad conjunta  $P(y)$ . Esta probabilidad conjunta es modelada por algunas restricciones representadas por probabilidades locales, llamadas potenciales. Los potenciales pueden ser interpretados como restricciones que penalizan o favorecen determinadas configuraciones de  $Y$ . La probabilidad conjunta es expresada según la Ecuación 4.1.

$$P(y) = \frac{1}{Z} * \exp^{-U_p(y)} \quad (4.1)$$

donde  $Z$  es la función de partición o constante de normalización y  $U_p(y)$  es la función de energía.  $U_p(y)$  es calculada utilizando los potenciales mencionados anteriormente, como se muestra en la Eq. 4.2.

$$U_p(y) = V_O(y) + \lambda \sum_I V_I(y, y') \quad (4.2)$$

$V_O(y)$  es el potencial de asociación (o potencial unario), el cual representa la información

proveniente de las observaciones.  $V_I(y, y')$  es el potencial de interacción (o potencial por pares) y modela la información obtenida de los vértices vecinos  $(y, y')$ .  $\lambda$  es una constante introducida para asignar pesos a las restricciones impuestas por las funciones de potenciales. La configuración óptima MAP <sup>1</sup>  $Y^*$  se obtiene minimizando el valor de  $U_p(y)$ . Entre los métodos más usados para obtener esta configuración óptima están las Modas Condicionales Iterativas (ICM, por sus siglas en inglés), el Recocido Simulado y la Propagación *Loopy Belief* (LBP por sus siglas en inglés). Aunque es posible utilizar en el modelo potenciales de orden superior (haciendo  $y_i$  dependiente de un número  $O$  de variables), se prefiere en este contexto usar solo los potenciales unarios y de pares, pues los potenciales de orden superior incrementan grandemente el costo computacional a la hora de encontrar la configuración óptima.

## 4.2. MRFs aplicados a la Visión por Computadora

Entre los enfoques que se han propuesto para utilizar los MRFs en el reconocimiento de objetos, se encuentra [Xiang 09], el cual presenta un MRF múltiple donde, en lugar de construir un único MRF, se construye un MRF para cada etiqueta de objeto, perteneciente a un vocabulario. Esto se hace con el objetivo de capturar distintas semánticas entre las etiquetas. La propuesta de [Llorente 10] explora las dependencias entre los rasgos y distintas etiquetas. En [Escalante 07] y en [Hernández-Gracidas 07] se propone utilizar como potencial de interacción de un MRF, la información de co-ocurrencia entre etiquetas y las probabilidades de ocurrencia de las relaciones espaciales entre pares de etiquetas respectivamente.

Aunque estos enfoques se han centrado en explorar las dependencias entre rasgos y etiquetas, y entre pares de etiquetas vecinas, no tienen en cuenta las relaciones padre-hijo que pueden existir en una jerarquía de segmentaciones de una imagen. La utilización de una jerarquía con modelos de MRF no es una idea nueva. En [Keuper 11] los autores proponen un método de segmentación que utiliza dos niveles de una jerarquía. La clasificación de las regiones se realiza independientemente en cada nivel y luego se combinan dentro del modelo MRF. Enfoques jerárquicos han sido utilizados también para la segmentación de texturas [Kim 06] y para eliminar ruido [Cao 11], donde la jerarquía consiste en dos o tres capas que representan distintas características de la imagen.

En este capítulo, el enfoque que se propone, a diferencia de los explicados anteriormente,

---

<sup>1</sup>Maximum A Posteriori

utiliza una jerarquía de segmentaciones de la imagen y se construye un MRF por cada nivel, que será nutrido con la información proveniente de los MRFs calculados en niveles adyacentes. De esta manera, se vincula la información espacial en cada nivel y la relación jerárquica entre niveles de la imagen.

### 4.3. Re-etiquetado de imágenes usando MRFs jerárquicos

La definición de los MRFs y la mayoría de sus aplicaciones en imágenes tratan con dos relaciones básicas: la relación entre los rasgos de una región (observación) y una etiqueta, y la relación entre dos etiquetas vecinas. En esta tesis se propone incluir una relación padre-hijo en el modelo del MRF, inspirada por la noción de que en una representación jerárquica, las regiones hijas pudieran tener una información relevante para influir en la clasificación de la región padre y viceversa.

Se propone en este trabajo de tesis (sustentado en la publicación [Morales-González 12]) construir un MRF por cada nivel de la pirámide irregular que representa una imagen, empezando de abajo hacia arriba. En cada nivel  $l$ , la información de la mejor configuración de etiquetas  $Y^*_{l-1}$  obtenida en el nivel  $l - 1$  es utilizada como información adicional para calcular la configuración de etiquetas  $Y^*_l$  en el nivel actual. Cuando se alcanza el nivel superior de la pirámide, se repite el mismo proceso de arriba hacia abajo, ahora utilizando  $Y^*_{l-1}$  y  $Y^*_{l+1}$  para calcular  $Y^*_l$ . El MRF en cada nivel tendrá la estructura del RAG subyacente de la pirámide irregular.

La vecindad Markoviana  $N(y_i^l)$  de la etiqueta  $y_i^l$  puede ser dividida en dos vecindades: la vecindad espacial y la vecindad jerárquica. La vecindad espacial de la etiqueta  $y_i^l$  correspondiente al vértice  $i$  está compuesta por las etiquetas asignadas a todos los vértices adyacentes a  $i$  en el RAG del nivel  $l$ . Esto es ilustrado en la Figura 4.1, donde se muestra el RAG de un nivel de la pirámide, con su MRF asociado. En el MRF, la vecindad Markoviana espacial del vértice rojo está compuesta por los vértices verdes.

La vecindad jerárquica está formada por todas las etiquetas asignadas al núcleo de contracción (CK, ver sección 2.1) del vértice  $i$  en el nivel  $l - i$  y por la etiqueta de su padre en el nivel  $l + 1$ . Gráficamente esta relación puede verse en la Figura 4.2, donde la vecindad jerárquica del vértice rojo está formada por los vértices verdes.

Se propone calcular la función de energía según la Ecuación 4.3.

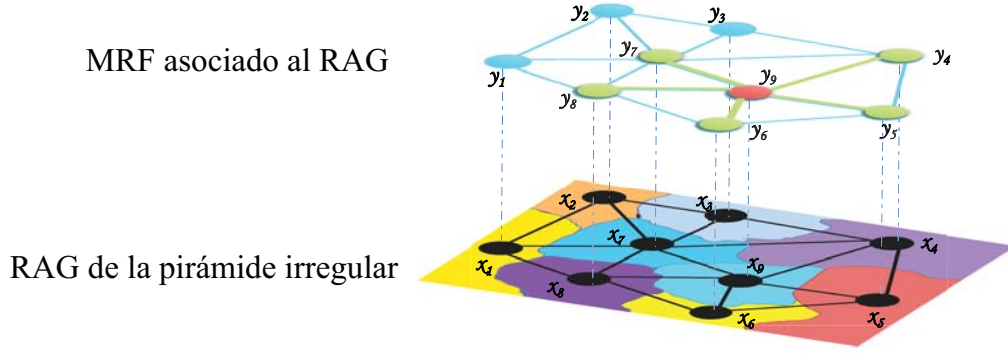


Figura 4.1: RAG de un nivel de la pirámide con su MRF asociado. Se ilustra la vecindad espacial Markoviana del vértice rojo, compuesta por los vértices verdes.

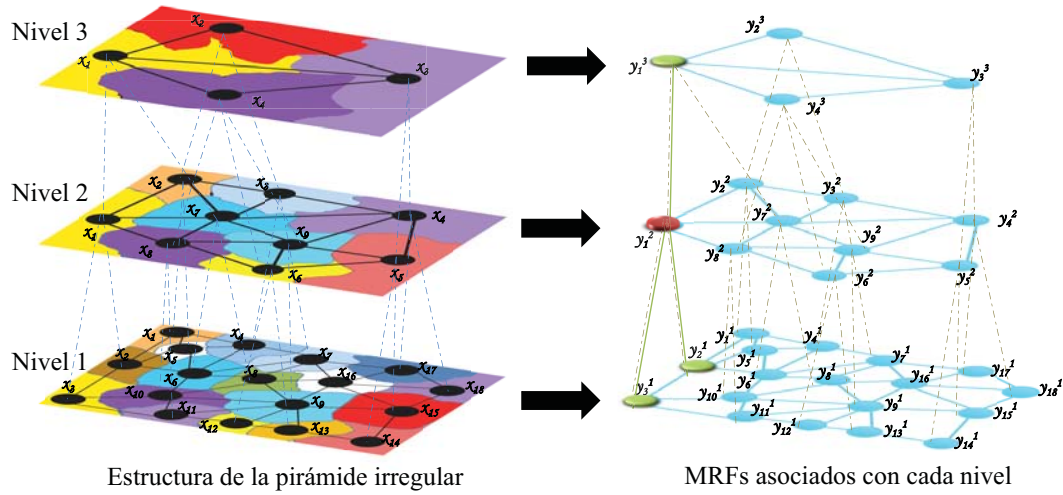


Figura 4.2: Niveles de la pirámide con sus MRFs asociados. Se ilustra la vecindad jerárquica Markoviana del vértice rojo, compuesta por los vértices verdes.

$$U_p(y_l) = \lambda_O V_O(y_i^l) + \lambda_I \sum_i V_I(y_i^l, y_j^l) + \lambda_H \left( \sum_k V_{Ch}(y_i^l, y_k^{l-1}) + V_P(y_i^l, y_m^{l+1}) \right) \quad (4.3)$$

Esto es una extensión de la Ecuación 4.2, introduciendo el término  $V_{Ch}(y_i^l, y_k^{l-1})$  como potencial jerárquico, que modela la relación de la etiqueta  $y_i^l$  (asignada al vértice  $i$  del nivel  $l$  de la pirámide) con su etiqueta hija  $y_k^{l-1}$ , y  $V_P(y_i^l, y_m^{l+1})$  que modela la relación de  $y_i^l$  con la etiqueta de su padre  $y_m^{l+1}$ . La etiqueta  $y_k^{l-1}$  fue asignada al vértice  $k$  (en el nivel  $l-1$  de la pirámide) por el MRF calculado en ese nivel. El vértice  $k$  pertenece al CK del vértice  $i$  ( $k \in CK(i)$ ). Los potenciales de asociación, interacción y jerárquicos son pesados

por  $\lambda_O$ ,  $\lambda_I$  y  $\lambda_H$  respectivamente, y  $\lambda_O + \lambda_I + \lambda_H = 1$ .

Una vez establecida la función de energía, el potencial de asociación  $V_O(y_i^l)$  se obtendrá mediante un clasificador base probabilístico, empleando la probabilidad *a posteriori*  $P_R(y_i^l|x_i^l)$  de una etiqueta dada una región, donde  $y_i^l \in Y, x_i^l \in X$ . El potencial de asociación es expresado según la Ecuación 4.4,

$$V_O(y_i^l) = \frac{1}{P_R(y_i^l|x_i^l)} \quad (4.4)$$

el potencial de interacción es definido en la Ecuación 4.5,

$$V_I(y_i^l, y_j^l) = \begin{cases} 0 & \text{if } y_i^l = y_j^l \\ 1 & \text{if } y_i^l \neq y_j^l \end{cases} \quad (4.5)$$

el potencial relacionado con la información de los hijos es definido en la Ecuación 4.6 y el potencial relacionado con la información del padre es presentado en la Ecuación 4.7.

$$V_{Ch}(y_i^l, y_k^{l-1}) = \begin{cases} 0 & \text{if } y_i^l = y_k^{l-1} \\ 1 & \text{if } y_i^l \neq y_k^{l-1} \end{cases} \quad (4.6) \quad V_P(y_i^l, y_m^{l+1}) = \begin{cases} 0 & \text{if } y_i^l = y_m^{l+1} \\ 1 & \text{if } y_i^l \neq y_m^{l+1} \end{cases} \quad (4.7)$$

donde  $k \in CK(i)$  e  $i \in CK(m)$ .

El potencial de interacción penaliza la configuración en que los vecinos de un vértice tengan etiquetas diferentes a dicho vértice, mientras que los potenciales jerárquicos penalizan las configuraciones en que los vértices hijos o padres tengan etiquetas distintas al vértice que se analiza.

Con el objetivo de encontrar la configuración óptima de etiquetas  $Y^*$ , se utiliza el algoritmo ICM, el cual es eficiente, y aunque usualmente ha sido criticado por converger a mínimos locales [Li 09], en este caso los resultados son muy similares a otros métodos más complejos. Una de las desventajas del algoritmo ICM es la necesidad de seleccionar un estado inicial adecuado [Jung 08]. En el algoritmo que se propone en este trabajo, la inicialización se hace utilizando un clasificador base que da una predicción inicial de las etiquetas de las regiones. Esta inicialización podría ser lo suficientemente buena como para ubicarse siempre en una posición cercana al óptimo global o al menos a un óptimo

local que se aproxime bastante al global. Teniendo esto en cuenta se podría pensar que el algoritmo ICM, en este caso, está convergiendo al mínimo global o a una solución cercana a este.

Este enfoque de re-etiquetado utilizando una jerarquía de MRFs fue nombrado HMRF-PYR y es detallado en forma de pseudocódigo en el Algoritmo 3, omitiendo los pasos enmarcados en rojo.

El algoritmo HMRF-PYR recibe como entrada la pirámide irregular que representa a una imagen, el nivel por el que se comenzará el proceso y el último nivel que intervendrá en el mismo ( $l_s$  y  $l_e$  respectivamente). Este proceso no se realiza para todos los niveles de la pirámide, pues los primeros niveles y los últimos serán niveles que presentan una sobre-segmentación y sub-segmentación extremas, lo cual entorpecería el proceso tanto en costo computacional como en introducción de segmentos ruidosos. Además se recibe como entrada el conjunto de rasgos  $\{X_{l_s}, X_{l_s+1} \dots X_{l_e}\}$  de cada nivel de la pirámide, los cuales se utilizarán para realizar la clasificación inicial de las regiones y para el cálculo de los potenciales de asociación.

Lo primero que se hace en este algoritmo es clasificar el nivel inicial  $l_s$  utilizando el clasificador base. Al comenzar este proceso, no se tiene información del etiquetado de los niveles superior o inferior de  $l_s$ , por lo que el primer re-etiquetado se realiza utilizando la Ecuación 4.3, pero omitiendo el potencial jerárquico.

Posteriormente se itera por todos los niveles, comenzando por el nivel  $l_s + 1$  hasta el nivel  $l_e$  realizando el mismo proceso: una clasificación inicial con el clasificador base y luego un re-etiquetado para corregir la primera clasificación. En este proceso que se realiza de abajo hacia arriba en cada nivel que se trabaja ya se conoce la información del etiquetado del nivel inferior, que fue re-etiquetado en una iteración anterior. Por esta razón en esta etapa, para resolver cada MRF, se utiliza la Ecuación 4.3 con el potencial jerárquico, pero solo incluyendo el término relacionado con la información de los hijos ( $V_{Ch}$ ).

Una vez que se alcanza el nivel superior, se realiza el mismo proceso de arriba hacia abajo, ya contando con la información de etiquetado del nivel padre y del nivel hijo de cada nivel analizado. Se vuelve a reconsiderar el etiquetado que se tiene hasta el momento resolviendo un nuevo MRF por nivel mediante la Ecuación 4.3 usando el potencial jerárquico completo, con los términos  $V_{Ch}$  y  $V_P$ .

El resultado del algoritmo es el conjunto de etiquetas  $\{Y_{l_s}^*, Y_{l_s+1}^* \dots Y_{l_e}^*\}$  que corresponden a las configuraciones más probables de etiquetas en cada nivel de la pirámide.

## 4.4. Re-etiquetado y segmentación simultáneos

El enfoque presentado en la sección 4.3 está limitado en cuanto a la eficacia del reconocimiento de los objetos debido a la segmentación subyacente. En la plataforma de pirámides irregulares que se utiliza, el criterio para decidir si dos regiones serán unidas en el nuevo nivel se basa únicamente en la similitud entre los colores medios de cada región. El color medio de una región es un rasgo que se vuelve menos discriminativo al hacerse las regiones más grandes, por lo que se considera que la combinación de rasgos de bajo nivel con la información semántica que se obtiene en el proceso de reconocimiento puede mejorar la segmentación de la imagen, y por consiguiente los resultados de etiquetado finales.

Para esta tarea se propone en esta tesis modificar el criterio empleado para crear los CK mediante el uso de la información de la clasificación en cada nivel y la información de bordes extraída de la imagen. Este enfoque fue previamente publicado por la autora de la tesis en [Morales-González 13b]. Se propone calcular un valor  $V_{contract}$  con el que se etiquetará cada arista de cada RAG de la pirámide. Este valor será una combinación de una medida semántica  $V_S$  y una medida de nivel bajo  $V_B$ .

Para calcular el valor semántico  $V_S(i, j)$ , el cual puede ser entendido como una atracción semántica entre los vértices  $v_i$  y  $v_j$ , se utiliza la información obtenida de las clases con que se etiquetó cada región y la probabilidad a priori dada por el clasificador empleado. Por cada vértice  $v_i$ , después del proceso de clasificación (y re-etiquetado usando el MRF) se tiene la siguiente información:

- Una clase  $C_i^{MRF}$  asignada al vértice  $v_i$  después que se halló la configuración más probable del MRF,
- La probabilidad a priori obtenida por el clasificador base para esta clase en este vértice  $P(C_i^{MRF})$
- Una lista de las  $n$  clases  $[C_{i,1}^{CB}, C_{i,2}^{CB}, \dots, C_{i,n}^{CB}]$ , ordenadas por la probabilidad a priori de cada clase para representar al vértice  $v_i$ , obtenida con el clasificador base (CB). De esta forma, se puede notar que la clase  $C_{i,1}^{CB}$  fue con la que el CB clasificó finalmente el vértice  $v_i$ .
- Una lista de todas las probabilidades a priori  $[P(C_{i,1}^{CB}), P(C_{i,2}^{CB}), \dots, P(C_{i,n}^{CB})]$  obtenidas por el CB indicando la probabilidad de que  $v_i$  pertenezca a cada clase.

Lo primero que se debe hacer es verificar si las clases anotadas para  $v_i$  y  $v_j$  son las mismas. Si este es el caso, el valor de  $V_S(i, j)$  es la suma de las probabilidades obtenidas por el CB para estas clases. Si las clases son distintas, existe la posibilidad de que el clasificador base haya cometido un error de clasificación, por tanto, es necesario chequear la confianza con que estas clases fueron asignadas a estos vértices. Se define la confianza de la clasificación como un valor lógico (verdadero/falso) dado por la Ecuación 4.8.

$$Confidence(C_i^{MRF}) = [(P(C_{i,1}^{CB}) - P(C_{i,2}^{CB})) > \delta] \quad (4.8)$$

Se considera que hay confianza en la clasificación de un vértice  $v_i$  con la clase  $C_i^{MRF}$  si la diferencia entre las dos probabilidades más altas asignadas por el CB a este vértice es mayor que un umbral  $\delta$ . Si hay confianza en la clasificación de ambos vértices (ya teniendo en cuenta que sus clases son diferentes), el valor  $V_S(i, j)$  será -1, indicando que estos vértices no deben ser unidos semánticamente. Pero si la clasificación para uno de los vértices no tiene confianza, se verifica si la primera o segunda clase asignada a él con mayores probabilidades corresponden con la clase del otro vértice. Si esto ocurre, se suman las probabilidades de ambas clases (las que resultaron iguales) para obtener  $V_S(i, j)$ . Este paso se muestra en la Ecuación 4.9, donde se cambia la nomenclatura haciendo  $V_S(i, j) = MisclassValue(i, j)$  para una mejor comprensión de lo que significa dicho valor en este caso en que se considera un error de clasificación.

$$MisclassValue(i, j) = \begin{cases} P(C_i^{MRF}) + P(C_{j,1}^{CB}) & \text{if } C_i^{MRF} = C_{j,1}^{CB} \\ P(C_i^{MRF}) + P(C_{j,2}^{CB}) & \text{if } C_i^{MRF} = C_{j,2}^{CB} \\ -1 & \text{en otro caso} \end{cases} \quad (4.9)$$

El proceso para calcular  $V_S(i, j)$  se resume en la Ecuación 4.10:

$$V_S(i, j) = \begin{cases} P(C_i^{MRF}) + P(C_j^{MRF}) & \text{if } C_i^{MRF} = C_j^{MRF} \\ MisclassValue(i, j) & \text{if } Confidence(C_i^{MRF}) = 0 \\ MisclassValue(j, i) & \text{if } Confidence(C_j^{MRF}) = 0 \\ -1 & \text{en otro caso} \end{cases} \quad (4.10)$$



Basado en la explicación anterior, se puede notar que intuitivamente, el valor  $V_S(i, j)$  representa la probabilidad de que dos regiones adyacentes pertenezcan a la misma clase y, por tanto, la posibilidad de que sean unidas en función de esto.

Por otro lado, el valor  $V_B(i, j)$  representa intuitivamente la probabilidad de unir dos vértices  $v_i$  y  $v_j$  tomando en cuenta la información de bordes cuando son dos regiones separadas y cuando son combinados en una sola región. Para esto se utiliza el detector de bordes de Canny [Canny 86] para extraer los bordes de la imagen. La máscara de bordes resultante se utilizará para evaluar la conveniencia de unir dos regiones adyacentes.

Se llamará  $B_{Canny}$  al conjunto de píxeles que representan bordes en la máscara de Canny. El conjunto de píxeles de bordes correspondiente al contorno del campo receptivo (CR) del vértice  $v_i$  será nombrado  $B_i$  y el conjunto de píxeles de bordes resultantes al unir los CRs de  $v_i$  y  $v_j$  se denominará  $B_{i \cup j}$ .

En primer lugar se calcula cuántos píxeles de  $B_{i \cup j}$  se corresponden con los píxeles de  $B_{Canny}$  (Ecuación 4.11), y luego se halla la cantidad de píxeles en la intersección entre  $B_{Canny}$  y la unión de  $B_i$  y  $B_j$  (Ecuación 4.12).

$$B_1(i, j) = |B_{i \cup j} \cap B_{Canny}| \quad (4.11)$$

$$B_2(i, j) = |(B_i \cup B_j) \cap B_{Canny}| \quad (4.12)$$

Finalmente, se propone calcular  $V_B(i, j)$  como se muestra en la Ecuación 4.13. En este caso, se puede notar que si  $B_2(i, j) > B_1(i, j)$ , existe un borde entre las regiones de  $v_i$  y  $v_j$  que está presente en la máscara de bordes de Canny y que sería eliminado si estas dos regiones se unieran. Esto no es deseable pues este es un borde que es necesario preservar, por tanto, en este caso el valor de  $V_B(i, j)$  es -1, invalidando la contracción de estos dos vértices. Si esta condición inicial no se cumple, el valor de  $V_B(i, j)$  es la relación entre  $B_{i \cup j}$  y la intersección de  $B_{i \cup j}$  con  $B_{Canny}$ . Esto es un indicador de cuántos píxeles de bordes de las regiones de  $v_i$  y  $v_j$  unidas se corresponden con los de la máscara de Canny, con respecto al total de píxeles de la unión. Si todos los píxeles de bordes de la unión de las dos regiones están presentes en la máscara de Canny, el valor de  $V_B(i, j)$  será 1.

$$V_B(i, j) = \begin{cases} -1 & \text{if } B_2(i, j) > B_1(i, j) \\ \frac{B_1(i, j)}{|B_{i \cup j}|} & \text{en caso contrario} \end{cases} \quad (4.13)$$

Una vez que se tienen  $V_S(i, j)$  y  $V_B(i, j)$ , se puede calcular  $V_{contract}(i, j)$  como se muestra en la Ecuación 4.14.

$$V_{contract}(i, j) = \begin{cases} 0 & \text{if } V_S(i, j) = -1 \text{ o } V_B(i, j) = -1 \\ V_S(i, j) + V_B(i, j) & \text{en caso contrario} \end{cases} \quad (4.14)$$

Cuando se han calculado los valores  $V_{contract}$  para todas las aristas en el grafo, cada vértice seleccionado para sobrevivir en el próximo nivel utilizará esta información para crear su CK, i.e, cuáles de sus vértices adyacentes tienen más posibilidades de unirse con él. Los valores mayores de  $V_{contract}$  corresponden a las aristas con más condiciones para ser contraídas, dada la información semántica y de bordes utilizada. Si  $V_{contract}$  es 0 para una arista, esta nunca será contraída, ya sea porque los vértices que conecta pertenecen a clases semánticas diferentes o porque existe un borde entre las regiones subyacentes que es necesario preservar. El proceso de combinación es ilustrado en la Figura 4.3.

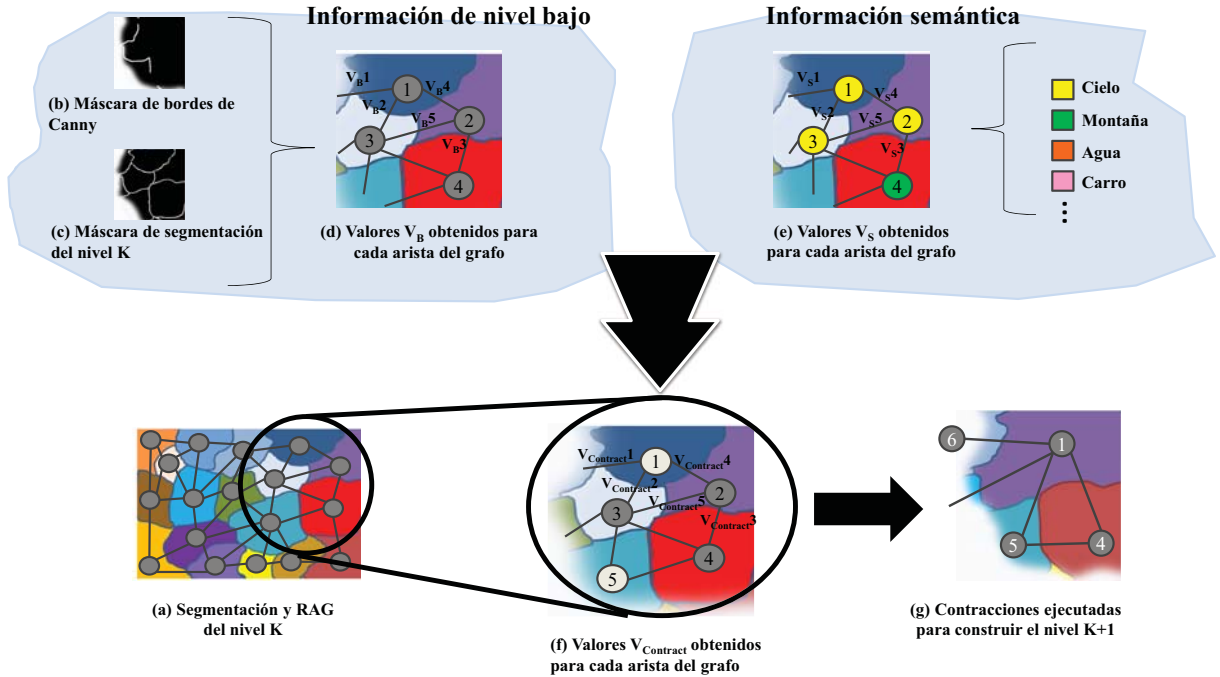


Figura 4.3: Combinación de la información de bajo nivel con la información semántica para construir un nuevo nivel de segmentación. En (f) los vértices blancos son los que sobrevivirán al nivel siguiente, y estos usan los valores  $V_{Contract}$  de sus aristas para determinar qué vértice no sobreviviente se unirá a él.

Usando este criterio de contracción se construye un nivel nuevo de la pirámide a partir

de uno existente. Esto es parte de un proceso iterativo donde un nivel es etiquetado inicialmente con un clasificador base, luego esta clasificación es refinada resolviendo su MRF asociado y finalmente, los nuevos CK son seleccionados utilizando la información de la clasificación y la información de bordes, dando lugar a un nuevo nivel de segmentación. Todo el proceso se vuelve a repetir hasta que se alcance un nivel donde no se permitan más contracciones. Este proceso fue denominado HMRF-PYRSEG (se muestra en pseudocódigo en el Algoritmo 3), el cual es una extensión del algoritmo HMRF-PYR .

## 4.5. Experimentos

Para validar esta propuesta de re-etiquetado de imágenes, se corrieron experimentos en un subconjunto de la colección de imágenes Corel. Específicamente, se utilizó el subconjunto CorelA seleccionado por [Carbonetto 03]. Este subconjunto contiene 205 imágenes de escenas naturales divididas en 2 subconjuntos. Uno contiene 137 imágenes para entrenar y el otro 68 para realizar pruebas. Todas las imágenes fueron segmentadas usando el algoritmo *Normalized Cuts* [Shi 97] y fueron etiquetadas manualmente con 22 clases semánticas.

Las pirámides irregulares calculadas para estas imágenes tienen un promedio de 20 niveles. Para los experimentos se utilizaron los niveles desde el 10 hasta el 16, de forma que se eviten sobre-segmentaciones y sub-segmentaciones extremas. En la Sección 4.6 se mostrará como influye la cantidad de niveles a utilizar en función de la eficacia y el tiempo.

Para estos experimentos se utilizaron los rasgos visuales descritos en el Epígrafe 2.2. Cuando se utilicen los rasgos de color y LBP (ver subepígrafe 2.2.1) se les denominará cLBP y cuando se utilicen los rasgos contextuales (ver subepígrafe 2.2.2) se les nombrará RCF.

Dado que el objetivo de los algoritmos HMRF-PYR y HMRF-PYRSEG es mejorar el etiquetado de las regiones de una imagen a partir de una clasificación inicial de las mismas, se utilizaron tres clasificadores base para mostrar dicha característica de los algoritmos propuestos en esta tesis. Los clasificadores base utilizados fueron el *Naïve Bayes* (NB), el *Random Forest* (RF) [Breiman 01] y las Máquinas de Soporte Vectorial (SVM) [Perronnin 12].

Fueron empleadas dos medidas para evaluar el desempeño de los algoritmos propuestos. La primera es la eficacia del reconocimiento, la cual mide la eficacia del etiquetado a

---

**Algorithm 3:** Algoritmo HMRF-Pyr (quitando los pasos enmarcados en rojo) y HMRF-PyrSeg (empleando los pasos enmarcados en rojo).

---

```

input : Grafo correspondiente al nivel inicial del proceso  $G_{l_s}$ ;
         nivel inicial  $l_s$ , nivel final  $l_e$ ;
         conjuntos de rasgos de cada nivel  $\{X_{l_s}, X_{l_s+1} \dots X_{l_e}\}$ ;
         máscara de bordes de Canny  $I_{Canny}$ 
output: Conjunto de configuraciones más probables de etiquetas de cada nivel
          $\{Y_{l_s}^*, Y_{l_s+1}^* \dots Y_{l_e}^*\}$ 

// Clasificar el primer nivel  $l_s$ 
1  $l \leftarrow l_s$ ;
2  $Y_l \leftarrow$  Clasificar  $X_l$  del nivel  $l$  con el clasificador base;
3  $Y_l^* \leftarrow$  Resolver el MRF para el nivel  $l_s$  usando ICM para minimizar la Ecuación
   4.3 (solo los términos  $V_O$  y  $V_I$ )
4  $G_{l+1} \leftarrow \text{CrearNuevoNivel}(l, G_l, Y_l^*, I_{Canny});$ 

// Iterar por los niveles de la pirámide de abajo hacia arriba
5 for  $l \leftarrow l_s + 1$  to  $l_e$  do
6    $Y_l \leftarrow$  Clasificar  $X_l$  del nivel  $l$  (grafo  $G_l$ ) con el clasificador base;
7    $Y_l^* \leftarrow$  Resolver el MRF para el nivel  $l$  usando ICM para minimizar la
   Ecuación 4.3 (solo los términos  $V_O$ ,  $V_I$  y  $V_{Ch}$ );
8    $G_{l+1} \leftarrow \text{CrearNuevoNivel}(l, G_l, Y_l^*, I_{Canny});$ 
9 end

// Iterar por los niveles de la pirámide de arriba hacia abajo
10 for  $l \leftarrow l_e - 1$  down to  $l_s$  do
11    $Y_l \leftarrow$  Clasificar  $X_l$  del nivel  $l$  (grafo  $G_l$ ) con el clasificador base;
12    $Y_l^* \leftarrow$  Resolver el MRF para el nivel  $l$  usando ICM para minimizar la
   Ecuación 4.3 con todos sus términos;
13 end
14 Procedure  $\text{CrearNuevoNivel}(l, G_l, Y_l^*, I_{Canny})$ 
15   Calcular  $V_B \forall e \in E$  del grafo  $G_l(V, E)$  usando  $I_{Canny}$  (Ecuación 4.13);
   Calcular  $V_S \forall e \in E$  del grafo  $G_l(V, E)$  usando  $Y_l^*$  (Ecuación 4.10);
   Calcular  $V_{contract} \forall e \in E$  del grafo  $G_l(V, E)$  (Ecuación 4.14);
   Seleccionar nuevos CKs para todos los vértices sobrevivientes en  $G_l(V, E)$ 
   usando los valores  $V_{contract}$ ;
   return  $G_{l+1} \leftarrow$  Ejecutar contracciones para obtener nuevo nivel;

```

---

nivel de píxel. La segunda medida es la Medida F (F-measure en inglés), que no es más que la media armónica entre la precisión y el recuerdo. La precisión se calcula como  $Prec = VP/(VP + FP)$  donde  $VP$  son los verdaderos positivos, es decir, los píxeles etiquetados correctamente con la clase  $c$  de acuerdo con el ground truth y  $FP$  son los falsos positivos, es decir, los píxeles etiquetados incorrectamente como  $c$ . El recuerdo se obtiene como  $Rec = VP/(VP + FN)$  donde  $FN$  son los falsos negativos, es decir, los píxeles que debieron etiquetarse con la clase  $c$  según el ground truth y no lo fueron. En resumen, la precisión mide qué fracción de lo clasificado fue correcta y el recuerdo expresa cuántos elementos, de los que debían ser encontrados, se clasificaron correctamente. Finalmente, la medida F se calcula como  $F = 2 * (Prec * Rec)/(Prec + Rec)$ . Según esta medida, mientras mayor es el valor F, mejor es el resultado de la clasificación.

Los resultados de eficacia de reconocimiento y de la medida F pueden observarse en las tablas 4.1 y 4.2 respectivamente. En estas tablas, las filas representan los algoritmos que se comparan, empleando primero los rasgos visuales de color y LBP, con lo cual se añade el prefijo cLBP a cada algoritmo, y posteriormente los rasgos de contexto, por lo que se incorpora el prefijo RCF a cada algoritmo. Los algoritmos cLBP-base y RCF-base se refieren a la clasificación de las regiones individuales utilizando únicamente estos rasgos visuales. Por las columnas se muestran los clasificadores base empleados en cada caso.

Tabla 4.1: Resultados obtenidos en el subconjunto CorelA en cuanto a eficacia de reconocimiento. Las filas representan los métodos evaluados mientras que las columnas representan los clasificadores base utilizados en cada caso.

Rasgos visuales-Algoritmo	Clasificador base		
	NB	RF	SVM
cLBP-base	15.9 %	38.8 %	35.4 %
cLBP-HMRF-PYR	15.7 %	40.4 %	36.0 %
cLBP-HMRF-PYRSEG	<b>19.1 %</b>	<b>48.7 %</b>	<b>42.4 %</b>
RCF-base	29.8 %	48.5 %	42.8 %
RCF-HMRF-PYR	29.9 %	50.1 %	43.0 %
RCF-HMRF-PYRSEG	<b>30.0 %</b>	<b>51.7 %</b>	<b>44.8 %</b>

En estas tablas se muestra que los algoritmos propuestos mejoran en todos los casos al clasificador base empleado. Particularmente, el algoritmo HMRF-PYRSEG exhibe un mejor desempeño que HMRF-PYR, lo que indica que la mejora en la segmentación incide positivamente en el resultado del reconocimiento.

Tabla 4.2: Resultados obtenidos en el subconjunto CorelA en cuanto a la medida F. Las filas representan los métodos evaluados mientras que las columnas representan los clasificadores base utilizados en cada caso.

Rasgos visuales-Algoritmo	Clasificador base		
	NB	RF	SVM
cLBP-base	0.225	0.261	0.149
cLBP-HMRF-PYR	0.203	0.263	0.150
cLBP-HMRF-PYRSEG	<b>0.239</b>	<b>0.333</b>	<b>0.163</b>
RCF-base	0.382	0.278	0.216
RCF-HMRF-PYR	<b>0.383</b>	0.275	0.215
RCF-HMRF-PYRSEG	0.368	<b>0.294</b>	<b>0.228</b>

Dado que para cada combinación de descripción visual y clasificadores base en los algoritmos HMRF-PYR y HMRF-PYRSEG se obtienen resultados para distintos valores de  $\lambda_I$  (potencial de interacción) y  $\lambda_H$  (potencial jerárquico), en las tablas 4.1 y 4.2 se muestran, para cada algoritmo, los mejores resultados obtenidos de entre todas las combinaciones de parámetros  $\lambda$  probados. Los coeficientes  $\lambda$  correspondientes a estos resultados se muestran en la Tabla 4.3.

Tabla 4.3: Parámetros con los que se obtuvieron los mejores resultados para cada combinación de descripción visual y clasificador base

Rasgos visuales - clasificador base	$\lambda_I$	$\lambda_H$
cLBP - NB	0.25	0.75
RCF - NB	0.75	0.25
cLBP - RF	0.25	0.25
RCF - RF	0.25	0.25
cLBP - SVM	0.5	0.25
RCF - SVM	0	0.25

Recordando que  $\lambda_O + \lambda_I + \lambda_H = 1$ , se puede notar en la Tabla 4.3 que los mejores resultados se obtienen en la mayoría de los casos con valores de  $\lambda_I \neq 0$  y  $\lambda_H \neq 0$ , las cuales representan relaciones espaciales y jerárquicas respectivamente. Esto muestra la importancia que juegan en este enfoque dichas relaciones y cómo funcionan como complemento de los rasgos que describen la apariencia de las regiones. Es importante notar, que cuando se utiliza el clasificador base *Random Forest*, coinciden los mejores

resultados para  $\lambda_I = \lambda_H = 0,25$ . De hecho, con los otros clasificadores base los resultados de esta configuración están muy cercanos a los mostrados en la Tabla 4.3. Es por esto que se recomienda emplear estos valores de  $\lambda_I$  y  $\lambda_H$  en futuras aplicaciones.

Para hacer un análisis de cómo se comportan los valores de eficacia del reconocimiento para los distintos niveles de la pirámide, es importante notar que, para el caso del algoritmo HMRF-PyrSeg, no se utilizan todos los niveles construidos inicialmente en la pirámide, sino que se comienza a partir de un nivel dado y se crean nuevos niveles de segmentación de la pirámide utilizando información semántica y de bordes. Se considera que la sobre-segmentación de una imagen es suficiente para realizar la clasificación de pequeños objetos o partes de objetos, por lo que se decidió comenzar el proceso de crear nuevos niveles a partir del nivel 10 de la pirámide original.

En el nivel 10 se etiquetan todas las regiones con el clasificador base, se calcula el MRF asociado a este nivel y se determinan los nuevos CKs para construir un nuevo nivel 11. Este proceso se repite hasta que se alcanza el nivel 20 para todas las pirámides.

Tabla 4.4: Resultados de eficacia obtenidos en el subconjunto CorelA para distintos niveles de la pirámide.

Algoritmo	Niveles de la pirámide						
	10	11	12	13	14	15	16
RF	47.9 %	48.4 %	47.7 %	46.3 %	45.4 %	44.7 %	44.4 %
HMRF-PYR	49.9 %	50.1 %	48.5 %	48.3 %	47.8 %	47.1 %	46.1 %
HMRF-PYRSEG	<b>51.7 %</b>	<b>51.2 %</b>	<b>50.1 %</b>	<b>49.7 %</b>	<b>49.0 %</b>	<b>48.0 %</b>	<b>46.9 %</b>

En la Tabla 4.4 se puede observar la comparación de los resultados de eficacia de reconocimiento obtenidos usando el clasificador base RF (pues fue el clasificador base con mejores resultados de eficacia), los resultados del algoritmo HMRF-PYR (el cual mantiene la segmentación original de la pirámide) y los resultados de HMRF-PYRSEG, que utiliza el mismo sistema de etiquetado que HMRF-PYR, pero mejora la segmentación mediante la creación de nuevos niveles. Esta comparación se puede observar gráficamente en la Figura 4.4.

Los algoritmos HMRF-PYR y HMRF-PYRSEG fueron comparados con otros métodos del estado del arte que han sido probados en este subconjunto de imágenes. Los resultados, en términos de eficacia global de reconocimiento, pueden observarse en la Tabla 4.5. No fue posible hacer una comparación con otros métodos en cuanto a la medida F, pues no

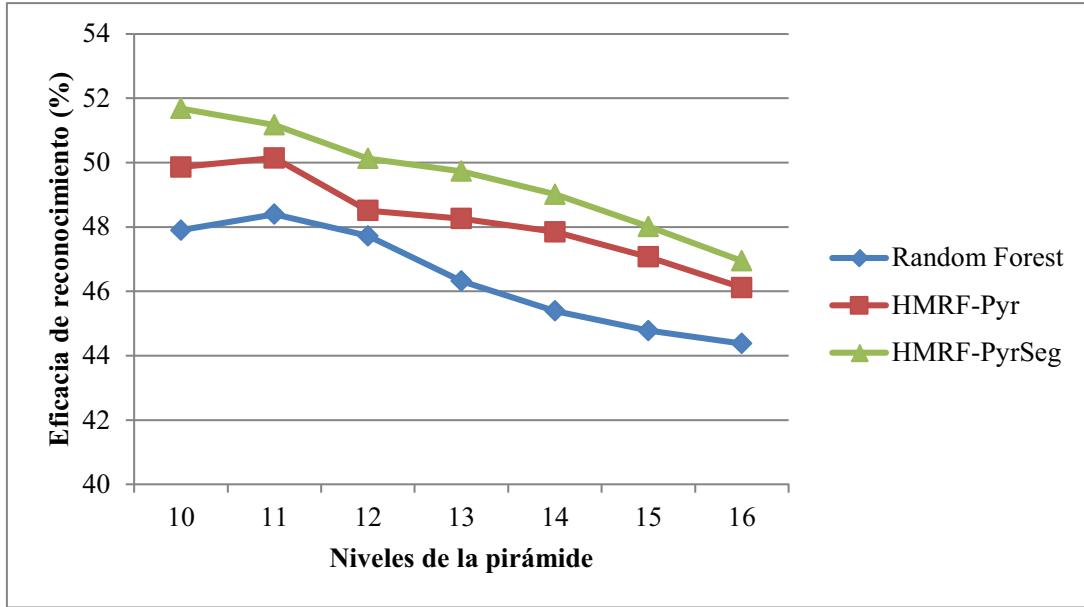


Figura 4.4: En esta figura se muestran los resultados de eficacia del clasificador base *Random Forest*, el algoritmo HMRF-PYR y el HMRF-PYRSEG.

se han encontrado resultados reportados en la literatura utilizando esta medida para este tipo de tarea específicamente.

Tabla 4.5: Comparación de HMRF-Pyr y HMRF-PyrSeg con otros algoritmos en el subconjunto CorelA.

Algoritmo	Eficacia global
gML1o [Carbonetto 03]	36.2 %
MRFs AREK [Hernández-Gracidas 07]	45.6 %
HMRF-Pyr	50.1 %
<b>HMRF-PyrSeg</b>	<b>51.7 %</b>

Para ilustrar la mejora en cuanto a segmentación y etiquetado de las imágenes, en la Figura 4.5 se muestra el mejor nivel segmentado para imágenes de ejemplo utilizando HMRF-PYR y HMRF-PYRSEG . Con estos resultados se puede notar la relevancia de tener una mejor segmentación subyacente a la hora de realizar el proceso de reconocimiento y cómo estos dos procesos se pueden combinar para aprovechar los resultados individuales de cada uno y obtener un mejor resultado final.



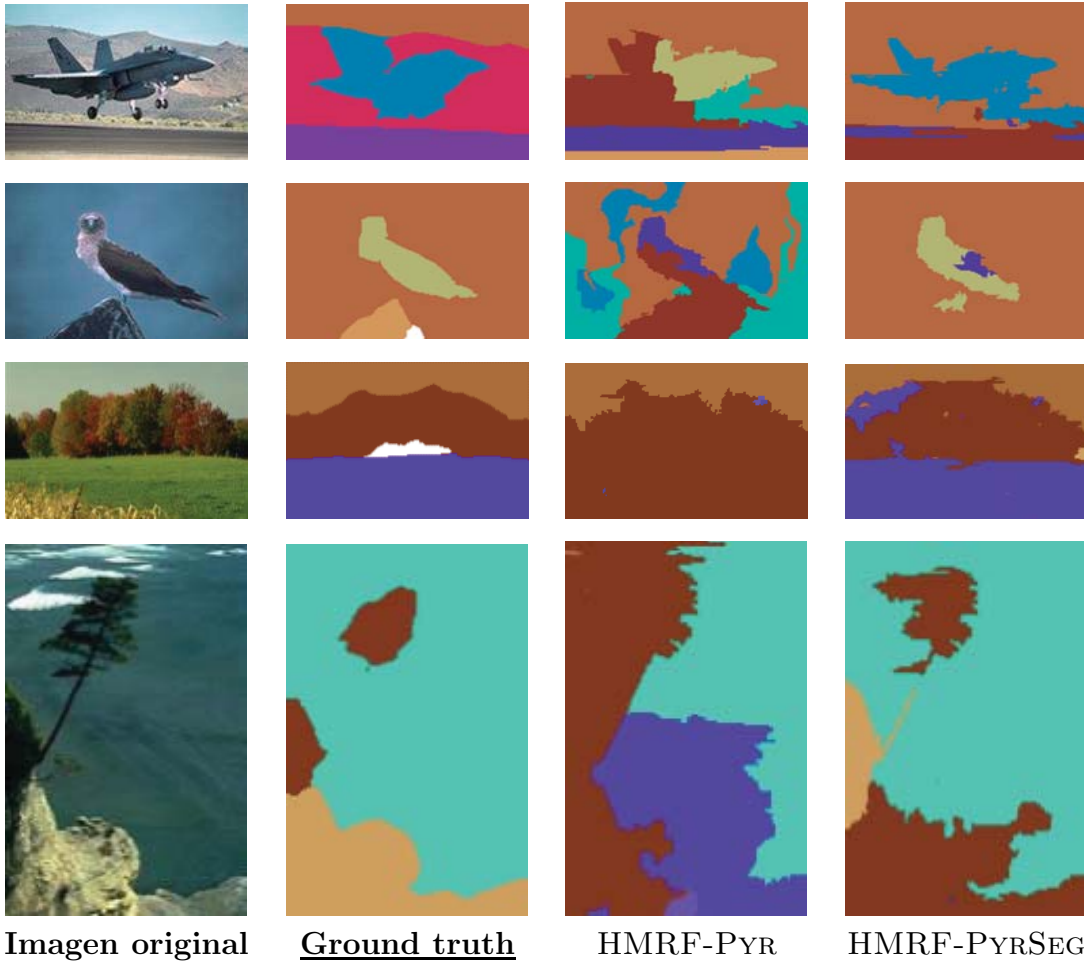


Figura 4.5: Segmentaciones de ejemplo utilizando HMRF-PYR y HMRF-PYRSEG . Los colores representan clases semánticas.

## 4.6. Complejidad computacional de HMRF-Pyr

La complejidad computacional del algoritmo ICM es  $O(nq)$ , donde  $n$  es la cantidad de vértices del grafo y  $q$  es la cantidad de etiquetas a asignar [Jung 08]. La función de energía es posible evaluarla en  $O(1)$ . La cantidad de etiquetas  $q$  es fija para este tipo de problemas por lo que la complejidad de ICM puede quedar como  $O(n)$ . Teniendo en cuenta que en HMRF-PYR este algoritmo se ejecuta por cada nivel de la pirámide, entonces la complejidad final de HMRF-PYR sería  $O(nl)$ , donde  $l$  es la cantidad de niveles. A medida que los niveles aumentan en la jerarquía, la cantidad de vértices disminuye, según lo mostrado en la Sección 2.4, por lo que puede ser complicado entender la relación entre la cantidad de niveles, el tiempo del algoritmo y cuánto esto representa en eficacia.

Para ejemplificar cómo varía el tiempo en función de la cantidad de niveles empleados, se muestra la Figura 4.6. Aquí se observan 10 experimentos realizados solo con 20 imágenes de prueba de la colección Corel. El eje  $x$  representa cada experimento y muestra la cantidad de niveles empleados en él. Esto quiere decir que se realizó un experimento con un solo nivel, otro experimento empleando 2 niveles de la jerarquía, luego 3 niveles y así sucesivamente hasta 10. En el eje  $y$  se muestra el tiempo promedio que tomó etiquetar cada imagen y en cada punto se puede ver el valor de eficacia alcanzado en dicho experimento.

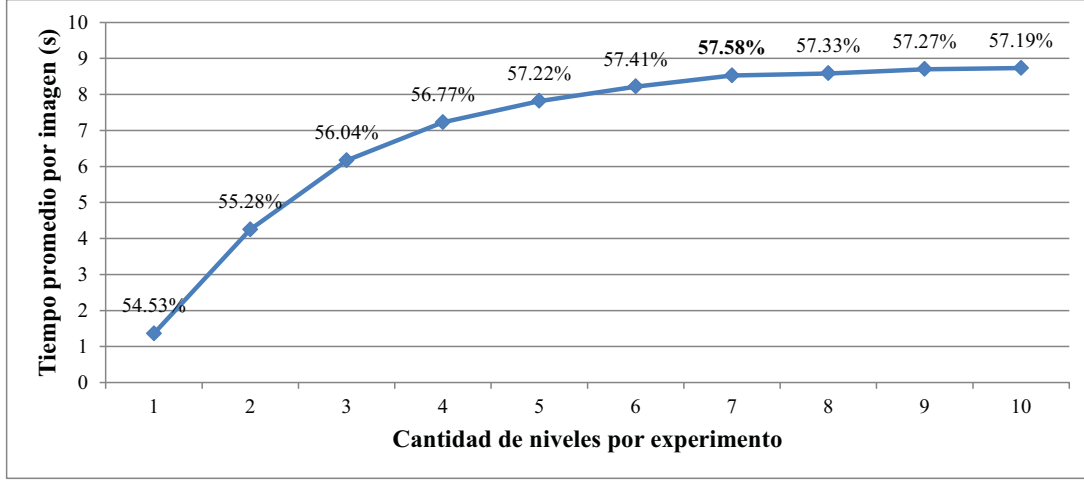


Figura 4.6: Se muestran 10 experimentos (eje  $x$ ) con HMRF-PYR, donde para cada experimento se emplearon distintas cantidades de niveles, desde 1 hasta 10. En el eje  $y$  se muestra el tiempo promedio que tomó etiquetar cada imagen en cada experimento. El dato que se muestra sobre cada punto es el valor de eficacia alcanzado en cada experimento.

Esta gráfica muestra un incremento logarítmico en el tiempo, lo que viene a sustentar la relación logarítmica que existe entre la cantidad de vértices de los niveles de la pirámide (altura logarítmica), mencionada en la Sección 2.4. También se puede observar el comportamiento de la eficacia, que va aumentando hasta alcanzar su máximo cuando se emplean 7 niveles, y a partir de ahí comienza a decaer. Esto indica que los niveles superiores son más ruidosos, y afectan la corrección del etiquetado de los niveles inferiores.

Como es de esperar, existe un compromiso entre la eficiencia y la eficacia, por lo que en problemas donde se priorice la rapidez, deberán escogerse menos niveles.

## 4.7. Conclusiones parciales

Al utilizar la representación jerárquica propuesta que incluye relaciones espaciales, se logró superar los resultados de eficacia en el etiquetado de regiones de una imagen. En los

experimentos realizados se puede ver que los mejores resultados se obtuvieron con pesos positivos para los potenciales de interacción (relaciones espaciales) y jerárquico, indicando que el desempeño logrado se debe a la combinación de estos dos factores con la descripción visual de las regiones.

Con el algoritmo HMRF-PYRSEG se mostró que la segmentación subyacente afecta los resultados del reconocimiento de objetos. Al desarrollar un método que realiza la segmentación a partir de los resultados de un proceso de reconocimiento de objetos previo, se logró mejorar la segmentación de la pirámide, y por tanto, se obtuvieron mejores resultados en el reconocimiento final. Esto indica que la segmentación guiada por pistas semánticas (y no solo por rasgos de bajo nivel) es una estrategia muy prometedora en esta área de investigación. La eficacia de reconocimiento de este algoritmo fue de 51.7 %, que mejora en un 1.6 % a HMRF-PYR y en un 3.2 % al clasificador base. Comparado con otros algoritmos del estado del arte, HMRF-PYRSEG logró superarlos al menos en 6.1 % de eficacia de reconocimiento.

# CONCLUSIONES Y RECOMENDACIONES

## Conclusiones

Mediante la propuesta de representación de las imágenes combinando información visual, espacial y jerárquica utilizada en 3 tareas de reconocimiento de objetos con objetivos distintos cada una, se logró, en cada caso, mejorar la eficacia de reconocimiento con respecto a los métodos existentes en el estado del arte.

En particular, el empleo de una representación basada en pirámides irregulares de grafos permitió combinar apariencia, configuración espacial y jerarquía, de forma que los vértices de cada grafo reciben atributos de rasgos visuales y un nuevo descriptor espacial es utilizado para etiquetar las aristas. Para utilizar esta nueva representación se propuso una medida de similitud para hallar aproximaciones entre estructuras de este tipo.

Al desarrollar un nuevo método para el reconocimiento de clases de objetos en escenarios simples, basado en el enfoque de correspondencia de grafos y que explota la representación propuesta, se logró obtener un 90.2 % de eficacia global en el experimento de categorización, que superó en 2.6 % al mejor resultado reportado en la base de datos utilizada. Además, mediante experimentos adicionales se evaluó el beneficio de usar las relaciones espaciales en este enfoque.

Al desarrollar un nuevo método para el reconocimiento de objetos específicos en escenarios simples, basado en bolsa de subgrafos (obtenidos mediante la representación propuesta), se obtuvieron muy buenos resultados de eficacia en el experimento de identificación de objetos (99.4 %). Además, este nuevo esquema de clasificación explora un campo de investigación muy poco desarrollado actualmente que consiste en la combinación de técnicas de clasificación de imágenes con técnicas de minería de datos.

Mediante el desarrollo de un nuevo método para el reconocimiento de objetos en escenarios complejos, aprovechando la representación propuesta para introducir relaciones jerárquicas en el etiquetado de regiones utilizando campos aleatorios de Markov, fue posible obtener 51.7 % de eficacia, que supera en 4.5 % al mejor resultado presentado en la

colección. Al extender este método para ser más tolerante a los errores de la segmentación inicial, mediante la combinación de los procesos de reconocimiento y segmentación simultánea e iterativamente, se obtuvo una eficacia global en el reconocimiento de 51.7 %, con lo cual se logró mejorar en 6.1 % al mejor resultado en el conjunto de prueba utilizado. Como resultado adicional, se logró mejorar la segmentación inicial de la pirámide.

## **Recomendaciones**

Con los resultados obtenidos no se concluye el trabajo en esta temática. Del estudio realizado se derivan algunos trabajos futuros a modo de recomendaciones:

1. Mejorar la evaluación de los niveles de la pirámide irregular que participarán en los procesos de reconocimiento, ya que se está realizando esta evaluación solo usando como base los bordes de la imagen (heredando por tanto los problemas relacionados con la detección de bordes). Pistas visuales como la homogeneidad de color y textura de las regiones y el tamaño promedio de las mismas, pueden ser útiles en este paso. Además, se podría considerar, desde el punto de vista semántico, el resultado de la clasificación en el enfoque de etiquetado de regiones, donde una mayor probabilidad conjunta de la clasificación en un nivel podría ser un indicador de que los objetos quedan mejor representados en dicho nivel.
2. Para el método BOFAS-PYR, se recomienda evaluar otras técnicas de selección de los grafos aproximados frecuentes (FAS), por ejemplo, FAS discriminativos, FAS representativos, FAS contrastantes, etc. Estas técnicas, además de mejorar la eficacia de la clasificación, podrían reducir la dimensionalidad de los rasgos finales.
3. Para todos los métodos propuestos, se recomienda extenderlos para ser utilizados en video, añadiendo información temporal a los esquemas propuestos. Además, sería conveniente analizar cómo trasladar y modificar una pirámide combinatoria a través de cuadros consecutivos del video, de manera que no haya que construirla completa en cada cuadro.

## REFERENCIAS BIBLIOGRÁFICAS

- [Acosta-Mendoza 12a] Niusvel Acosta-Mendoza, Andrés Gago Alonso, and José E. Medina-Pagola. “Frequent approximate subgraphs as features for graph-based image classification.”. *Knowl.-Based Syst.*, 27:381–392, 2012.
- [Acosta-Mendoza 12b] Niusvel Acosta-Mendoza, Annette Morales-González, Andrés Gago Alonso, Edel B. García Reyes, and José E. Medina-Pagola. “Image classification using frequent approximate subgraphs.”. In *Iberoamerican Congress on Pattern Recognition (CIARP) (7441) of Lecture Notes in Computer Science*, 292–299. Springer, 2012.
- [Akçay 07] Huseyin Gokhan Akçay and Selim Aksoy. “Automated detection of objects using multiple hierarchical segmentations.”. In *IGARSS*, 1468–1471. IEEE, 2007.
- [Antúnez 12] Esther Antúnez, Yll Haxhimusa, Rebeca Marfil and Walter G. Kropatsch, and Antonio Bandera. “Artificial visual attention using combinatorial pyramids”. In *Robotics and Vision: Technologies for Machine Learning and Vision Applications* (Jose García Rodríguez and Miguel Cazorla, eds.), 439–457. IGI Global, 2012.
- [Antúnez 13] Esther Antúnez, Rebeca Marfil, Juan P. Bandera, and Antonio Bandera. “Part-based object detection into a hierarchy of image segmentations combining color and topology”. *Pattern Recogn. Lett.*, 34(7):744–753, May 2013.
- [Arbelaez 12] Pablo Arbelaez, Bharath Hariharan, Chunhui Gu, Saurabh Gupta, Lubomir D. Bourdev, and Jitendra Malik. “Semantic

- segmentation using regions and parts.”. In *CVPR*, 3378–3385. IEEE, 2012.
- [Arif 09] Thawar Arif, Ziad Shaaban, Lala Krekor, and Sami Baba. “Object classification via geometrical, zernike and legendre moments”. *Journal of Theoretical and Applied Information Technology*, 7(1):31–37, 2009.
- [Aydemir 11] Alper Aydemir, Kristoffer Sjöö, John Folkesson, Andrzej Pronobis, and Patric Jensfelt. “Search in the real world: Active visual object search based on spatial relations.”. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2011, 2818–2824. IEEE, 2011.
- [Bay 08] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. “Speeded-up robust features (surf)”. *Comput. Vis. Image Underst.*, 110(3):346–359, June 2008.
- [Biederman 72] Irving Biederman. “Perceiving real-world scenes”. *Science*, 177:77–80, 1972.
- [Breiman 01] Leo Breiman. “Random forests”. *Mach. Learn.*, 45(1):5–32, October 2001.
- [Brun 01] L. Brun and W. Kropatsch. “Introduction to combinatorial pyramids”. In *Digital and image geometry: advanced lectures* (2243) of *LNCS*, 108–128. Springer-Verlag New York, Inc., 2001.
- [Brun 03] Luc Brun and Walter Kropatsch. “Contraction kernels and combinatorial maps”. *Pattern Recogn. Lett.*, 24(8):1051–1057, 2003.
- [Brun 06] L. Brun and W. Kropatsch. “Contains and inside relationships within combinatorial pyramids”. *Pattern Recogn.*, 39(4):515–526, 2006.
- [Brun 08] Luc Brun and Jean-Hugues Pruvot. “Hierarchical matching using combinatorial pyramid framework.”. In *ICISP* (Abderrahim Elmoataz, Olivier Lezoray, Fathallah Nouboud,

- and Driss Mammass, eds.) (5099) of *Lecture Notes in Computer Science*, 346–355. Springer, 2008.
- [Bunke 08] Horst Bunke and Kaspar Riesen. “Graph classification based on dissimilarity space embedding”. In *Proceedings of the 2008 Joint IAPR International Workshop on Structural, Syntactic, and Statistical Pattern Recognition*, SSPR & SPR ’08, 996–1007, Berlin, Heidelberg, 2008. Springer-Verlag.
- [Canny 86] J Canny. “A computational approach to edge detection”. *IEEE Trans. Pattern Anal. Mach. Intell.*, 8(6):679–698, June 1986.
- [Cao 11] Y. Cao, Y. Luo, and S. Yang. “Image denoising based on hierarchical markov random field”. *Pattern Recogn. Lett.*, 32(2):368–374, January 2011.
- [Carbonetto 03] P. Carbonetto. “Unsupervised statistical models for general object recognition”. tech. report, The Faculty of Graduate Studies, Department of Computer Science, The University of British Columbia, West Mall Vancouver, BC Canada, 2003.
- [Caruana 08] Rich Caruana, Nikos Karampatziakis, and Ainur Yessenalina. “An empirical evaluation of supervised learning in high dimensions”. In *Proceedings of the 25th International Conference on Machine Learning*, ICML ’08, 96–103, New York, NY, USA, 2008. ACM.
- [Cheriet 07] Mohammed Cheriet, Nawwaf Kharma, Cheng-lin Liu, and Ching Suen. “Character recognition systems: A guide for students and practitioners”. Wiley-Interscience, 2007.
- [Choi 12] Myung Jin Choi, Antonio Torralba, and Alan S. Willsky. “A tree-based context model for object recognition”. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(2):240–252, February 2012.
- [Conte 06] Donatello Conte. *Detection, Tracking, and Behaviour Analysis of Moving People in Intelligent Video Surveillance Systems : A Graph Based Approach*. PhD thesis, Université Lyon, Francia, 2006. Directores - Dr. J-M. Jolion, Dr. M. Vento.



- [Csurka 04] Gabriella Csurka, Christopher R. Dance, Lixin Fan, Jutta Willamowski, and Cédric Bray. “Visual categorization with bags of keypoints”. In *Workshop on Statistical Learning in Computer Vision, European Conference on Computer Vision*, 1–22, 2004.
- [DiCarlo 12] James J DiCarlo, Davide Zoccolan, and Nicole C Rust. “How does the brain solve visual object recognition?”. *Neuron*, 73:415–34, 2012 Feb 9 2012.
- [Dickinson 09] Sven J. Dickinson, Ales Leonardis, Bernt Schiele, and Michael J. Tarr. “Object categorization: Computer and human vision perspectives”. Cambridge University Press, New York, NY, USA, 1st edition, 2009.
- [Duchenne 11] Olivier Duchenne, Armand Joulin, and Jean Ponce. “A graph-matching kernel for object categorization”. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 1792–1799. IEEE, 2011.
- [Duval 10] Miguel A. Duval, Sandro Vega-Pons, and Eduardo Garea Llano. “Experimental comparison of orthogonal moments as feature extraction methods for character recognition”. In *Proceedings of the 15th Iberoamerican Congress on Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications (CIARP)*, 394–401, 2010.
- [Duygulu 02] P. Duygulu, Kobus Barnard, J. F. G. de Freitas, and David A. Forsyth. “Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary”. In *Proceedings of the 7th European Conference on Computer Vision-Part IV, ECCV ’02*, 97–112, London, UK, UK, 2002. Springer-Verlag.
- [Egenhofer 93] M. J. Egenhofer, J. Sharma, and D. M. Mark. “A critical comparison of the 4-intersection and 9-intersection models for spatial relations: Formal analysis”. In *Autocarto 11*, 1–11, 1993.
- [Escalante 07] H. J. Escalante, M. Montes, and E. Sucar. “Word co-occurrence and markov random fields for improving automatic image

annotation”. In *Proceedings of the 18th British Machine Vision Conference (BMVC-2007)*, September 2007.

- [Everingham 08] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. “The PASCAL Visual Object Classes Challenge 2008 (VOC2008) Results”. <http://www.pascal-network.org/challenges/VOC/voc2008/workshop/index.html>, 2008.
- [Everingham 11] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. “The PASCAL Visual Object Classes Challenge 2011 (VOC2011) Results”. <http://www.pascal-network.org/challenges/VOC/voc2011/workshop/index.html>, 2011.
- [Everingham 12] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. “The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results”. <http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html>, 2012.
- [Fei-Fei 04] Li Fei-Fei, Rob Fergus, and Pietro Perona. “Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories”. *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 12:178, 2004.
- [Felzenszwalb 13] Pedro Felzenszwalb, Ross Girshick, David McAllester, and Deva Ramanan. “Visual object detection with deformable part models”. *Commun. ACM*, 56(9):97–105, September 2013.
- [Feng 11] Jie Feng, Yichen Wei, Litian Tao, Chao Zhang, and Jian Sun. “Salient object detection by composition”. *Computer Vision, IEEE International Conference on*, 0:1028–1035, 2011.
- [Fergus 03] R. Fergus, P. Perona, and A. Zisserman. “Object class recognition by unsupervised scale-invariant learning”. In *Proceedings of the IEEE Conference on Computer Vision and*

*Pattern Recognition* (2), 264–271, Madison, Wisconsin, June 2003.

- [Fischer 04] Benedikt Fischer, Christian Thies, Mark O. Güld, and Thomas M. Lehmann. “Content-based image retrieval by matching hierarchical attributed region adjacency graphs”. In *Proc. SPIE–Medical Imaging: Image Processing* (5370), 598–606, 2004.
- [Fischler 81] Martin A. Fischler and Robert C. Bolles. “Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography”. *Commun. ACM*, 24(6):381–395, June 1981.
- [Fouquier 12] Geoffroy Fouquier, Jamal Atif, and Isabelle Bloch. “Sequential model-based segmentation and recognition of image structures driven by visual features and spatial relations”. *Comput. Vis. Image Underst.*, 116(1):146–165, January 2012.
- [Galleguillos 08] Carolina Galleguillos, Andrew Rabinovich, and Serge Belongie. “Object categorization using co-occurrence, location and appearance.”. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, 2008.
- [Galleguillos 10] Carolina Galleguillos and Serge Belongie. “Context based object categorization: A critical survey”. *Computer Vision and Image Understanding (CVIU)*, 114:712–722, 2010.
- [Gerstmayer 11] Michael Gerstmayer, Yll Haxhimusa, and Walter G. Kropatsch. “Hierarchical interactive image segmentation using irregular pyramids”. In *Graph-Based Representations in Pattern Recognition - 8th IAPR-TC-15 International Workshop, GbRPR 2011, Münster, Germany, May 18-20, 2011. Proceedings* (Xiaoyi Jiang, Miquel Ferrer, and Andrea Torsello, eds.) (6658) of *Lecture Notes in Computer Science*, 245–254. Springer, 2011.
- [Gibson 50] J. J. Gibson. “The Perception of the Visual World”. Houghton Mifflin, Boston, MA, 1950.

- [Glantz 04] Roland Glantz, Marcello Pelillo, and Walter G. Kropatsch. “Matching segmentation hierarchies”. *International Journal of Pattern Recognition and Artificial Intelligence*, 18(3):397–424, 2004.
- [González-Díaz 09] Rocío González-Díaz, Adrian Ion, Mabel Iglesias Ham, and Walter G. Kropatsch. “Irregular graph pyramids and representative cocycles of cohomology generators.”. In *GbRPR* (Andrea Torsello, Francisco Escolano, and Luc Brun, eds.) (5534) of *Lecture Notes in Computer Science*, 263–272. Springer, 2009.
- [González 03] Rafael C. González, Richard E. Woods, and Steven L. Eddins. “Digital image processing using matlab”. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 2003.
- [Grauman 05] Kristen Grauman and Trevor Darrell. “Pyramid match kernels: Discriminative classification with sets of image features”. Tech. Report MIT-CSAIL-TR-2005-017, Massachusetts Institute of Technology, Cambridge, 2005.
- [Grauman 07] Kristen Grauman and Trevor Darrell. “The pyramid match kernel: Efficient learning with sets of features”. *J. Mach. Learn. Res.*, 8:725–760, May 2007.
- [Grauman 11] Kristen Grauman and Bastian Leibe. “Visual object recognition”. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2011.
- [Guting 94] R. Hartmut Guting, P. Informatik Iv, and F. Hagen. “An introduction to spatial database systems”. *VLDB Journal*, 3:357–399, 1994.
- [Haxhimusa 04] Yll Haxhimusa and Walter G. Kropatsch. “Segmentation graph hierarchies”. In *Proceedings of Joint International Workshops on Structural, Syntactic, and Statistical Pattern Recognition S+SSPR 2004* (Ana Fred, Terry Caelli, Robert P.W. Duin, Aurelio Campilho, and Dick de Ridder, eds.) (LNCS 3138) of

*Lecture Notes in Computer Science*, 343–351, Lisbon, Portugal, 2004. Springer, Berlin Heidelberg, New York.

- [Haxhimusa 06] Yll Haxhimusa. “Structurally optimal dual graph pyramid and its application in image partitioning.”. Dissertations in Artificial Intelligence. IOS Press and Akadademische Verlagsgesellschaft AKA, Berlin, 2006.
- [He 09] Xiaofei He, Ming Ji, and Hujun Bao. “Graph embedding with constraints”. In *Proceedings of the 21st international joint conference on Artifical intelligence*, IJCAI’09, 1065–1070, San Francisco, CA, USA, 2009. Morgan Kaufmann Publishers Inc.
- [Hedau 12] Varsha Hedau, Derek Hoiem, and David A. Forsyth. “Recovering free space of indoor scenes from a single image.”. In *CVPR*, 2807–2814. IEEE, 2012.
- [Heikkilä 09] Marko Heikkilä, Matti Pietikäinen, and Cordelia Schmid. “Description of interest regions with local binary patterns”. *Pattern Recognition*, 42(3):425–436, 2009.
- [Hernández-Gracidas 07] Carlos Arturo Hernández-Gracidas and Luis Enrique Sucar. “Markov random fields and spatial information to improve automatic image annotation.”. In *PSIVT* (4872) of *Lecture Notes in Computer Science*, 879–892. Springer, 2007.
- [Hernández-Saura 13] Eric Hernández-Saura. “Recuperación de imágenes y videos por contenido utilizando matchpyr”. Tesis de Diploma Curso 20122013, Licenciatura en Ciencias de la Computación, Universidad de la Habana, Julio 2013. Asesor: Ing. Annette Morales González-Quevedo.
- [Hodé 07] Yann Hodé and Aline Deruyver. “Qualitative spatial relationships for image interpretation by using semantic graph.”. In *GbRPR* (4538) of *Lecture Notes in Computer Science*, 240–250. Springer, 2007.
- [Hu 13] Bo Hu and JianRui Zhang. “Spatial content-based scene matching using a relaxation method”. In *IEEE International*

*Conference on Anti-Counterfeiting, Security and Identification (ASID)*, 1 – 3. IEEE, 2013.

- [Humphreys 89] Glyn W. Humphreys and Vicki Bruce. “Visual Cognition: Computational, Experimental and Neuropsychological Perspectives”. Lawrence Erlbaum Associates, East Sussex, GBR, 1989.
- [Hurtut 08] T. Hurtut, Y. Gousseau, and F. Schmitt. “Adaptive image retrieval based on the spatial organization of colors”. *Comput. Vis. Image Underst.*, 112(2):101–113, 2008.
- [Iglesias-Ham 07] M. Iglesias-Ham, Y. Bazán-Pereira, and E. B. García-Reyes. “A multiple substructure matching algorithm for fingerprint verification”. In *Proceedings of the Iberomaerican Congress on Pattern Recognition, CIARP’07* (4756) of *LNCS*, 172–181. Springer-Verlag, 2007.
- [Ion 14] Adrian Ion, João Carreira, and Cristian Sminchisescu. “Probabilistic joint image segmentation and labeling by figure-ground composition”. *International Journal of Computer Vision*, 107(1):40–57, 2014.
- [Jia 11] Yi Jia, Jintao Zhang, and Jun Huan. “An efficient graph-mining method for complicated and noisy data with real-world applications.”. *Knowl. Inf. Syst.*, 28(2):423–447, 2011.
- [Johnson 80] KO Johnson. “Sensory discrimination: decision process.”. *J Neurophysiol*, 43(6):1771–92, 1980.
- [Jung 08] Ho Yub Jung, Kyoung Mu Lee, and Sang Uk Lee. “Toward global minimum through combined local minima.”. In *ECCV (4)* (David A. Forsyth, Philip H. S. Torr, and Andrew Zisserman, eds.) (5305) of *Lecture Notes in Computer Science*, 298–311. Springer, 2008.
- [Keuper 11] M. Keuper, T. Schmidt, M. Rodriguez-Franco, W. Schamel, T. Brox, H. Burkhardt, and O. Ronneberger. “Hierarchical markov random fields for mast cell segmentation in electron

- microscopic recordings”. In *Proceedings of the 8th IEEE International Symposium on Biomedical Imaging, ISBI 2011*, 973–978, 2011.
- [Kim 06] D. H. Kim, I. D. Yun, and S. U. Lee. “New mrf parameter estimation technique for texture image segmentation using hierarchical gmrf model based on random spatial interaction and mean field theory”. In *Proceedings of ICPR 2006 - Volume 02*, ICPR ’06, 365–368, Washington, DC, USA, 2006. IEEE Computer Society.
- [Kropatsch 04] Walter G. Kropatsch, Yll Haxhimusa, and Pascal Lienhardt. “Cognitive Vision Systems: Sampling the Spectrum of Approaches”, ch. 13. Hierarchies relating Topology and Geometry 13. Hierarchies relating Topology and Geometry. Lecture Notes in Computer Science. Springer, Berlin Heidelberg, Dagstuhl, September 2004.
- [Kropatsch 05] Walter G. Kropatsch, Yll Haxhimusa, Zygmunt Pizlo, and Georg Langs. “Vision pyramids that do not grow too high”. *Pattern Recognition Letters*, 26(3):319–337, 2005.
- [Lazebnik 06] S. Lazebnik, C. Schmid, and J. Ponce. “Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories”. In *CVPR ’06: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2169–2178. IEEE Computer Society, 2006.
- [Leibe 03] B. Leibe and B. Schiele. “Analyzing appearance and contour based methods for object categorization”. In *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR’03)*, 409–415, 2003.
- [Li 09] Stan Z. Li. “Markov random field modeling in image analysis”. Springer Publishing Company, Incorporated, 3rd edition, 2009.
- [Liao 12] Jiawen Liao, Jianzhong Cao, and Linao Tang. “Research on video stabilization algorithm based on sift and improved ransac”. In *Proceedings of the 2012 Second International*

*Conference on Electric Information and Control Engineering - Volume 03*, ICEICE '12, 755–758, Washington, DC, USA, 2012. IEEE Computer Society.

- [Lin 03] P. L. Lin and W. H. Tan. “An efficient method for the retrieval of objects by topological relations in spatial database systems”. *Inf. Process. Manage.*, 39(4):543–559, 2003.
- [Liu 12] Ying-Ho Liu, Anthony J. T. Lee, and Fu Chang. “Object recognition using discriminative parts”. *Comput. Vis. Image Underst.*, 116(7):854–867, July 2012.
- [Llorente 10] A. Llorente, R. Manmatha, and S. Rüger. “Image retrieval using markov random fields and global image features”. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, CIVR '10, 243–250. ACM, 2010.
- [Lowe 85] David G. Lowe. “Perceptual organization and visual recognition”. Kluwer Academic Publishers, Norwell, MA, USA, 1985.
- [Lowe 04] David G. Lowe. “Distinctive image features from scale-invariant keypoints”. *Int. J. Comput. Vision*, 60(2):91–110, November 2004.
- [Maire 13] Michael Maire, Stella X. Yu, and Pietro Perona. “Hierarchical scene annotation”. In *British Machine Vision Conference (BMVC)*, 2013.
- [Malisiewicz 07] Tomasz Malisiewicz and Alexei A. Efros. “Improving spatial support for objects via multiple segmentations.”. In *BMVC*. British Machine Vision Association, 2007.
- [Marée 05] Raphaël Marée, Pierre Geurts, Justus Piater, and Louis Wehenkel. “Decision trees and random subwindows for object recognition”. In *ICML workshop on Machine Learning Techniques for Processing Multimedia Content (MLMM2005)*, 2005.



- [Markman 00] A. B. Markman and D. Gentner. “Structure mapping in the comparison process”. *American Journal of Psychology*, 113(4):501–538, 2000.
- [Marr 80] David Marr and E. Hildreth. “Theory of edge detection”. *Proceedings of the Royal Society of London Series B*, 207:187–217, 1980.
- [Marr 82] David Marr. “Vision: A computational investigation into the human representation and processing of visual information”. Henry Holt and Co., Inc., New York, NY, USA, 1982.
- [Morales-González 09] Annette Morales-González and Edel García-Reyes. “Content-based image retrieval using topological descriptors”. Tech. Report RT\_039, Serie Azul, Centro de Aplicaciones de Tecnologías de Avanzada, La Habana, Cuba, 2009.
- [Morales-González 10a] Annette Morales-González and Edel García-Reyes. “Assessing the role of spatial relations for the object recognition task”. In *Proceedings of the 15th Iberoamerican Congress Conference on Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, CIARP’10, 549–556, Berlin, Heidelberg, 2010. Springer-Verlag.
- [Morales-González 10b] Annette Morales-González, Edel B. García Reyes, and Luis E. Sucar. “Segmentation based on level combination of irregular pyramids.”. In *Proceedings of the Automatic Image Annotation and Retrieval Workshop 2010* (719), 1–10. CEUR Workshop Proceedings, 2010.
- [Morales-González 12] Annette Morales-González, Edel B. García Reyes, and Luis Enrique Sucar. “Hierarchical markov random fields with irregular pyramids for improving image annotation.”. In *13th Ibero-American Conference on Artificial Intelligence (IBERAMIA)* (7637) of *Lecture Notes in Computer Science*, 521–530. Springer, 2012.
- [Morales-González 13a] Annette Morales-González and Edel B. García-Reyes. “Simple object recognition based on spatial relations and visual features

represented using irregular pyramids”. *Multimedia Tools Appl.*, 63(3):875–897, April 2013.

- [Morales-González 13b] Annette Morales-González, Edel B. García Reyes, and Luis Enrique Sucar. “Improving image segmentation for boosting image annotation with irregular pyramids.”. In *Iberoamerican Congress on Pattern Recognition (CIARP)* (8258) of *Lecture Notes in Computer Science*, 399–406. Springer, 2013.
- [Morales-González 14] Annette Morales-González, Niusvel Acosta-Mendoza, Andrés Gago-Alonso, Edel B. García-Reyes, and José E. Medina-Pagola. “A new proposal for graph-based image classification using frequent approximate subgraphs”. *Pattern Recogn.*, 47(1):169–177, January 2014.
- [Morioka 08] Nobuyuki Morioka. “Learning object representations using sequential patterns”. In *Proceedings of the 21st Australasian Joint Conference on Artificial Intelligence: Advances in Artificial Intelligence*, AI ’08, 551–561, 2008.
- [Nene 96] S. A. Nene, S. K. Nayar, and H. Murase. “Columbia Object Image Library (COIL-100)”. Tech. Report CUCS-006-96, Department of Computer Science, Columbia University, Feb 1996.
- [Noma 12] Alexandre Noma, Ana B. V. Graciano, Roberto M. Cesar Jr, Luis A. Consularo, and Isabelle Bloch. “Interactive image segmentation by matching attributed relational graphs”. *Pattern Recogn.*, 45(3):1159–1179, March 2012.
- [Nomiya 09] H. Nomiya and K. Uehara. “Data mining and knowledge discovery in real life applications”, ch. 9 9, 157–166. IN-TECH, 2009.
- [Obdržálek 02] Stepán Obdržálek and Jiri Matas. “Object recognition using local affine frames on distinguished regions”. In *Proceedings of the British Machine Vision Conference 2002*. British Machine Vision Association, 2002.

- [Ojala 96] T. Ojala, M. Pietikainen, and D. Harwood. “A comparative study of texture measures with classification based on featured distribution”. *Pattern Recognition*, 29(1):51–59, 1996.
- [Olson 01] C. R. Olson. “Object-based vision and attention in primates”. *Current Opinion in Neurobiology*, 11:171–179, 2001.
- [Pantofaru 08] Caroline Pantofaru, Cordelia Schmid, and Martial Hebert. “Object recognition by integrating multiple image segmentations”. In *Proceedings of the 10th European Conference on Computer Vision: Part III, ECCV ’08*, 481–494, Berlin, Heidelberg, 2008. Springer-Verlag.
- [Perronnin 12] Florent Perronnin, Zeynep Akata, Zaïd Harchaoui, and Cordelia Schmid. “Towards good practice in large-scale learning for image classification.”. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3482–3489. IEEE, 2012.
- [Peterson 94] M. A. Peterson. “Object recognition processes can and do operate before figure ground organization.”. *Current Directions in Psychological Science*, 3:105–111, 1994.
- [Pham 10] Trong-Ton Pham, Philippe Mulhem, Loic Maisonnasse, Eric Gaussier, and Joo-Hwee Lim. “Visual graph modeling for scene recognition and mobile robot localization”. *Multimedia Tools and Applications*, 1–23–23, September 2010.
- [Pizlo 01] Zygmunt Pizlo. “Perception viewed as an inverse problem”. *Vision Research*, 41(24):3145 – 3161, 2001.
- [Punitha 06] P. Punitha and D. S. Guru. “An effective and efficient exact match retrieval scheme for symbolic image database systems based on spatial reasoning: A logarithmic search time approach”. *IEEE Trans. on Knowl. and Data Eng.*, 18(10):1368–1381, 2006.
- [Rao 10] Ch.Srinivasa Rao, S.Srinivas Kumar<sup>2</sup>, and B.Chandra Mohan<sup>3</sup>. “Content based image retrieval using exact legendre moments and support vector machine”. *International Journal of Multimedia and Its Application*, 2(2):69–79, 2010.

- [Ren 14] Yi Ren, Aurélie Bugeau, and Jenny Benois-Pineau. “Bag-of-Bags of Words - Irregular Graph Pyramids vs Spatial Pyramid Matching for Image Retrieval”. In *4th International Conference on Image Processing Theory, Tools and Applications*, In Press, October 2014.
- [Riesenhuber 00] Maximilian Riesenhuber and Tomaso Poggio. “Computational models of object recognition in cortex: A review”. tech. report, and 190, Artificial Intelligence Laboratory and Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, 2000.
- [Roberts 63] Lawrence G. Roberts. “Machine perception of three-dimensional solids”. Outstanding Dissertations in the Computer Sciences. Garland Publishing, New York, 1963.
- [Roth 06] Volker Roth and Björn Ommer. “Exploiting low-level image segmentation for object recognition”. In *Pattern Recognition, 28th DAGM Symposium, Berlin, Germany, September 12-14, 2006, Proceedings* (4174) of *Lecture Notes in Computer Science*, 11–20. Springer, 2006.
- [Russakovsky 14] O. Russakovsky, J. Deng, J. Krause, A. Berg, and F.F. Li. “Results of ILSVRC2013”. <http://www.image-net.org/challenges/LSVRC/2013/results.php>, 2014.
- [Russell 14] Chris Russell, Lúbor Ladicky, Pushmeet Kohli, and Philip H. S. Torr. “Associative hierarchical random fields”. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(6):1–1, 2014.
- [Saux 05] Bertrand Le Saux and Horst Bunke. “Feature selection for graph-based image classifiers”. In *IbPRIA (2)* (3523) of *Lecture Notes in Computer Science*, 147–154. Springer, 2005.
- [Shahiduzzaman 10] Mohammad Shahiduzzaman, Dengsheng Zhang, and Guojun Lu. “Improved spatial pyramid matching for image classification.”. In *10th Asian conference on computer vision*

- (*ACCV*) 2010. Part IV (6495) of *Lecture Notes in Computer Science*, 449–459. Springer, 2010.
- [Shi 97] J. Shi and J. Malik. “Normalized cuts and image segmentation”. In *Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR ’97)*, CVPR ’97, 731–, Washington, DC, USA, 1997. IEEE Computer Society.
- [Sivic 03] J. Sivic and A. Zisserman. “Video Google: A text retrieval approach to object matching in videos”. In *Proceedings of the International Conference on Computer Vision (2)*, 1470–1477, October 2003.
- [Sjöö 12] Kristoffer Sjöö, Alper Aydemir, and Patric Jensfelt. “Topological spatial relations for active visual search”. *Robot. Auton. Syst.*, 60(9):1093–1107, September 2012.
- [Skurikhin 09] A.N. Skurikhin. “Hierarchical image feature extraction by an irregular piramid of polygonal partitions”. In *Proceedings of the Annual Conference - American Society for Photogrammetry and Remote Sensing (2)*, 775–786. American Society for Photogrammetry and Remote Sensing , Bethesda, Md., 2009.
- [Sokal 58] R. R. Sokal and C.D. Michener. “A statistical method for evaluating systematic relationships”. *University of Kansas Science Bulletin*, 38:1409–1438, 1958.
- [Song 10] Yi-Zhe Song, Pablo Arbelaez, Peter Hall, Chuan Li, and Anupriya Balikai. “Finding semantic structures in image hierarchies using laplacian graph energy”. In *Proceedings of the 11th European conference on Computer vision: Part IV*, ECCV’10, 694–707, Berlin, Heidelberg, 2010. Springer-Verlag.
- [Spitzer 71] F. Spitzer. “Random fields and interacting particle systems: Notes on lectures given at the 1971 maa summer seminar, williamstown - mass”. 1971.
- [Su 11] Bor-Yiing Su, Tasneem G. Brutch, and Kurt Keutzer. “A parallel region based object recognition system”. In *Proceedings*

of the 2011 IEEE Workshop on Applications of Computer Vision (WACV), WACV '11, 81–88, Washington, DC, USA, 2011. IEEE Computer Society.

- [Takala 05] Valtteri Takala, Timo Ahonen, and Matti Pietikäinen. “Block-based methods for image retrieval using local binary patterns”. In *SCIA (3540) of Lecture Notes in Computer Science*, 882–891, 2005.
- [Tegen 14] A. Tegen, R. Weegar, L. Hammarlund, M. Oskarsson, F. Jiang, D. Medved, P. Nugues, and K. Åström. “Image segmentation and labeling using free-form semantic annotation”. In *International Conference on Pattern Recognition*, 2281–2286, 2014.
- [Todorovic 08] Sinisa Todorovic and Narendra Ahuja. “Region-based hierarchical image matching”. *Int. J. Comput. Vision*, 78(1):47–66, June 2008.
- [Torrent 13] Albert Torrent, Xavier Lladó, Jordi Freixenet, and Antonio Torralba. “A boosting approach for the simultaneous detection and segmentation of generic objects”. *Pattern Recogn. Lett.*, 34(13):1490–1498, October 2013.
- [Torres 10] Fuensanta Torres, Rebeca Marfil, Yll Haxhimusa, and Antonio Bandera. “Combining regular decimation and dual graph contraction for hierarchical image segmentation”. In *3rd International Workshop on Computational Topology in Image Context*, 97–104. University of Sevilla, Spain, November 2010. ISSN: 1885-4508.
- [Tousch 12] Anne-Marie Tousch, StéPhane Herbin, and Jean-Yves Audibert. “Semantic hierarchies for image annotation: A survey”. *Pattern Recogn.*, 45(1):333–345, January 2012.
- [Tsapatsoulis 07] N. Tsapatsoulis and S. Petridis. “Classifying images from athletics based on spatial relations”. In *International Workshop on Semantic Media Adaptation and Personalization*, 92–97. IEEE Computer Society, 2007.

- [Tsotsos 88] John K. Tsotsos. “How does human vision beat the computational complexity of visual perception?”. In *Computational Processes in Human Vision* (Z. W. Pylyshyn, ed.). Ablex, Norwood, NJ, 1988.
- [Ullman 07] Shimon Ullman. “Object recognition and segmentation by a fragment-based hierarchy.”. *Trends in cognitive sciences*, 11(2):58–64, February 2007.
- [van de Sande 11] Koen E. A. van de Sande, Jasper R. R. Uijlings, Theo Gevers, and Arnold W. M. Smeulders. “Segmentation as selective search for object recognition”. In *Proceedings of ICCV ’11*, 1879–1886. IEEE Computer Society, 2011.
- [Vecera 98] Shaun P. Vecera and Randall C. O’Reilly. “Figure-ground organization and object recognition processes: An interactive account.”. *Journal of Experimental Psychology: Human Perception and Performance*, 24:441–462, 1998.
- [Vieux 10] Remi Vieux, Jenny Benois-Pineau, Jean-Philippe Domenger, and Achille Braquelaire. “Segmentation-based multi-class semantic object detection”. *Multimedia Tools and Applications*, 1–22–22, October 2010.
- [Vieux 12] R. Vieux, J. Benois-Pineau, J.P. Domenger, and A. Braquelaire. “Segmentation-based multi-class semantic object detection”. *Multimedia Tools Appl.*, 60(2):305–326, September 2012.
- [Viola 04] Paul Viola and Michael J. Jones. “Robust real-time face detection”. *Int. J. Comput. Vision*, 57(2):137–154, May 2004.
- [Wang 06] Yong Wang and Shaogang Gong. “Tensor discriminant analysis for view-based object recognition”. In *Proceedings of the 18th International Conference on Pattern Recognition - Volume 03*, ICPR ’06, 33–36, 2006.
- [Wang 13] Jim Jing-Yan Wang, Halima Bensmail, and Xin Gao. “Joint learning and weighting of visual vocabulary for bag-of-feature based tissue classification.”. *Pattern Recognition*, 46(12):3249–3255, 2013.

- [Weiss 12] I. Weiss. “Robust model-based object recognition using a dual-hierarchy graph”. In *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series* (8391) of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, May 2012.
- [Wertheimer 23] Max Wertheimer. “69 - untersuchungen zur lehre von der gestalt.”. *Psychologische Forschung: Zeitschrift für Psychologie und ihre Grenzwissenschaften*, 4:301–350, 1923.
- [Xiang 09] Y. Xiang, X. Zhou, T. Chua, and C. Ngo. “A revisit of generative model for automatic image annotation using markov random fields”. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1153–1160, 2009.
- [Yang 11] Michael Ying Yang and Wolfgang Förstner. “A hierarchical conditional random field model for labeling and classifying images of man-made scenes”. In *ICCV Workshop on Computer Vision for Remote Sensing of the Environment*, 196 – 203, 2011.
- [Yao 10] Bangpeng Yao and Fei-Fei Li. “Grouplet: A structured image representation for recognizing human and object interactions.”. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 9–16. IEEE, 2010.
- [Yoon 11] Kuk-Jin Yoon and Min-Gil Shin. “Reducing ambiguity in object recognition using relational information”. In *Proceedings of the 10th Asian conference on Computer vision - Volume Part IV*, ACCV’10, 293–306, Berlin, Heidelberg, 2011. Springer-Verlag.
- [Zagoris 11] Konstantinos Zagoris, Savvas A. Chatzichristofis, and Avi Arampatzis. “Bag-of-visual-words vs global image descriptors on two-stage multimodal retrieval”. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR ’11, 1251–1252, New York, NY, USA, 2011. ACM.
- [Zankl 12] Georg Zankl, Yll Haxhimusa, and Adrian Ion. “Interactive labeling of image segmentation hierarchies.”. In *DAGM/OAGM*



*Symposium* (Axel Pinz, Thomas Pock, Horst Bischof, and Franz Leberl, eds.) (7476) of *Lecture Notes in Computer Science*, 11–20. Springer, 2012.

- [Zhang 03] Bin Zhang and Sargur N. Srihari. “Binary vector dissimilarity measures for handwriting identification”. In *Proceedings of Document Recognition and Retrieval Conference (DRR)* (5010) of *SPIE Proceedings*, 28–38, 2003.
- [Zhang 07] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. “Local features and kernels for classification of texture and object categories: A comprehensive study”. *Int. J. Comput. Vision*, 73:213–238, June 2007.
- [Zhang 12] Junge Zhang, Xin Zhao, Yongzhen Huang, Kaiqi Huang, and Tieniu Tan. “Semantic windows mining in sliding window based object detection.”. In *International Conference on Pattern Recognition (ICPR)*, 3264–3267. IEEE, 2012.
- [Zhang 13] Shu Zhang and Mei Xie. “Beyond sliding windows: Object detection based on hierarchical segmentation model”. In *International Conference on Communications, Circuits and Systems (ICCCAS)*, 263 – 266. IEEE, 2013.

## Producción científica de la autora sobre el tema de la tesis

### Artículos en Revistas de Impacto (referenciadas en Web of Science)

1. **A. Morales-González**, N. Acosta-Mendoza, A. Gago-Alonso, E.B. García-Reyes, and J.E. Medina-Pagola. “A New Proposal for Graph-Based Image Classification using Frequent Approximate Subgraphs”. Pattern Recognition, 2013. DOI: <http://dx.doi.org/10.1016/j.patcog.2013.07.004>. (**Factor de impacto: 2.632**)
2. **A. Morales-González**, E. García-Reyes, “Simple object recognition based on spatial -relations and visual features represented using irregular pyramids”. Multimedia Tools Appl. Vol. 63, No. 3, pp. Vol. 63, No. 3, pp. 875-897, 2013. (**Factor de impacto: 1.014**)

### Artículos en Memorias de Eventos (referenciados en SCOPUS)

3. **A. Morales-González**, E. García-Reyes, and L. E. Sucar, “Improving Image Segmentation for Boosting Image Annotation with Irregular Pyramids”, In Proceedings of CIARP 2013, Springer LNCS 8258, pp. 399-406, 2013.
4. **A. Morales-González**, E. García-Reyes, and L. E. Sucar, “Hierarchical Markov Random Fields with Irregular Pyramids for Improving Image Annotation”, In Proceedings of IBERAMIA 2012, Springer LNAI 7637, pp. 521–530, 2012.
5. N. Acosta-Mendoza, **A. Morales-González**, A. Gago-Alonso, E. B. García-Reyes, and J. E. Medina-Pagola, “Image Classification Using Frequent Approximate Subgraphs”, In Proceedings of CIARP 2012, Springer LNCS 7441, pp. 292–299, 2012.
6. **A. Morales-González**, E. García-Reyes, “Assessing the Role of Spatial Relations for the Object Recognition Task”. 15th Iberoamerican Congress on Pattern Recognition (CIARP 2010), Springer LNCS 6419, pp. 549-556, November 08-11, São Paulo, Brazil, 2010.

7. **A. Morales-González**, E. García-Reyes, L. E. Sucar, “Segmentation based on level combination of irregular pyramids”. Proceedings of the Automatic Image Annotation and Retrieval Workshop (AIAR’2010), Puebla, Mexico, volume 1, issue 1, pp. 1-10, 2010.

#### **Otras publicaciones**

8. **A. Morales-González**, E. García-Reyes, “Content-Based Image Retrieval using Topological Descriptors”. Technical Report, Serie Azul, CENATAV, RT\_0039, 2009.

#### **Asesoría de trabajo de diploma**

9. Hernández, Eric J.: “Recuperación de imágenes y videos por contenido utilizando MatchPyr”. Tesis de Diploma Curso 2012-2013, Licenciatura en Ciencias de la Computación, Asesor: Ing. **Annette Morales González-Quevedo**. (2013.).

## Glosario de acrónimos

CK	Núcleos de Contracción, del inglés <i>Contraction Kernels</i>
CM	Mapas Combinatorios, del inglés <i>Combinatorial Maps</i>
CRF	Campos Aleatorios Condicionales, del inglés <i>Conditional Random Fields</i>
FAS	Subgrafos Aproximados Frecuentes, del inglés <i>Frequent Approximate Subgraphs</i>
ICM	Modas Condicionales Iterativas, del inglés <i>Iterative Conditional Modes</i>
KNN	Clasificador basado en los K vecinos más cercanos, del inglés <i>K-Nearest Neighbors</i>
LBP	Patrones Binarios Locales, del inglés <i>Local Binary Patterns</i>
MRF	Campos Aleatorios de Markov, del inglés <i>Markov Random Fields</i>
MST	Árbol de Expansión Mínima, del inglés <i>Minimum Spanning Tree</i>
RAG	Grafo de Adyacencia de Regiones, del inglés <i>Region Adjacency Graph</i>
RCF	Rasgos Contextuales basados en Regiones, del inglés <i>Region-based Context Features</i>
RF	Clasificador basado en árboles de decisión, del inglés <i>Random Forest</i>
SIFT	Descriptor basado en puntos de interés, del inglés <i>Scale Invariant Feature Transform</i>
SURF	Descriptor basado en puntos de interés, del inglés <i>Speeded Up Robust Features</i>
SVM	Clasificador basado vectores de soporte, del inglés <i>Support Vector Machines</i>



## Glosario de términos

<b>algoritmo húngaro</b>	Es un algoritmo de optimización que resuelve problemas de asignación (encontrar un emparejamiento de peso máximo en un grafo bipartido ponderado).
<b>bounding box</b>	Región rectangular que delimita o enmarca un objeto o área de interés en una imagen.
<b>bolsa de palabra visuales</b>	Enfoque de reconocimiento de objetos que se basa en la construcción de vocabularios visuales a partir del agrupamiento de rasgos de bajo nivel, para luego utilizar la ocurrencia de cada palabra visual del vocabulario como rasgos en la clasificación.
<b>clasificador boosting</b>	Clasificador basado en la combinación de clasificadores débiles para obtener un clasificador fuerte.
<b>contexto espacial</b>	Se refiere al área que rodea a un objeto en el espacio de la imagen.
<b>cuantización</b>	Consiste en reducir el número de valores a utilizar en un rango de valores, agrupándolos por bloques de valores (ej. la cuantización del rango de valores de una imagen en escala de grises (256 valores por píxel) consiste en reducir el número de valores utilizados para representar dicha imagen, i.e. se puede representar la imagen con 128 valores por píxel, donde cada valor representa 2 valores del rango original)

<b>etiquetado de imágenes</b>	Proceso mediante el cual se le asignan etiquetas o clases a distintas regiones de la imagen, o a la imagen completa.
<b>graph embedding</b>	Esquema que se utiliza para proyectar la estructura de un grafo en una superficie (ej. en un espacio vectorial).
<b>ground truth</b>	Conjunto de medidas que se sabe que son más acertadas que las medidas del sistema que se prueba. En el campo específico de las imágenes, se entienden como <u>ground truth</u> de etiquetado y/o segmentaciones, a etiquetados y/o segmentaciones ideales (usualmente generadas manualmente por humanos), que se utilizan para verificar la eficacia de los algoritmos de etiquetado y/o segmentación.
<b>imagen retinal</b>	Información visual que recibe como entrada un sistema visual (biológico o sintético).
<b>jerarquías de abstracción</b>	Se refiere a la estructuración de los tipos o clases de objetos según su generalidad (niveles superiores de la jerarquía) o especificidad (niveles inferiores de la jerarquía), donde las clases más específicas heredan características o atributos de las clases más generales. Ej. en una jerarquía de abstracción los elementos “perro”, “gato” y “delfín” serían hijos del elemento “mamífero”, que se encuentra en un nivel superior.
<b>jerarquías parte/todo</b>	Se refiere a la descomposición de un objeto en partes y a sus relaciones de agregación. Ej. en una jerarquía parte/todo, los elementos “ojos”, “boca” y “nariz” serían hijos del elemento “rostro” que se encuentra en un nivel superior de la jerarquía.
<b>k-means</b>	Es un algoritmo de agrupamiento que particiona $n$ observaciones en $k$ grupos, y cada observación pertenece al grupo de la media más cercana.
<b>loopy belief propagation</b>	Es un algoritmo de paso de mensajes para realizar inferencia en modelos gráficos, tales como redes bayesianas y campos aleatorios de Markov.

<b>objetos deformables</b>	Son objetos que pueden ser definidos por el conjunto de sus partes, las cuales pueden aparecer en distintas configuraciones espaciales.
<b>recocido simulado</b>	Es un algoritmo de búsqueda meta-heurística para problemas de optimización global.
<b>reconocimiento de objetos específicos</b>	Consiste en identificar instancias de clases de objetos. Ej. el capitolio, mi perro Robin, la botella de mi refresco preferido.
<b>reconocimiento de clases de objetos</b>	Consiste en identificar categorías genéricas de objetos y no instancias de los mismos. Ej. edificios, perros, botellas
<b>segmentación jerárquica o multiresolución</b>	Proceso de segmentación de imágenes que genera distintas particiones de la misma a distintas resoluciones, donde los niveles más bajos quedan muy sobre-segmentados (muchas regiones pequeñas) mientras que los niveles superiores están sub-segmentados (pocas regiones grandes).
<b>sistema visual ventral</b>	Dentro del sistema visual, la vía ventral comprende regiones que procesan el color y la forma. De forma general, proporciona información de “qué” es lo que se ve.





## Anexos



# Anexo 1: Ejemplo de las ventajas que proporciona el descriptor espacial propuesto

Características del descriptor espacial propuesto (Ver Figura A.1):

1. Un descriptor binario que codifica configuraciones espaciales complejas a partir de relaciones simples.
2. Se puedan calcular valores de similitud entre dichos descriptores, es decir, se obtiene una medida de cuán similares o disimilares son las distintas configuraciones espaciales representadas

H	L	R	V	T	B	A	C	I
H – Alineados horizontalmente			V – Alineados verticalmente			A – Adyacente		
L – A la izquierda de			T – Arriba de			C – Contiene		
R – A la derecha de			B – Debajo de			I – Es contenido por		

Figura A.1: Descriptor espacial que combina relaciones topológicas y de orientación.

Con este descriptor es posible diferenciar configuraciones espaciales complejas, como se puede observar en la Figura A.2. En esta imagen se puede ver que desde el punto de vista topológico, estas dos relaciones serían equivalentes, representando *contiene a*, pero al combinarlas en el descriptor espacial con relaciones de dirección es posible diferenciarlas.

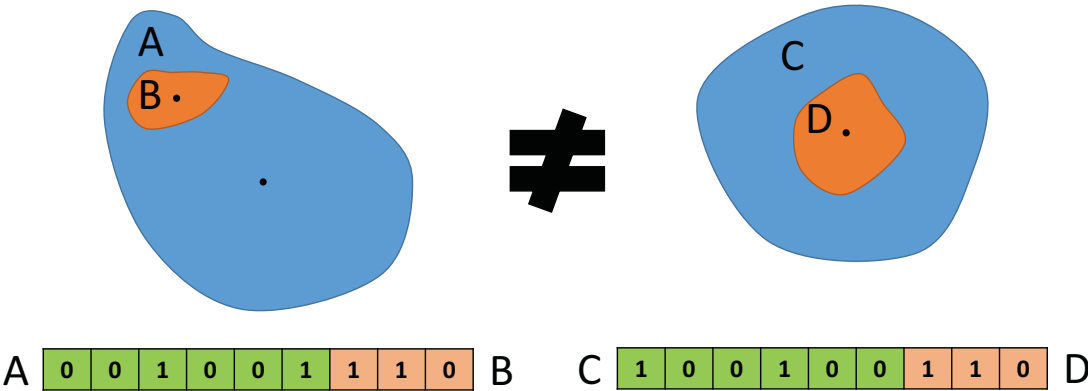


Figura A.2: Diferencia en la representación de dos configuraciones espaciales.

Se pueden calcular valores cuantitativos de la similitud entre estos descriptores, mediante las coincidencias de relaciones básicas que cada vector representa. Esto se puede observar en la Figura A.3. A la izquierda se muestran las configuraciones espaciales entre 3 pares de regiones: A-B, C-D y E-F. A la derecha se pueden ver los descriptores espaciales de cada configuración y el valor de similitud que es posible hallar entre ellas. En este ejemplo se puede ver que la configuración espacial entre las regiones C-D es más parecida a la configuración de A-B que a la de E-F

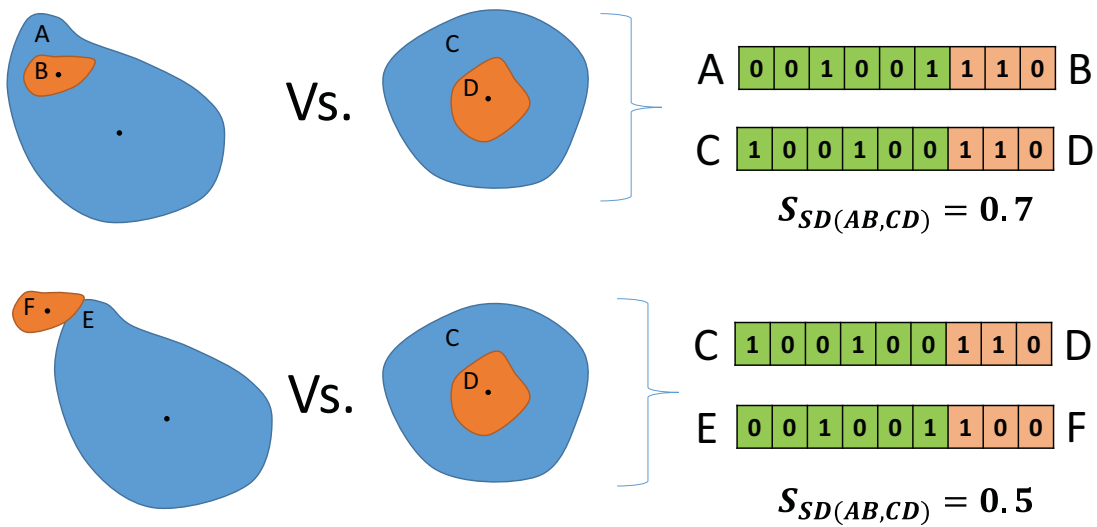


Figura A.3: Valores de similitud entre configuraciones espaciales.

## Anexo 2: Pruebas del algoritmo MATCH-PYR para el reconocimiento de objetos en escenarios reales de videovigilancia

En el trabajo de diploma [Hernández-Saura 13], se presentó como objetivo general optimizar la implementación del método MATCH-PYR de forma que fuera más eficiente y pudiera ser adaptado a la recuperación de objetos en video. La solución propuesta disminuyó el tiempo de corrida del algoritmo en al menos 4 veces y se creó una estrategia para aplicarlo eficientemente en videos. Su adaptación a video fue probada en archivos de videovigilancia tomados por cámaras ubicadas en calles de Cuba.

La optimización de la implementación del método, que ya se encontraba desarrollado en C++, estuvo dirigida a la utilización de bibliotecas más eficientes, como OpenCV (para el manejo de las imágenes y su información visual) y Boost para el manejo de los grafos. También se optimizaron los procesos de extracción de rasgos visuales y bordes de las regiones, de manera que se realizaran únicamente en los niveles de la pirámide que fueran útiles para el proceso de reconocimiento, y no en toda la pirámide.

Se hicieron experimentos de recuperación de videos por contenido, donde se considera que un cuadro de video es relevante si contiene el objeto que se desea recuperar. Se propuso dividir cada video en segmentos, luego a cada segmento se le extraen cuadros representativos. De esta forma se construye un índice que asocia los cuadros representativos al segmento que este representa. Para reducir el espacio de los cuadros representativos se propuso agrupar los cuadros representativos que sean muy parecidos de modo que un cuadro puede pasar a representar varios segmentos. Esto último presenta gran utilidad en videos donde hay segmentos que se repiten de manera periódica en el tiempo, por ejemplos los videos de vigilancia.

Fueron utilizados videos capturados por una cámara de videovigilancia en los cuales aparecen objetos de distintas categorías, entre las que se encuentran automóviles, peatones y motocicletas. El video es tomado en una calle en el horario del día, y el tamaño de los cuadros del video es de 300x300 píxeles. Para tener objetos que sirvan de ejemplos a recuperar, fueron extraídos del video distintos cuadros que contienen objetos de distintas clases. El objetivo es buscar estos objetos dentro del video para evaluar la capacidad del método para recuperarlos aún cuando estos se encuentren en distintos lugares, bajo diferentes poses e incluso cuando alguna parte de estos objetos esté solapada por otro

objeto. Los resultados se resumen en la Figura A.1. En el eje horizontal se muestran los  $k$  primeros cuadros recuperados y en el eje vertical se muestra la precisión de la recuperación.

Un ejemplo de confusión en la clasificación se puede ver en la Figura A.2. El error de clasificación ocurrió entre un peatón y una motocicleta en posición frontal, la cual tiene características muy parecidas a un peatón.

Un ejemplo de que el algoritmo es robusto ante oclusiones se puede ver en la Figura A.3, donde se encuentra a la misma mujer, en distinta pose y ocluida por un automóvil.

En la Figura A.4 se muestra una captura de pantalla de la herramienta desarrollada para realizar experimentos en videos.

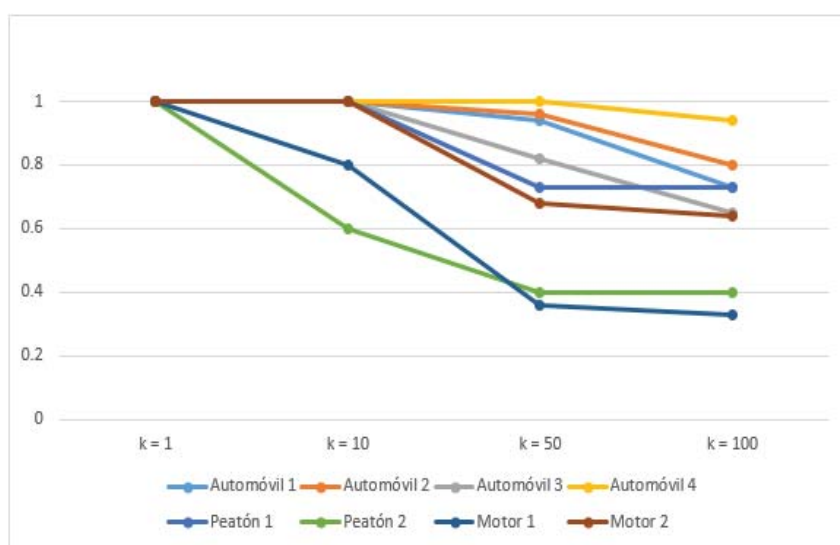


Figura A.1: Resultados de recuperación de objetos en video.

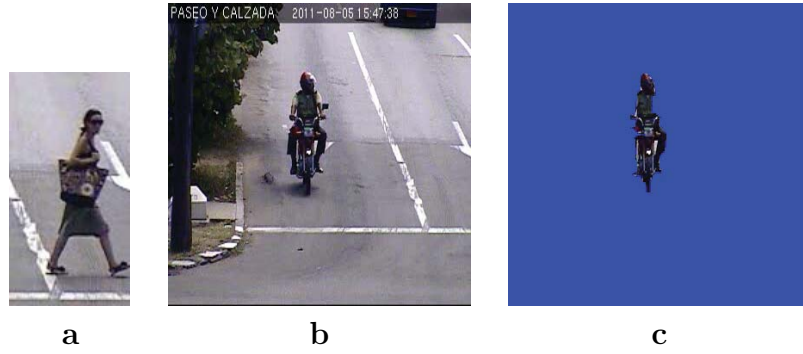


Figura A.2: Ejemplo de confusión de clases en el reconocimiento. a) Objeto de interés que se busca, b) cuadro de video que el algoritmo encontró como relevante para la presencia del objeto de interés, c) máscara que muestra la correspondencia hallada entre el objeto de interés en a) y el cuadro de video en b).

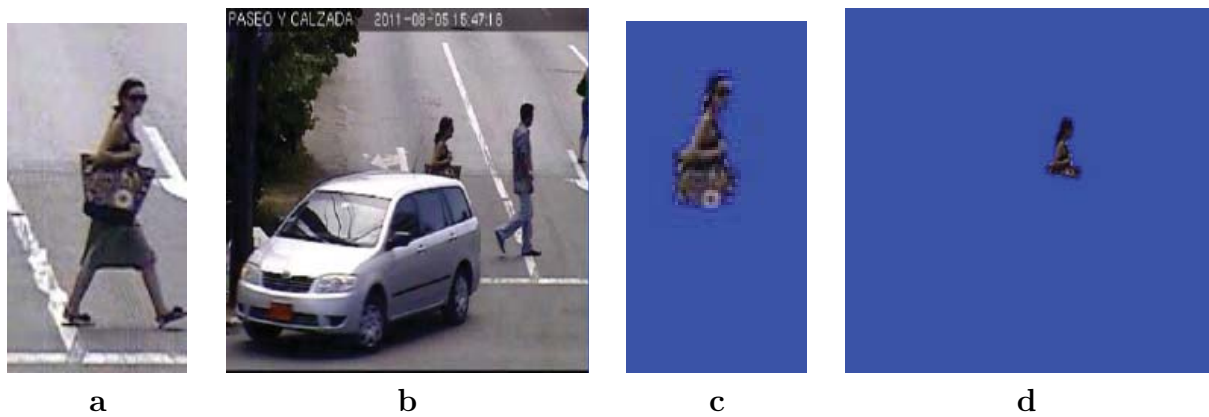


Figura A.3: Ejemplo de reconocimiento correcto ante oclusiones. a) Objeto de interés que se busca, b) cuadro de video que el algoritmo encontró como relevante para la presencia del objeto de interés, c) y d) muestran las máscaras de las regiones que correspondieron entre el objeto de interés en a) y el cuadro de video en b).





Figura A.4: Pantalla de la aplicación desarrollada para realizar experimentos en video.