Revista Internacional de Investigación e Innovación Tecnológica

Página principal: www.riiit.com.mx

Descripciones Semánticas de escenas en Imágenes mediante categorización, relaciones y distribuciones espaciales

Semantic descriptions of scenes in Images through categorization, relations and spatial distributions statistics

Vázquez-Rodríguez, C.A.*, Pinto-Elías, R.

Tecnológico Nacional de México/CENIDET, Depto. Ciencias Computacionales, Interior Internado Palmira S/N, Col. Palmira, C.P. 62490, Cuernavaca, Morelos, México.
*cvazquez@cenidet.edu.mx; rpinto@cenidet.edu.mx.

Innovación Tecnológica: Descripción semántica de imágenes con verificación sintáctica.

Área de aplicación Industrial: Automotriz, automatización de inventarios, Sistemas de seguridad.

Enviado: 19 diciembre 2019.
Aceptado: 15 marzo 2020.

Abstract

Artificial intelligence and automotive are getting closer to everyday life. It has been seen how cars are taking characteristics for a natural machine-person interaction, it is just that moment when virtual assistants come into play, it is now possible that the virtual assistant of the car reads a text message received to the cell phone, but there is a problem when what you receive is an image, then only read "an image has been received" or in the best case scenario it reads the name of the image, but it does not provide a description of their content. The images transmit information in different ways, the description can be referring to the shape of objects, colors, textures; but they can also contain concepts, stories like a stage, activities, parties, etc. which are formed by all the elements that exist in an image. For humans, semantically describing this type of concept is an easy task, however, for virtual assistants, computers, phones and machines in general it is a complex task. This paper presents an approach to semantic description of images based on categories, relationships and spatial distributions additionally applies a lexical-syntactic analysis to semantically describe the content of an image.

Key Words: Coupling, flexibility, implementation inheritance, interface inheritance.

Resumen

La inteligencia artificial y la automoción cada vez se acercan más a la vida cotidiana; se ha visto como los coches van tomando características para una interacción máquina-persona mucho más natural, ahí fue cuando los asistentes virtuales entraron en juego, ahora es posible que el asistente virtual de coche lea un mensaje de texto recibido al celular, pero existe un problema cuando lo que se recibe es una imagen, entonces, solo leerá "se ha recibido una imagen" o en el mejor de los casos leerá el nombre de la imagen pero no proporcionara una descripción del contenido de la misma. Las imágenes transmiten información de diferentes maneras, la descripción puede ser referente a la forma de los objetos, colores, texturas; pero también pueden contener conceptos, tales como un escenario, actividades, fiestas etc. los cuales están formados por todos los elementos que existan en una imagen. Para los humanos describir semánticamente este tipo de conceptos es una tarea fácil, sin embargo, para los asistentes virtuales, computadoras, celulares y máquinas en general es una tarea compleja. En este trabajo se presenta un enfoque de descripción semántica de imágenes basado en categorías, relaciones y distribuciones espaciales aplicando un análisis léxico-sintáctico para describir semánticamente el contenido de una imagen.

Palabras clave: Semántica del contenido de imágenes, análisis léxico del contenido de imágenes, representación y estructuración del conocimiento.

1. Introducción

En el sector automotriz y comercial los avances tecnológicos han revolucionado la manera de interactuar con las máquinas y la automatización de algunas tareas. En el caso de los coches, la experiencia de viaje va más allá de trasladarse de un punto a otro. En la actualidad, los coches cuentan con módulos de entretenimiento para los pasajeros y facilidades para el conductor mediante asistentes virtuales, ahora es posible reproducir música mediante voz, sin necesidad de distraerse quitando la vista del camino, ser guiados en lugares desconocidos por el sistema de navegación del coche, atender llamadas sin necesidad de tomar el teléfono, etc. Sin embargo, parece que solo un 10 % de las grandes empresas automovilísticas está implantando proyectos de inteligencia artificial. Algunos ejemplos de empresas que la emplean son: Continental, que utiliza simulaciones de IA capaces de generar cada hora datos de pruebas de vehículos equivalentes al recorrido de 5.000 millas, frente a los datos equivalentes a un recorrido de 6.500 millas que se obtenían cada mes con las pruebas anteriores de conducción física. El uso de Volkswagen de machine learning para elaborar estimaciones de ventas precisas con 250 modelos de automóviles en 120 países. Las pruebas de Mercedes-Benz con un sistema de reconocimiento visual basado en IA para los servicios de entrega de paquetería, que puede reducir el tiempo de carga de los vehículos un 15 % (Capgemini, 2019).

El recibir mensajes de texto y que un asistente virtual lea su contenido es una de los intentos de volver a los coches inteligentes con técnicas de IA, sin embargo la problemática comienza al recibir una imagen, actualmente los asistentes inteligentes de coches no describen semánticamente el contenido que tiene una imagen, ya que no se trata de una tarea trivial, La descripción de las imágenes es una tarea que ha intentado resolverse computacionalmente por muchos años, dicha tarea muchas veces es confundida con la

clasificación o reconocimiento de objetos (Mottaghi et al., 2014), el cual consiste en identificar los elementos y objetos de las imágenes (Álvarez, Salzmann & Barnes, 2014), mientras que la descripción de las imágenes va enfocada a describir el contenido de ellas como un todo, es decir, interpretar la escena presente (Xu, Schwing & Urtasun, 2014).

La descripción no ha sido una tarea fácil de realizar a lo largo de los años, se han realizado desde descripciones de bajo nivel tales como análisis del color, textura y forma (Agpal, Singh, Kaleka & Sharma, 2012); sin embargo, se detectó que esto no bastaba y se continuó el trabajo describiendo las imágenes basándose en el modelo Content-Based Image Retrieval en la década de 1990. Mediante este modelo se incorporaba exclusivamente el código lingüístico (Gudivada & Raghavan, 1995) y la tarea de recuperación se reducía esencialmente a la búsqueda de las palabras clave empleadas al describir imágenes (Zhou, Li & Tian, 2017).

Posteriormente al ver las limitaciones con el uso de etiquetas para las descripciones de imágenes, se comenzó a trabajar la denominada recuperación de información visual basada en la semántica (Álvarez et al., 2014). Este enfoque se caracteriza, por la unión de la descripción del contenido visual como colores, texturas formas etc. y la descripción lingüística en una imagen dada de la descripción humana como vendrían siendo algunos etiquetados, Ground Truth, etc. (Zhao, Shi & Wang, 2017).

La descripción semántica de imágenes mediante la recuperación de información visual basada en la semántica resultó ser una mejora al igual que los modelos ontológicos (Kong, Lin, Bansal, Urtasun & Fidler, 2014), pero en ocasiones, las descripciones podrían tener errores sintácticos de las etiquetas descriptivas, lo que llevaría a un erróneo procesamiento de la información.

Al profundizar en el campo de las propiedades de las imágenes, aumenta más la complejidad de la traducción o interpretación de la relación entre los atributos intrínsecos y objetos de una imagen (Cruz, Pérez & Cantero, 2009). Para los seres humanos describir lo que ven en una imagen, foto, escena cotidiana o interpretar alguna situación es una tarea sencilla, gracias al empleo de las palabras; en cambio, la dificultad va aumentando conforme trata de expresar con palabras los atributos de más bajo nivel como lo son el color, forma, textura. Eso provoca finalmente una disfunción en el momento de la recuperación de imágenes.

En la Figura 1 se muestra la importancia de la ubicación de los objetos dentro de una imagen, donde, se puede apreciar qué en ambas imágenes se encuentran los mismos objetos, la diferencia radica en el lugar que se encuentran cada uno de ellos lo que modifica como se relacionan.



Figura 1. Imagen con los mismos objetos, pero diferente significado.

A continuación, se describe el contenido de este trabajo. En la sección 2 el estado del arte para entrar en contexto y conocer los intentos por describir semánticamente imágenes, mostrando que aún queda campo de desarrollo y aplicación en las descripciones semánticas en imágenes. En la sección 3 se muestra como está estructurado el sistema, sus etapas y el proceso para que esté trabajo

sea reproducible. En la sección 4 se muestran la experimentación y resultados mostrando la diferencia entre una descripción de bajo nivel y un acercamiento a descripciones de alto nivel que es lo que se pretende lograr. En la sección 5 las conclusiones sobre los hallazgos y finalmente los agradecimientos y las referencias que dan soporte a este trabajo.

2. Estado del arte

Para describir imágenes es necesaria una explicación de los elementos que lo conforman. Son muy extensas las posibilidades de descripción, desde los atributos visuales básicos como el color, composición, textura, distribución de los elementos que aparecen en ella, hasta los sentimientos subjetivos, representación de una ideología o pensamiento.

Las propiedades intrínsecas de la imagen corresponden a rasgos visuales que caracterizan toda la imagen como su color, textura, forma y las relaciones espaciales; a este tipo de características suelen denominarse descripciones de bajo nivel (Civelek, Lüy & Mamur, 2017). Las propiedades extrínsecas de la imagen corresponden a todo lo contenido en la imagen no propiamente visual, estas propiedades están divididas en nivel medio y nivel alto. En el nivel medio se trabaja la detección automática de límites, contornos, objetos como rostros, sillas, coches etc. y de conceptos extraídos de la imagen como la identificación de día o de noche, si es verano o invierno. En el nivel alto se trabajan los elementos que se incluyen en los metadatos como autor, título, localización geográfica etc. (Iancu, 2018).

Si se desea expresar una consulta o realizar una búsqueda mediante un código textual, y las descripciones de las imágenes están expresadas mediante un código visual, se crea una brecha; este fenómeno recibe el nombre de vacío semántico. Para eliminar o reducir este vacío semántico es necesario favorecer la equiparación entre el código visual y las correspondientes propiedades de alto nivel mediante técnicas de retroalimentación entre otras para mejorar la respuesta del sistema (García & Cillán, 1998), También es necesario representar el conocimiento de manera estructurada.

Pueden plantearse diversas maneras de representar el conocimiento para posteriormente lograr descripciones semánticas, mediante, ontologías, lógica de primer orden, lógica difusa, etc. (Ducrot, 2000). La representación del conocimiento y el razonamiento juegan un papel central en la IA para eliminar la brecha entre humanos y máquinas y tener una comunicación natural. El tema de las descripciones semánticas se ha abordado mediante diversos enfoques a lo largo de los años, sin embargo, sigue siendo un problema sin resolver por completo, es sencillo describir los elementos que conforman una imagen, pero interpretar la escena dados los elementos es una tarea más compleja (Pérez & Merino, 2018).

Existen trabajos enfocados en describir imágenes de manera semántica o temáticas similares, el más cercano es en el trabajo *Hierarchical Scene Parsing by Weakly Supervised Learning with Image Description* se enfocan en clasificar y etiquetar semánticamente los objetos que aparecen en una imagen. Estructuran la información en jerarquías de objetos y su posición, mediante redes neuronales convolucional (CNN) con la cual extraen los objetos y los clasifican. Adicionalmente usan una red neuronal recursiva (RsNN) que identifica la sobreposición de los objetos para deducir el orden de aparición en los planos de la imagen

(Zhang, Liang, Wang, Wang & Zuo, 2017). En la Figura 2 se muestra el etiquetado semántico que realizan.

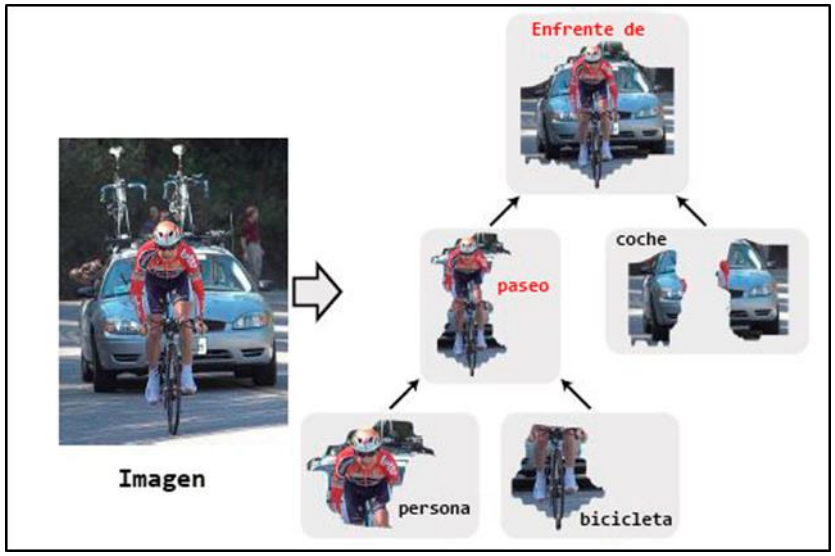


Figura 2. Configuración de escena estructurada.

En la Figura 2 se muestra como estructuran la información de la imagen, las etiquetas en negro son objetos, mientras que las etiquetas en rojo corresponden a relaciones de interacción entre los objetos.

En la siguiente sección se describirá como está construido el sistema de descripción semántica y el proceso que sigue una imagen para lograr ser descrita.

3. Estructura del sistema

El sistema recibe información de tres formas: la primera es mediante Ground Truth de Pascal VOC (Everingham, Eslami, Van Gool, Williams, Winn & Zisserman, 2015), las cuales son imágenes con colores definidos asignados para cada clase. La segunda es mediante la clasificación de imágenes proporcionada por el clasificador YOLOv3, (Redmon, Joseph, Farhadi & Ali, 2018) utiliza redes neuronales convolucionales. Y la tercera forma es mediante archivos de .txt proporcionados por cualquier sistema de clasificación. Cabe mencionar que este trabajo no es de clasificación, sino, el paso siguiente, estructurar la información obtenida por los clasificadores para que tenga un sentido coherente como conjunto y realizar descripciones semánticas. En la Figura 3 se muestra el diagrama del proceso de la metodología realizada para la descripción semántica.

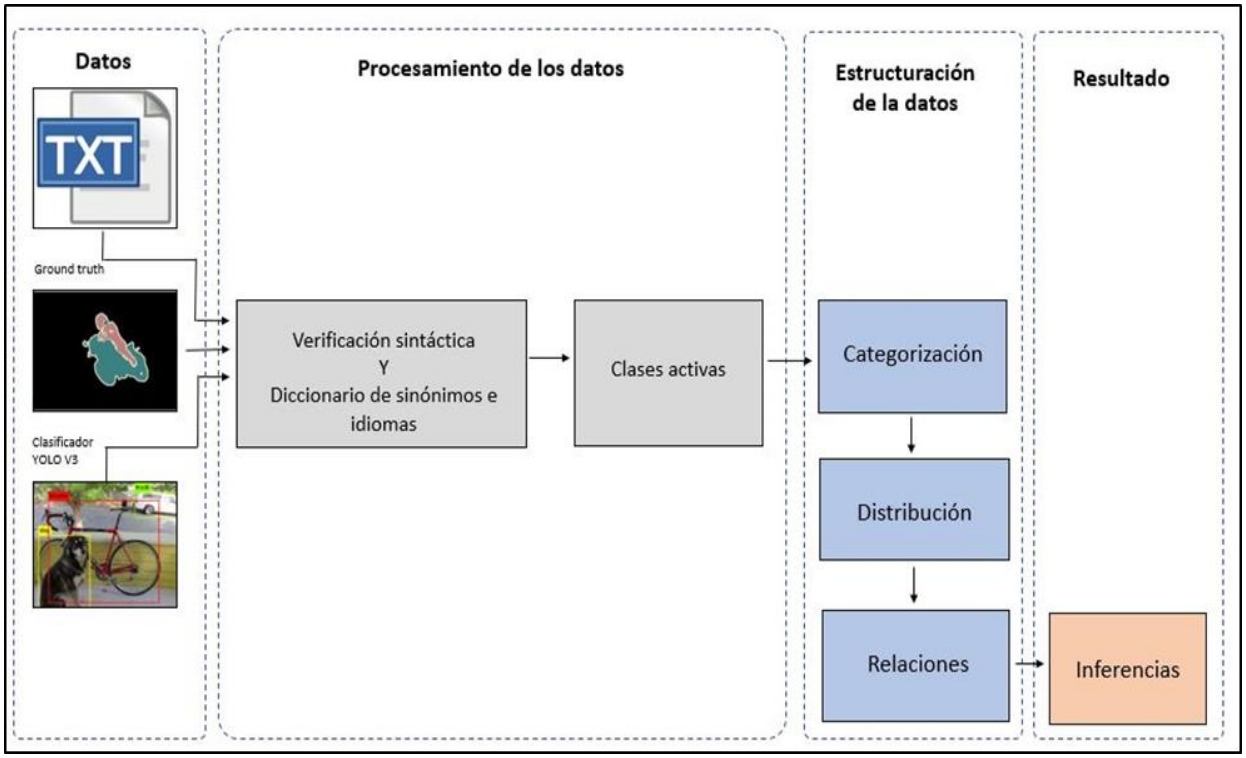


Figura 3. Diagrama del proceso de la descripción semántica en imágenes mediante la estructuración de los datos.

El proceso ha sido dividido en cuatro secciones: datos, procesamiento de los datos, estructuración de los datos y finalmente la descripción semántica de las escenas contenidas en una imagen.

3.1. Datos

La recepción de datos se divide en tres. La primera es mediante un archivo de texto de cualquier sistema clasificador que desee usar este proceso para la descripción semántica del contenido, a partir de ahí la información pasa al procesamiento de los datos. La segunda mediante Ground truth de Pascal Voc, en este caso lo que el sistema recibe es una imagen validada con etiquetado por código de color RGB, donde, cada color representa una clase, esa información es extraída y almacenada en un archivo de texto plano y a partir de ahí pasa a la siguiente etapa de procesamiento de datos. Finalmente, la tercera entrada al sistema es utilizando el clasificador YOLOv3, el cual está integrado en el sistema y es capaz de reconocer alrededor de 10,000 clases diferentes (Redmon et al, 2018). Incluso es posible tomar una foto y describirla inmediatamente.

El contar con diversas maneras de alimentar el sistema brinda un abanico de posibilidades en cuanto aplicaciones para la descripción de imágenes, no solamente en el sector automotriz sino también podrían ser en video vigilancia, como herramienta de apoyo a personas invidentes de nacimiento o que paulatinamente ha perdido el sistema de la vista, en descripción de la escena en guarderías para detectar un inminente peligro por objetos, animales o personas, etc.

3.2. Procesamiento de los datos

Se divide en dos procesos: El primero la verificación sintáctica y diccionario de sinónimos e idiomas. En esta etapa al recibir los datos se realiza una verificación si alguno de los datos recibidos corresponde a alguna de las 10,000 clases con las que se está trabajando, tras terminar la búsqueda si no encuentra ninguna clase, se verifica en el diccionario de sinónimos y en el diccionario de idiomas, por el momento únicamente se está trabajando con español e inglés. Se agrego este módulo debido a la riqueza y variabilidad en el lenguaje incluso en el mismo idioma, por ejemplo, en el norte de México se utilizan definiciones diferentes con respecto al sur para algunos objetos. Y este fenómeno repite el patrón en otros idiomas.

El segundo activa las clases detectadas en el paso anterior y carga sus estructuras de datos, se está trabajando con árboles.

3.3. Estructuración de los datos

Se divide en tres etapas. La primera etapa de la estructuración de los datos es la categorización, se creó una plantilla genérica de estructura de datos para las clases, de esta manera el software es adaptable y expandible. En la Figura 4 se muestra una imagen representativa del contenido de la estructura que almacena las distribuciones.

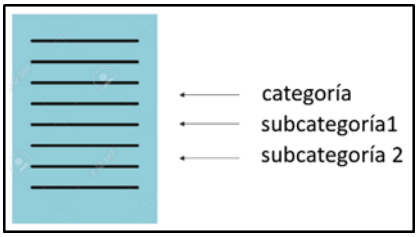


Figura 4. Imagen representativa de la estructura de la platilla de categorización.

En esta etapa del proceso, se estructura la información, con categorías y subcategorías, en algunos casos no aparecerá la categoría principal en una imagen, pero si alguna subcategoría y es importante saber a qué categoría pertenece, para poder realizar búsquedas semánticas en las imágenes. En la Figura 5 se muestra un ejemplo de cómo está estructurada la información para la clase entorno, al ser un árbol puede seguir expandiéndose para agregar más categorías y subcategorías.

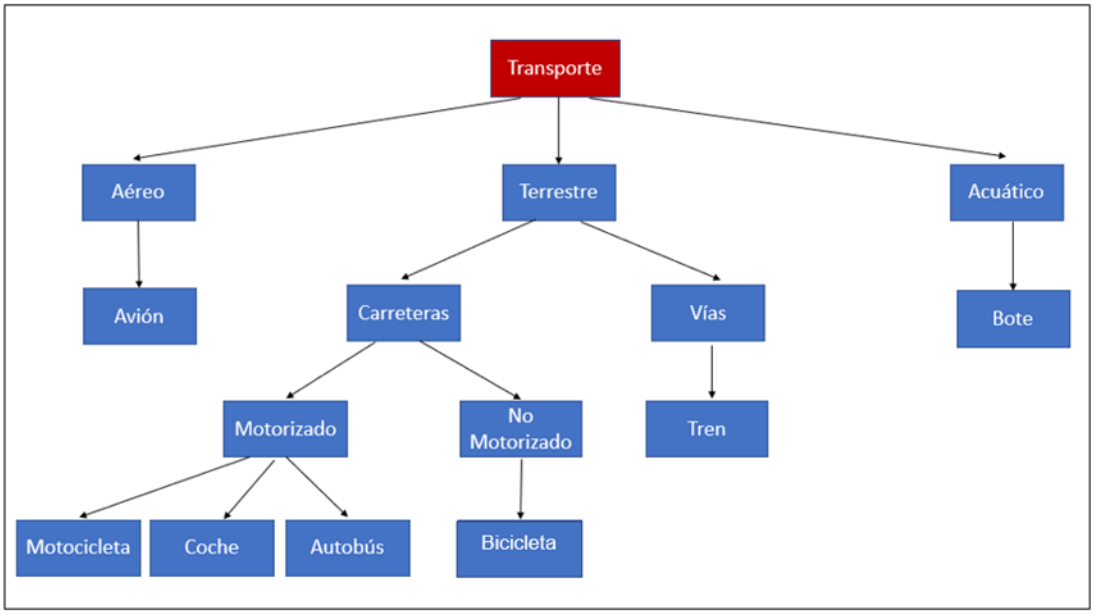


Figura 5. Árbol de estructuración de la clase transporte.

La información es cargada a la RAM mediante el gestor de base de datos REDIS (Nelson, 2016) para su rápido acceso y consulta de información, ya que Redis es un almacén de estructura de datos en memoria de código abierto, que se utiliza como agente de base de datos, caché y mensaje y admite estructuras de datos como cadenas, hashes, listas, conjuntos, conjuntos ordenados con consultas de rango, mapas de bits, hyperloglog, índices geoespaciales, etc., facilita el proceso de trabajar con millones de datos.

La segunda etapa de la estructuración de los datos es la distribución, en esta etapa se utilizan las coordenadas superiores e inferiores que pertenecen a cada objeto y con ellos se calcula su centroide para tener el objeto ubicado en *x*, *y* en la escena de la imagen. En la Figura 6 se muestra una imagen representativa del contenido de la estructura que almacena las distribuciones espaciales.

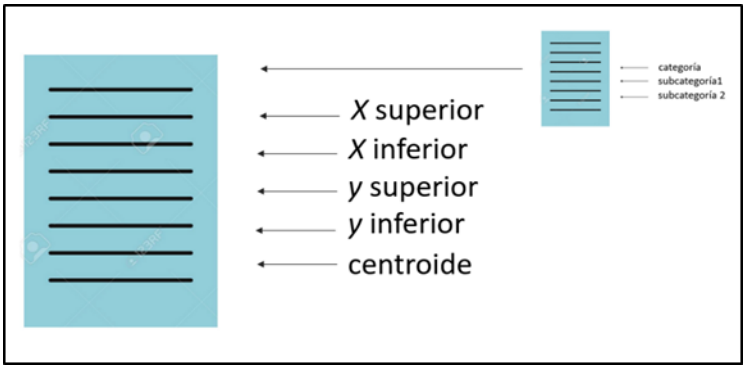


Figura 6. Imagen representativa de la estructura de la platilla de distribuciones espaciales.

La tercera etapa de la estructuración de los datos son las relaciones, es aquí donde todos los árboles de todas las categorías se unen y son recorridos por listas circulares y mediante teoría de conjuntos se evalúan las clases que se interceptan entre sí, posteriormente se calcula la distancia euclidiana (Liwei, Yan & Jufu, 2005) entre cada conjunto perteneciente a las categorías y posteriormente realizar inferencias de lo que está sucediendo en la escena mediante una descripción semántica. En la Figura 7 se muestra una imagen representativa del contenido de la estructura que almacena las relaciones.

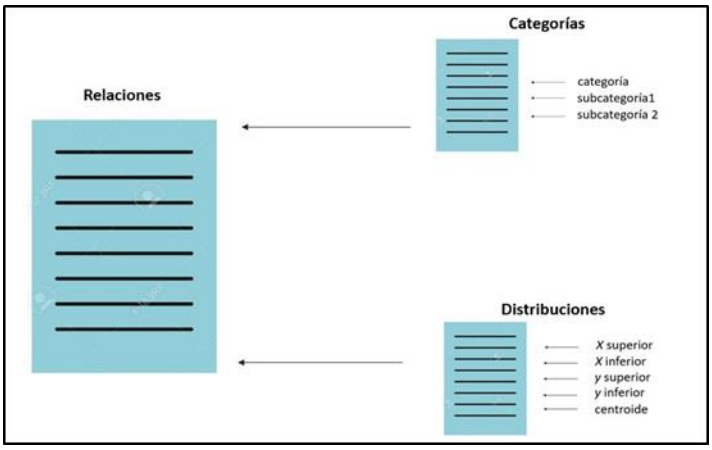


Figura 7. Imagen representativa de la estructura de la platilla de relaciones.

En la sección 4 se presenta la cuarta etapa del proceso, el cual corresponde al resultado. Se realizó experimentación con diversas imágenes para conseguir sus descripciones como un todo y no como un listado de objetos que aparecen. Dadas las estructuras de datos que contienen la categorización, distribución y relaciones fue posible realizar interpretaciones sobre las escenas en imágenes infiriendo las acciones y comportamientos mediante lógica de predicados y lógica proposicional.

4. Experimentación y resultados

La experimentación se llevó a cabo con diversas imágenes con mayor y menor riqueza para mostrar la importancia de la información para describir semánticamente, no todas las imágenes se pueden describir de esta manera, habrá casos donde únicamente se tiene un objeto y no se pueda inferir o describir más allá de su categoría como se muestra en la Figura 8, lo que representa descripciones de bajo nivel, que no proporcionan una interpretación.

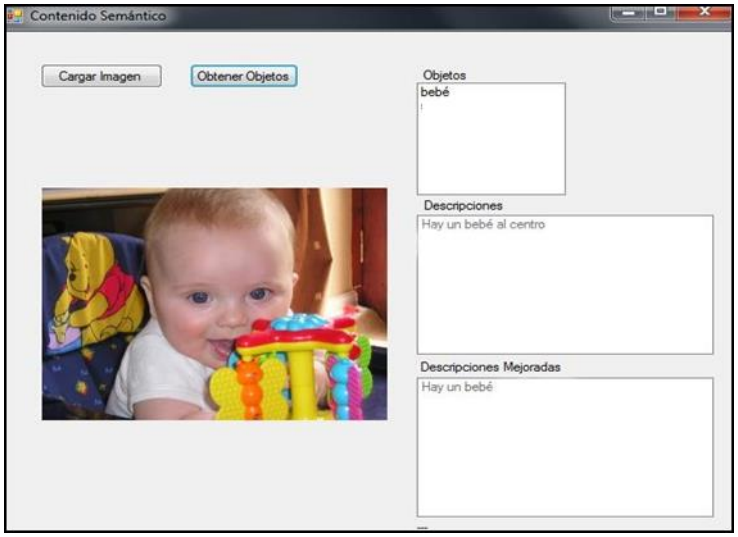


Figura 8. Imagen de prueba con escasa información para una descripción semántica.

Sin embargo, en otros casos las imágenes cuentan con varios objetos y relacionados entre ellos como se muestra en la Figura 9.

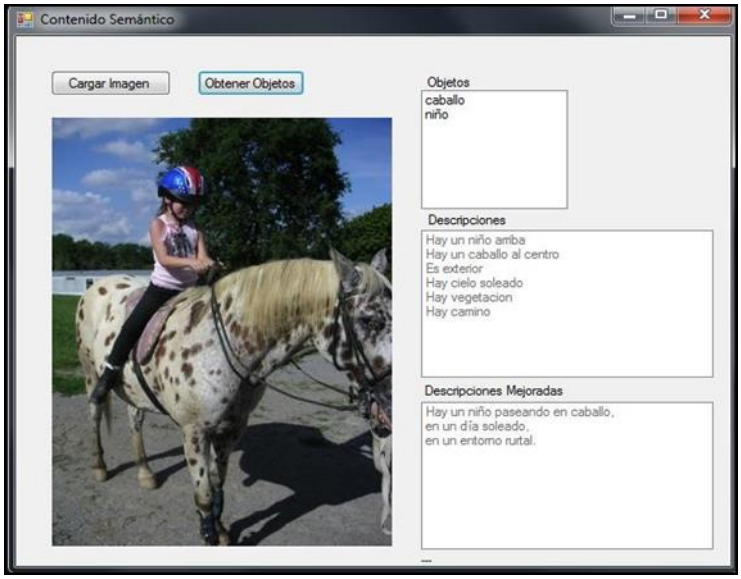


Figura 9. Imagen de prueba con información de entorno y diversas categorías para una descripción semántica.

Como se puede apreciar en la Figura 9 al tener mayor riqueza de categorías es posible describir mejor una imagen. En este caso se consideró la variable entorno, la cual, es muy importante porque permite deducir en qué escenario se está llevando a cabo el evento enmarcado en una imagen. En la Figura 10 se muestra el resultado de la descripción de otra imagen con riqueza de categorías.

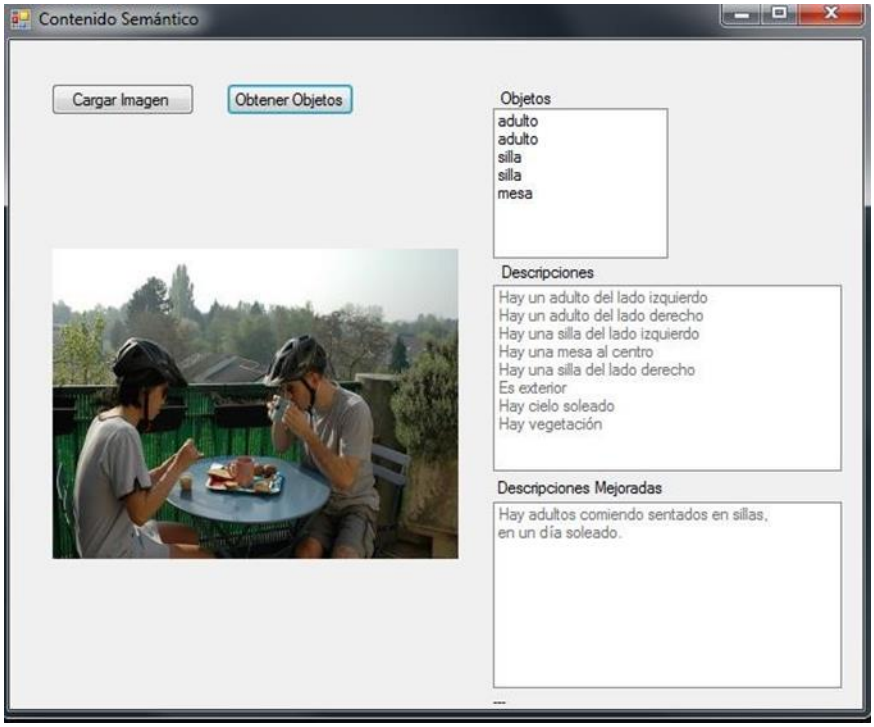


Figura 10. Imagen de prueba con información de entorno y diversas categorías para una descripción semántica.

Como se puede apreciar en la Figura 10 la descripción de la escena es más natural, similar a como lo realizan las personas; esto es lo que se busca al describir las imágenes de manera semántica.

En todos los casos se muestra en el primer recuadro los objetos encontrados, en el segundo se describe su entorno y sus distribuciones espaciales y finalmente al tomar en cuenta como se relacionan dadas sus proximidades e interacciones es posible inferir que está sucediendo en la escena. Estas inferencias son realizadas mediante lógica proposicional y cálculo de predicados, de la misma manera que Prolog (Shaalan, K., 2019) realiza inferencias por pares de información.

Los resultados obtenidos muestran que al tomar en cuenta las imágenes como un todo y no como objetos aislados se puede obtener información importante para describir la imagen como un todo, incluso podrían realizarse inferencias sobre lo que sucede en la escena que conforman la imagen, por ello es muy importante la verificación sintáctica ya que de esta manera se evitó añadir ruido al sistema y se trabajó solo con elementos válidos.

5. Conclusiones

La riqueza de las descripciones será mayor cuando haya más objetos en las imágenes, cuando tengan visibles detalles del entorno tales como: vegetación, cielo, edificios, etc. Se lograron descripciones de imágenes de una manera más natural con respecto a enunciar cada uno de los elementos que logran aparecer en las imágenes, si bien queda mucho trabajo por realizar este nivel de descripción puede ser muy útil en diversos ámbitos, como en asistentes

virtuales en automóviles, sin embargo, su campo de aplicación son infinitos, desde búsquedas semánticas, automatización, video vigilancia, en sistemas tutores inteligentes y en móviles con cámara apoyado de sintetizador de voz como herramienta de ayuda a personas con discapacidad visual, para que puedan conocer que hay a su alrededor.

6. Agradecimientos

Al CONACyT por su apoyo para la realización de este trabajo enmarcado en el Doctorado en Ciencias de la Computación en Tecnológico Nacional de México/ CENIDET.

7. Referencias

1. Agpal, J., Singh, J., Kaleka, S. & Sharma, R., (2012). Different Approaches of CBIR Techniques. International Journal of Computers & Distributed Systems 1(2), 76-78.

2. Capgemini Research Institut., (2019). Accelerating automotive's AI transformation.

3. Civelek, Z., Lüy, M. & Mamur, H., (2017). A new fuzzy logic proportional controller approach applied to individual pitch angle for wind turbine load mitigation, Renewable Energy, 111, 708-717.

4. Cruz, M. N., Perez, M. P. & Cantero, T. C., (2009). Influencia de la motivación intrínseca y extrínseca sobre la transmisión de conocimiento. CIRIEC-España, Revista de Economía Pública, Social y Cooperativa, 187-211.

5. Ducrot, O., (2000). La elección de las descripciones en semántica argumentativa léxica. Revista iberoamericana de discurso y sociedad, 2(4), 23-44.

6. García, M., Cillán, J., (1998). Las señales de calidad: atributos extrínsecos del producto. Revista anual de estudios económicos y empresariales, 81-116.

7. Gudivada, V. N. & Raghavan V. V., (1995). Content based image retrieval systems. Computer, 28(9), 18-22. doi: 10.1109/2.410145.

8. Iancu, I., (2018). Heart disease diagnosis based on mediative fuzzy logic. Artificial Intelligence in Medicine, 89, 51-60. doi: 10.1016/j.artmed.2018.05.004.

9. Kong, C., Lin, D., Bansal, M., Urtasun, R. & Fidler, S., (2014). What Are You Talking About? Text-to-Image Coreference. In 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 3558-3565. Doi: 10.1109/CVPR.2014.455.

10. Liwei Wang, Yan Zhang and Jufu Feng., (2005). On the

Euclidean distance of images, in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 8, pp. 1334-1339. doi: 10.1109/TPAMI.2005.165.

11. Mottaghi, R., Chen, X., Liu, X., Cho, N. G., Lee, S. W., Fidler, S. & Yuille, A., (2014). The role of context for object detection and semantic segmentation in the wild. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 891-898. IEEE Computer Society. doi.org/10.1109/CVPR.2014.119.

12. Nelson, J., (2016). Mastering Redis, Packet Publishing Ltd, ISBN: 978178398819.

13. Pérez, P. & Merino, M., (2018). Definición de semántica. Recuperado a partir de https://definicion.de/semantica.

14. Redmon, Joseph and Farhadi & Ali., (2018). YOLOv3: An Incremental Improvement., Computer Science arXiv.

15. Ruimao Zhang, Liang Lin, Guangrun Wang, Meng Wang, and Wangmeng Zuo., (2018). Hierarchical Scene Parsing by Weakly Supervised Learning with Image Descriptions., IEEE Transactions on Pattern Analysis and Machine Intelligence.

16. Shaalan, K., (2019). Extending prolog for better natural language analysis. In Proceeding of the 1st Conference on Language Engineering, Egyptian Society of Language Engineering (ELSE) (pp. 225-236).

17. Zhao, H., Shi & J., Wang., (2017). Pyramid Scene Parsing Network. IEEE Conference on Computer Vision and Pattern Recognition, 6230-6239. doi:10.1109/CVPR.2017.660.

18. Zhou, W., Li, H. & Tian, Q., (2017). Recent Advance in Content-based Image Retrieval: A Literature Survey. International Journal Advanced Networking and Applications, 10(1), 3741-3757.