EDUCACIÓN
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLÓGICO NACIONAL DE MÉXICO

Tecnológico Nacional de México

Centro Nacional de Investigación y Desarrollo Tecnológico

Tesis de Doctorado

Modelo para la descripción del contenido semántico de imágenes

presentada por

M.C. Catalina Alejandra Vázquez Rodríguez

como requisito para la obtención del grado de

Doctora en Ciencias de la Computación

Director de tesis

Dr. Raúl Pinto Elías

Cuernavaca, Morelos, México. Marzo 2022.

cenidet
Centro Nacional de Investigación y Desarrollo Tecnológico

EDUCACIÓN
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLÓGICO NACIONAL DE MÉXICO

Centro Nacional de Investigación y Desarrollo Tecnológico
Departamento de Ciencias Computacionales

ESC\FORDOC09

Cuernavaca, Morelos, 08/febrero/2022

ASUNTO: ACEPTACIÓN DEL TRABAJO DE TESIS DOCTORAL

DR. JUAN GABRIEL GONZÁLEZ SERNA
JEFE DEL DEPARTAMENTO DE CIENCIAS COMPUTACIONALES
PRESENTE

Los abajo firmantes, miembros del Comité Tutorial de la Tesis Doctoral de la alumna **M.C. CATALINA ALEJANDRA VÁZQUEZ RODRÍGUEZ** manifiestan que después de haber revisado su trabajo de tesis doctoral titulado **"MODELO PARA LA DESCRIPCIÓN DEL CONTENIDO SEMÁNTICO DE IMÁGENES"**, realizado bajo la dirección del **Dr. Raúl Pinto Elías,** el trabajo se ACEPTA para proceder a su impresión.

ATENTAMENTE
"Excelencia en Educación Tecnológica®
"Educación Tecnológica al Servicio de México"

DR. RAÚL PINTO ELÍAS
CENIDET

DRA. ANDREA MAGADÁN SALAZAR
CENIDET

DR. DANTE MÚJICA VARGAS
CENIDET

DR. NOÉ ALEJANDRO CASTRO SÁNCHEZ
CENIDET

DR. JORGE ALBERTO FUENTES PACHECO
INST. DE INVESTIGACIÓN EN CIENCIAS BÁSICAS

C.c.p.: Lic. Silvia del Carmen Ortiz Fuentes/ Jefa del Depto. de Servicios Escolares
Dr. Carlos Manuel Astorga Zaragoza / Subdirector Académico
Expediente

EDUCACIÓN
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLÓGICO NACIONAL DE MÉXICO

Centro Nacional de Investigación y Desarrollo Tecnológico
Subdirección Académica

Cuernavaca, Mor., 10/marzo/2022
No. De Oficio: SAC/56/2022
Asunto: Autorización de impresión de tesis

CATALINA ALEJANDRA VÁZQUEZ RODRÍGUEZ
CANDIDATA AL GRADO DE DOCTORA EN CIENCIAS DE LA COMPUTACIÓN
P R E S E N T E

Por este conducto, tengo el agrado de comunicarle que el Comité Tutorial asignado a su trabajo de tesis titulado **"MODELO PARA LA DESCRIPCIÓN DEL CONTENIDO SEMÁNTICO DE IMÁGENES"**, ha informado a esta Subdirección Académica, que están de acuerdo con el trabajo presentado. Por lo anterior, se le autoriza a que proceda con la impresión definitiva de su trabajo de tesis.

Esperando que el logro del mismo sea acorde con sus aspiraciones profesionales, reciba un cordial saludo.

EDUCACIÓN
11 MAR 2022
CENTRO NACIONAL DE INVESTIGACIÓN Y DESARROLLO TECNOLÓGICO
SERVICIOS ESCOLARES
RECIBIDO
EBN

A T E N T A M E N T E
Excelencia en Educación Tecnológica®
"Educación Tecnológica al Servicio de México"

EDUCACIÓN
CENTRO NACIONAL DE INVESTIGACIÓN Y DESARROLLO TECNOLÓGICO
SUBDIRECCIÓN ACADÉMICA

DR. CARLOS MANUEL ASTORGA ZARAGOZA
SUBDIRECTOR ACADÉMICO

C. c. p. Departamento de Ciencias Computacionales
Departamento de Servicios Escolares

CMAZ/CHG

DEDICATORIA

A mis padres por su apoyo incondicional, formación, por siempre alentarme a seguir adelante y crecer. Gracias, infinitas gracias.

A Luis por apoyarme cuando no encontraba la salida.

A mis niñas Lilian y Lupita por su compañía en los últimos semestres y entretenerme con sus ocurrencias cuando se ponía pesado.

A Martin por siempre estar ahí apoyándome en todos los aspectos de la vida y motivándome.

A mi médico de cabecera, Roció, por su diagnóstico certero y seguimiento incondicional que me permitieron continuar con este doctorado, atendiéndome siempre con mucho cariño.

A todos ustedes infinitas gracias, sin ustedes esta etapa aun estaría sin concluir.

AGRADECIMIENTOS

Al Dr. Raúl Pinto Elías por su seguimiento, orientación y consejos, que sin duda alguna han sido de gran apoyo en mi formación doctoral.

A mi comité revisor por dedicarle tiempo a cada una de las etapas de mi doctorado, por sus revisiones y retroalimentaciones.

Al Consejo Nacional de Ciencia y Tecnología, quienes me brindaron el subsidio económico, que sin él todo hubiera sido aún más difícil.

Al Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET) de México, que me dio la oportunidad de continuar con mi formación académica.

Resumen

La descripción semántica de imágenes es una tarea que se ha intentado resolver de diferentes formas a lo largo de los años debido a la complejidad que representa para una computadora describir conceptos. La complejidad no solo radica en lograr una descripción semántica a través de los sistemas informáticos, sino que también existe la complejidad en la descripción semántica que dan los humanos y las variaciones que se pueden mostrar entre ellos; estas diferencias están condicionadas al conocimiento y otros factores tales como: la edad, el grado de estudio, el pensamiento analítico, el entorno, etc.

El modelo propuesto ha mostrado la capacidad de describir semánticamente el contenido de las imágenes con diferentes niveles de detalle mediante estructuración, jerarquización de la información y un mecanismo de inferencia formado por reglas de producción, de esta manera se proporciona una descripción del contenido de las imágenes.

Los resultados obtenidos con el modelo se compararon con las respuestas extraídas del banco de representación del conocimiento RIDeCS, en el que las personas pueden describir lo que ven en las imágenes y registrar sus edades; luego de un análisis comparativo y de medición con la métrica “similitud de oraciones basada en fragmentos y contenido”, se concluyó que el modelo propuesto es capaz de describir imágenes de manera similar a los humanos. Al contemplar jerarquías, las descripciones tuvieron una similitud con las respuestas de niños de 7 a 14 años y las descripciones generales tuvieron una similitud que corresponde a personas de 38 a 42 años, finalmente el parecido entre el modelo de descripción semántica y las descripciones de humanos en RIDeCS fue cercano al 98% en promedio.

Las descripciones semánticas proporcionadas por el modelo también tuvieron similitud con otros trabajos, sin embargo, en algunos casos para una misma imagen la descripción fue diferente entre humanos, el modelo propuesto y otros autores, esto se debe a la interpretación de la escena de parte de cada uno.

Abstract

The semantic description of images is a task that has been trying to solve in different ways over the years due to the complexity for a computer to describe concepts. The complexity not only lies in achieve a semantic description through the computer systems, but also exist the complexity in the sematic description given by humans and variations that can be shown between one and another; these differences are conditioned to the knowledge of each one, as well as to other factors, such as age, degree of study, and analytical thinking, environment.

The proposed model has shown the ability to describe the content of the images semantically and with different levels of detail by structuring the information and using it to describe through an inference mechanism formed by production rules including an interpretation of the content of the images and creating a hierarchical semantic description.

The results obtained with the model were compared to human responses taken from the RIDeCS knowledge representation bank, in which people can describe what they see in the images and record their ages. After measurement with "Sentence similarity based on fragments and content," it was possible to conclude the similarity between the proposed model and answers by humans. When hierarchies were considered, the descriptions have a similarity to the children aged 7 to 14 years old; the general descriptions have a similarity that corresponds to people aged 38 to 42 years. Finally, the resemblance between the model for semantic description and descriptions of humans in RIDeCS was near to 98% the percent of similarity was diverse depended on characteristics in the images.

The semantic descriptions provided by the model also had similarities with other works; however, in some cases for the same picture, the description was different between humans, the proposed model, and other authors, this is due to the interpretation of the scene by each one.

Contenido

Índice de figuras

Índice de tablas

Glosario

BOV	Bag of visualterms
CBIR	Content-based Image Retrieval
CLD	Colour Layout Descriptor
EHD	Edge Histogram Descriptor
HSV	Hue, Saturation, Value
HSL	Hue, Saturation, Lightness
HTML	HyperText Markup
IA	Atributos de Imagen
KNN	k-nearest neighbors
LCSP	Combining locality- constrained sparse coding
MCM	Mínimo común múltiplo
MPEG	Moving Picture Experts Group
PHOW	Pyramid histogram of visual words
RGB	Red, green, blue
RIDeCS	Repositorio de Imágenes para la Descripción de Contenido Semántico
RNA	Redes neuronales artificiales
SCD	Scalable Colour Descriptor
SPP	Spatial Pyramid Pooling
SURF	Speeded Up Robust Features
SVM	Máquinas de vector soporte
VIRS	Visual Information Retrieval Systems
VQ	Cuantificación vectorial
YCrCb	Espacios de color usada en sistemas de vídeo y fotografía digital

CAPÍTULO 1

INTRODUCCIÓN

La descripción semántica de imágenes es una tarea que se ha intentado resolver de diferentes formas a lo largo de los años debido a la complejidad que representa para una computadora describir conceptos. La complejidad no solo radica en lograr una descripción semántica a través de los sistemas informáticos, sino que también existe la complejidad en la descripción semántica que dan los humanos y las variaciones que se pueden mostrar entre ellos; estas diferencias están condicionadas al conocimiento de cada uno, así como a otros factores como la edad, el grado de estudio, el pensamiento analítico, el entorno, etc. En el presente trabajo se propone un modelo para la descripción semántica del contenido de las imágenes, el cual, considera como parte importante la jerarquización de la información para dar descripciones con diversos niveles de detalle, similares a como lo harían personas de diferentes edades acorde al conocimiento que han adquirido a lo largo de su vida.

1.1 Antecedentes

La descripción semántica de imágenes es una tarea que se ha trabajado de diversas maneras debido a la complejidad que conlleva para una computadora describir conceptos, los cuales, suelen ser adquiridos mediante el conocimiento que los humanos incorporan en los bancos de información. Inicialmente existieron las descripciones de bajo nivel, donde, los protagonistas eran el color, textura y forma [1]; sin embargo, se detectó que esto no era suficiente, lo que dio paso al modelo de recuperación de imágenes basada en contenido a inicios de los noventa [2]. Mediante este modelo se incorporaba código lingüístico y la tarea de recuperación se reducía a la búsqueda de las palabras clave o etiquetas [3], sin embargo, las limitaciones iniciaron los trabajos de recuperación de información visual basada en la semántica [4], este enfoque se caracteriza por la unión de la descripción del contenido visual como colores, texturas formas y la descripción lingüística de una imagen usando etiquetas [5].

La complejidad del contenido semántico de las imágenes no solo radica en la descripción automática de los sistemas de cómputo, sino en la descripción que puede darse entre diferentes personas para una misma imagen y la cual estará condicionada al conocimiento de cada uno, así como a otros factores, tales como: la edad, grado de estudio, pensamiento analítico, entorno, etc. [6]. Cada persona selecciona y organiza la información de manera diferente en virtud de lo percibido y de sus particulares estructuras cerebrales. Una primera diferencia está en la cantidad de información que la persona capta del estímulo percibido, esto es, desde su amplitud de aprehensión la cantidad de información correctamente identificada y recordada tras una exposición a imágenes, sonidos y sabores que generan una representación fáctica de la realidad en la mente [7].

Como los órganos sensoriales encargados de la percepción presentan diferencias en cada persona, también se tiene distinta capacidad de captación de los estímulos [8] y como no es posible aprehender toda la inmensa cantidad de información disponible sobre un suceso, los seres humanos mediante el mecanismo de la atención captan

sólo una parte concreta de las escenas, la parte que se considera más importante o reconocida por el individuo y que puede procesar con eficacia y, en virtud de ésta, obtendrá más o menos información, por ejemplo, en una habitación una persona puede captar diez objetos en los que fija su atención, mientras que otra puede fijarla en veinte, lo que incrementa los datos de entrada de información a procesar [9], lo mismo ocurre al momento de describir semánticamente imágenes, cada persona centrará su atención en lo que conoce y es de su interés, lo que conlleva a diversas interpretaciones de una misma escena, por ello surge la problemática de cómo medir la descripción de una imagen dado que la misma puede tener diversas respuestas correctas.

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 [10]. 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 realiza 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, verano o invierno, etc.

En el nivel alto se trabajan los elementos que se incluyen en los metadatos como autor, título, localización geográfica, fecha, formato y propiedades de carácter subjetivo denominadas semánticas extraídas a través de la observación de la imagen y dando una interpretación de ellas, las cuales suelen incorporarse en el apartado de descripciones o notas en los metadatos. El elemento de consultas permite al usuario expresar su necesidad informativa, existen varios métodos que permiten introducir la consulta: puede ser mediante texto, ejemplos, o navegación por la colección, entre otros [11].

1.2 Antecedentes de esta investigación

A continuación, se mencionan algunos de los trabajos relacionados culminados en CENIDET, referentes a la semántica de las imágenes.

La tesis "*Recuperación Automatizada de Imágenes mediante la Implementación de Descriptores del Estándar MPEG-7"* de Carlos Pérez Lara [20], realizada en 2014, la cual consiste en el desarrollo de un sistema CBIR (*Content-Based Image Retrieval*) empleando descriptores visuales del estándar MPEG-7. En esta investigación se seleccionaron dos descriptores del estándar MPEG-7, uno de color y otro de textura, *Color Layout Descriptor & Edge Histogram Descriptor.*

Para cada una de las pruebas realizadas se utilizaron tres *data sets* de imágenes mencionados en la literatura:

- *Common Color Dataset (CCD), Corel_1k*
- *Uncompressed Colour Image Dataset (UCID)*
- Imágenes capturadas en CENIDET

La tesis "*Indexado y Recuperación de Imágenes por Contenido"*, de Perla Troncoso Rey, en la cual se presentó una metodología para la recuperación automática de imágenes, y se desarrolló un sistema que realiza caracterización mediante descripciones parciales de los elementos de la imagen mediante estructuras de árboles. Incluye información sobre categorías, subcategorías e imágenes que pertenecen a cada una de ellas y se dividen en dos niveles:

- El primero se enfoca en determinar el contenido de la imagen, para lo cual se utiliza un algoritmo de votación.

- El segundo nivel tiene como objetivo clasificar la escena presente en la imagen a partir del contenido encontrado en ella y utilizando las reglas obtenidas a partir del índice [21].

Las tesis antes descritas han trabajado la descripción de imágenes desde el enfoque de etiquetado, tomando en cuenta la textura y el color, así como forma.

1.3 Descripción del problema

La descripción semántica de las imágenes es una tarea compleja debido a que no existe un modelo que proporcione el protocolo a seguir para realizar dicha tarea; además, una misma imagen puede tener diversos niveles de descripción, algunos más extensos que otros y no existe algún algoritmo, regla o norma que dicte cuál es el nivel de detalle de descripción correcto.

1.3.1 Complejidad del problema

La complejidad radica en la similitud que existe entre imágenes debido a su contenido, es decir, dos imágenes pueden tener exactamente los mismos elementos; sin embargo, lo que las podría hacer diferentes es el tamaño de los objetos, posición de los objetos dentro de la imagen, posición de los objetos entre ellos mismos y el cómo se relacionan.

1.4 Propuesta de solución

El problema presentado se resolvió mediante el diseño de un modelo para generar descripciones semánticas que permite describir y estructurar el contenido de las imágenes mediante: categorización, distribuciones espaciales, relación entre objetos y jerarquías, logrando generar descripciones de las imágenes desde una perspectiva semántica con base en el análisis de los objetos contenidos y sus interacciones, además al jerarquizar el contenido de las imágenes es posible obtener descripciones con diferentes niveles de detalle y/o diversas descripciones para una misma imagen.

1.5 Planteamiento de la solución

En este trabajo se presenta el proceso realizado para las pruebas con el modelo propuesto y consta de cuatro etapas: Entrada, Procesamiento de los datos, Modelo y Salida. Es importante mencionar que la aportación de este trabajo es el modelo.

La entrada: corresponde a los datos que ingresan al modelo, se implementó que la recepción de los mismo fuera de diversas maneras considerando texto, *Ground truth*, y

clasificador, esto para mostrar la versatilidad del modelo al trabajar con diferentes tipos de entradas, es importante aclarar que el procesamiento de la imagen para obtener las características y demás información no es parte de este trabajo, se utilizan para alimentar el modelo.

Procesamiento de los datos: Una vez que los datos han sido recibidos, es necesario comprobar que las categorías pertenecen al universo de datos utilizado. En este caso para la experimentación se trabajó con 702 categorías organizadas en seis súper categorías: *personas, animales, transporte, alimentos, objetos y entorno*. Los datos que entran al modelo llevan un proceso de verificación sintáctica, se evalúa si la categoría recibida se encuentra entre las registradas en español, si no hay coincidencia, entonces se revisa el diccionario de sinónimos e idiomas, por el momento los idiomas con los que se está trabajando son inglés y español. Si alguna categoría coincide se carga el árbol de esta, de lo contrario se envía un mensaje de notificación de coincidencias nulas.

Modelo: En esta etapa es donde se encuentra el aporte de este trabajo. El modelo está dividido en cinco etapas.

1. Categorías: La categorización consiste en organizar la información de las categorías activas que ingresaron al modelo y verificar a cuál super categoría pertenece para cargar sus árboles y sus respectivas reglas.

2. Jerarquías: Se encarga de estructurar la información de las categorías, en algunos casos se tendrá mayor detalle que en otros, esto dependerá de la cantidad de información que reciba el modelo, por ejemplo, en el caso de una computadora, puede ser laptop o de escritorio y a su vez cada una de ellas tener jerarquías tales como marca, color, etc., una vez que la información ha sido estructurada, es enviada a la etapa de distribuciones espaciales.

3. Distribución espacial: Se encarga de almacenar la posición de cada uno de los objetos de las imágenes, independientemente de que sea la categoría principal o

una subcategoría se almacena su posición superior, inferior y el centroide de la región que contiene al objeto.

4. Relaciones: Se relaciona toda la información contenida en una imagen, y se encarga de que cada una de las categorías consulte sus jerarquías y su distribución espacial para ser relacionada aquí se recibe información de todas las demás secciones de manera estructurada. Estos datos son los que consulta el mecanismo de inferencias.

5. Mecanismo de inferencias: El mecanismo de inferencias está basado en reglas de producción, analiza la información y mediante inferencias realiza una descripción semántica de la imagen, tomando en cuenta varios factores tales como la prioridad de las categorías respecto al protagonismo dentro de la imagen; la prioridad en las categorías está considerada de la siguiente manera: personas, animales, transporte, alimentos, objetos y finalmente entorno. También se consideran las capacidades de las clases, es decir, algunas pueden realizar acciones y otras no.

En la Figura 1.1, se muestra el proceso del sistema implementado para las pruebas del modelo propuesto.

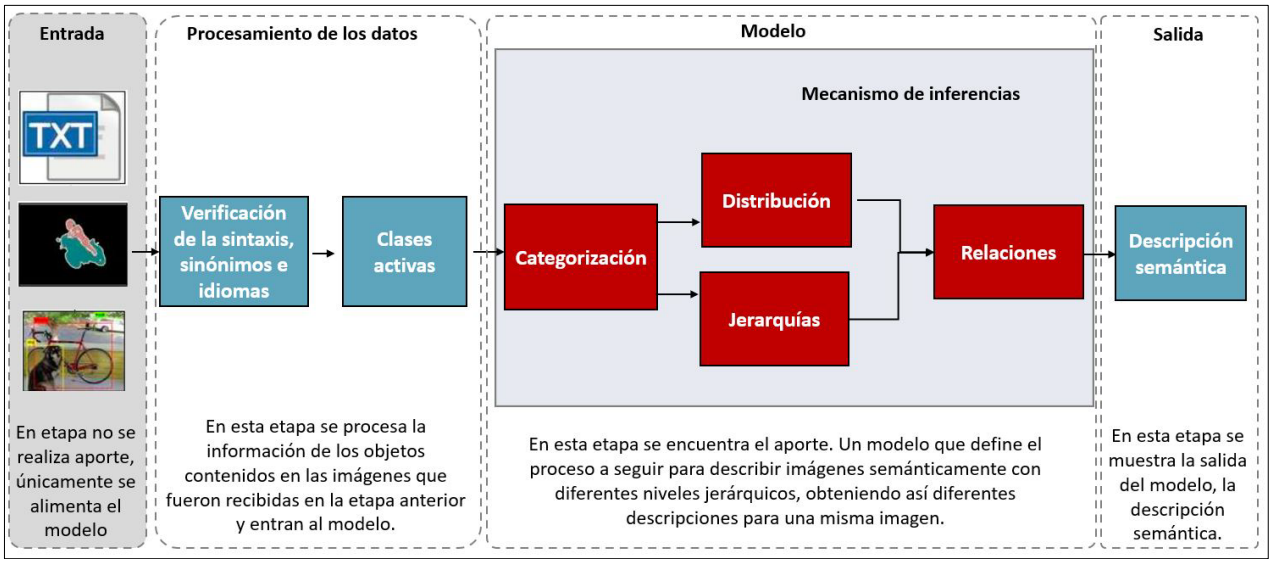


Figura 1.1 Sistema implementado para pruebas del modelo propuesto

1.6 Objetivo general

Proponer e implementar un modelo para la descripción semántica del contenido en imágenes.

1.6.1 Objetivos específicos

- Proponer e implementar un elemento dentro del modelo para la categorización de objetos.
- Proponer e implementar un elemento dentro del modelo basado en distribuciones espaciales de los elementos que conforman una imagen.
- Proponer e implementar un elemento dentro del modelo basado en jerarquías para que la información que conforma una imagen sea capaz de ser descrita con diferentes niveles de detalle.
- Proponer e implementar un elemento dentro del modelo de relaciones para que toda la información de una imagen pueda ser relacionada entre sí.
- Proponer e implementar un mecanismo de inferencias para que las descripciones puedan ser realizadas de manera natural usando verbos en lugar de posiciones, mediante el análisis de categorías, distribuciones, jerarquías y relaciones.

1.7 Alcances y limitaciones

El alcance de este proyecto de investigación fue lograr descripciones semánticas en imágenes, sin olvidar que el detalle de las mismas es definido por la cantidad de objetos presentes; si una imagen tiene pocos objetos o poca información las descripciones serán sencillas mientras que, para imágenes con mucha información serán detalladas. En este trabajo no se realizó el proceso de clasificación ni las partes que la componen (detección de objetos, segmentación semántica ni reconocimiento de objetos).

Adicionalmente, se creó un banco de conocimiento llamado RIDeCS [19], el cual consta actualmente de 126 imágenes en el que las personas pueden describir lo que observan en cada una de ellas, esta información sirve para medir la similitud entre las descripciones humanas y las proporcionadas por el modelo propuesto, así como con el trabajo de otros autores, esta información también puede ser usada para retroalimentar el mecanismo de inferencias.

Se añadió un sintetizador de voz al sistema donde se realizaron las pruebas del modelo con la finalidad de que este pudiera ser idóneo para personas con problemas visuales.

1.8 Justificación

Existe una gran problemática al momento de interpretar una imagen debido a que podría considerarse que dos imágenes son iguales por estar compuesta por los mismos elementos, sin embargo, aunque las imágenes tengan los mismos elementos pueden tener otro significado; por ejemplo, si se tuvieran dos imágenes con una persona cruzando una puerta, en dichas imágenes se tienen los mismos elementos (la persona, la puerta) pero son diferentes debido a que en una se podría observar un ingreso y en otra la salida del individuo. Se propone un modelo para la descripción de imágenes por contenido semántico, dado que no se identificó en la literatura un modelo que dicte el proceso a seguir en este tipo de trabajos considerando toda la escena de una imagen para ser descrita y no una lista de objetos que aparecen en la misma.

1.9 Organización de la tesis

A continuación, se describe la organización de los capítulos y su contenido en este documento de tesis. En el Capítulo 1 se expone la introducción, los antecedes, objetivos, problemática, así como su posible solución.

El Capítulo 2 corresponde al marco teórico, se abordarán algunas de las técnicas utilizadas en la representación del conocimiento y descripciones semánticas de imágenes para preparar al lector al Capítulo 3.

En el Capítulo 3 se encuentra el estado del arte donde se abordan trabajos relacionados, así como una breve comparativa entre lo ya reportado por otros autores y lo que se propone en esta investigación.

En el Capítulo 4 se describe el modelo propuesto para la descripción del contenido semántico en imágenes y se realiza el análisis de la solución propuesta.

El Capítulo 5 corresponde a la experimentación y resultados, se muestran los experimentos realizados con cada uno de los elementos del modelo, así como del modelo completo, comparativas con otros trabajos y con las respuestas de descripciones de imágenes dadas por humanos tomadas de RIDeCS.

CAPÍTULO 2

MARCO TEÓRICO

En este capítulo se abordarán algunas de las técnicas utilizadas en la representación del conocimiento y descripciones semánticas de imágenes para introducir al lector al tema de la descripción semántica de imágenes, donde se abordan términos como brecha semántica, redes semánticas, ontologías, lógica, así como su funcionamiento y usos dentro de la descripción semántica de imágenes.

2.1 Descripción semántica de imágenes

Para describir imágenes es necesaria una explicación de los elementos de estas, pero 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.

Para realizar una interpretación o recuperación de las imágenes mediante sistemas computacionales, es necesario realizar un proceso de descripción y uno de consultas. El proceso de descripción tiene como objetivo representar numérica o textualmente las características o propiedades de las imágenes que ingresan para formar el banco de conocimiento y estas pueden tener propiedades intrínsecas y extrínsecas [11].

Los sistemas de recuperación de imagen pueden representar de manera automática y con facilidad las características de bajo nivel o propiedades intrínsecas de las imágenes, ya que estos atributos son inherentes a la imagen. El problema se complica cuando se aborda la representación automática de las propiedades extrínsecas ya que, para describir el concepto de *frío, rostro, mascota, etc*. en términos de color no es tarea sencilla, y mucho más complejo sería expresar mediante estos atributos de bajo nivel un concepto subjetivo como el amor, la alegría, el dolor o la intranquilidad. En consecuencia, conforme más se adentra en el campo de las propiedades de la imagen, más aumenta la complejidad de la traducción o interpretación de la relación entre los atributos intrínsecos y objetos de una imagen [12].

Los seres humanos expresan con relativa facilidad los atributos de alto nivel incluyendo las propiedades semánticas, al observar una imagen es muy sencillo interpretar amor, tristeza, calor, etc. 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 cuando se realiza una búsqueda [11]. 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 o brecha semántica*; 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 [13].

En la evolución de la recuperación de imágenes pueden distinguirse tres grandes etapas: la primera etapa se basa en representaciones textuales de las características. La segunda se basa en rasgos visuales. En la tercera etapa se emplean simultáneamente el código visual y el código textual para representar, interpretar y recuperar imágenes. La primera etapa abarca desde los orígenes de la incorporación de imágenes como unidad de descripción documental a las colecciones digitales hasta la década de 1990. Durante la primera etapa, los sistemas de recuperación de imágenes aplican las mismas técnicas que se empleaban con los documentos textuales. En los sistemas de esta primera etapa que incorporan exclusivamente el código lingüístico, la tarea de recuperación se reduce esencialmente a la búsqueda de las palabras clave empleadas al describir la imagen, pero surgieron varios problemas como la dependencia de analistas humanos, la inconsistencia de la descripción entre analistas, el volumen de la documentación y la dificultad de descripción de las propiedades de bajo nivel mediante el código lingüístico aplicado a imágenes [14].

A fin de resolver estos inconvenientes, en la década de 1990 se desarrolló una segunda etapa denominada recuperación de imagen basada en contenido (*Content-Based Image Retrieval* o CBIR). La principal novedad de este nuevo enfoque consistió en la adopción de un código propiamente visual completamente distinto al lingüístico. La imagen se describe ahora mediante sus propiedades perceptuales, esto es, mediante las características psicofísicas percibidas por el ojo humano, principalmente el color, la textura, la forma y de igual forma que en la etapa anterior surgieron algunos problemas como las limitaciones formales. A raíz de los problemas planteados por la recuperación de imagen basada en contenido (CBIR), surge una tercera etapa que trata de superarlos, es la denominada recuperación de información visual basada en la semántica (*Semantic-Based Visual Information Retrieval* o SBVIR). Este enfoque se

caracteriza por la unión de la descripción del contenido visual y la descripción lingüística en una imagen [15].

La inteligencia artificial (IA) no sólo se ocupa de mecanismos generales relacionados con la búsqueda de soluciones en un espacio dado, o de cómo representar y utilizar el conocimiento de un determinado dominio de discurso, otro aspecto, es el que corresponde a los mecanismos y/o procesos inferenciales, que se consideran como el punto de partida de los llamados modelos de razonamiento. En cualquier dominio, la propagación del conocimiento por medio de programas de IA se efectúa siempre siguiendo un modelo de razonamiento bien definido, estos modelos de razonamiento forman parte del motor de inferencias si se habla de sistemas de producción, o de las estructuras de control del conocimiento o de cualquier otro tipo de sistemas, y contribuyen de manera decisiva a organizar correctamente la búsqueda de soluciones [16].

Pueden plantearse diversas maneras de representar el conocimiento para, posteriormente lograr descripciones semánticas mediante ontologías, lógica de primer orden, lógica difusa, etc. [17]. Casi todas las tareas que puede realizar un ser humano que se considera que requieren inteligencia, también, en ocasiones se basan en una gran cantidad de conocimiento. Por ejemplo, la mayoría de las declaraciones humanas son ambiguas, y esta ambigüedad es una característica esencial, no sólo del lenguaje, sino también de los procesos de clasificación, del establecimiento de taxonomías, jerarquías, y de los procesos de razonamiento en sí mismos [18].

2.2 Tipos de representación del conocimiento utilizados en imágenes

Para la recuperación de características basada en contenido son necesarios varios pasos tales como la segmentación de imagen, tomar en cuenta la característica de color, de textura, de forma, localización espacial y medida de similitud posteriormente será posible trabajar en la brecha entre niveles. J. Eakins, M. y Graham mencionan tres niveles de recuperación basada en contenido [3].

Representación del conocimiento Nivel 1: Recuperación por características primitivas como el color, la textura, la forma o la ubicación espacial de los elementos de la imagen, por ejemplo, consultar mediante la búsqueda, “encontrar imágenes como esta”.

Representación del conocimiento nivel 2: Recuperación de objetos de determinado tipo identificados por características derivadas, con algún grado de inferencia lógica. Por ejemplo, "encontrar una imagen de una flor”.

Representación del conocimiento nivel 3: Recuperación por atributos abstractos, involucrando una cantidad significativa de razonamiento de alto nivel sobre el propósito de los objetos o escenas representadas. Esto incluye la recuperación de eventos con nombre, de imágenes con significado emocional o religioso, etc. ejemplo de consulta, "encuentre imágenes de una multitud alegre".

La extracción de características de imagen de bajo nivel es la base de los sistemas de reconocimiento por contenido, las características de la imagen se pueden extraer de toda la imagen o de regiones, la mayoría de los sistemas actuales de reconocimiento por contenido están basados en regiones, la recuperación basada en características globales es comparativamente más simple. La representación de las imágenes a nivel de la región es similar al de percepción humana [22].

2.2.1 Brecha semántica

Los seres humanos expresan con relativa facilidad los atributos de alto nivel incluyendo las propiedades semánticas, al observar una imagen es muy sencillo interpretar amor, tristeza, calor etc. 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 cuando se realiza una búsqueda, 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. La brecha semántica entre las características visuales de bajo nivel y la semántica de alto nivel es un desafío bien conocido en la recuperación de información multimedia basada en contenido. Con la rápida popularización de los medios de comunicación social, que permite a los usuarios asignar etiquetas para describir imágenes y videos, es natural que la atención se centre en estos metadatos para superar la brecha semántica [23].

2.2.2 Redes semánticas

Las Redes Semánticas fueron tratadas en la época filosófica Aristotélica como una forma de ilustrar los métodos de Aristóteles para definir categorías. Los primeros esquemas de representación formalizados fueron dados por Quillian y Raphael en 1968, en 1971 Shapiro y Woddmansee dieron continuidad a estos trabajos [24, 25, 26].

Las redes semánticas han sido muy utilizadas en Inteligencia Artificial, los elementos que forman cualquier tipo de red semántica son: estructuras de datos en nodos, que representan conceptos, ligadas por arcos que representan las relaciones entre ellos y un conjunto de procedimientos de inferencia que operan sobre las estructuras de datos y se pueden distinguir tres categorías de redes semánticas:

1.- Redes IS-A, en las que los enlaces entre nodos están etiquetados, estas redes son el resultado de la observación y anotaciones del conocimiento humano y se basa en la adscripción de un subconjunto de elementos como parte de otro más general [27].

2.- Grafos conceptuales: en los que existen dos tipos de nodos, de conceptos y de relaciones. Los grafos conceptuales, propuestos por John Sowa en 1984, se diferencian de las redes IS-A en que los arcos no están etiquetados, y los nodos son de dos tipos: Nodos de concepto, que pueden representar tanto una entidad como un estado o proceso y nodos de relación, que indican cómo se relacionan los nodos de concepto. Por tanto, son los nodos de relación los que hacen la función de enlaces entre las entidades [28].

3.- Redes de marcos: en las que los puntos de unión de los enlaces son parte de la etiqueta del nodo, la representación basada en marcos constituye en gran medida la base del modelo orientado al objeto actual, ya que contiene casi todos los conceptos que éste presenta, aunque estas ideas fueron aplicadas en principio a los lenguajes de programación más que a lenguajes de representación y consulta [27].

2.2.3 Ontologías

El término ontología es utilizado en filosofía para hablar acerca de una teoría sobre la existencia y fue adoptado por la comunidad de inteligencia artificial para definir una categorización y las relaciones entre sus términos.

En el contexto de la ingeniería web, una ontología representa una taxonomía y un conjunto de reglas de inferencia. La taxonomía define las clases de objetos y de relaciones entre dichos objetos. Las clases, subclases y relaciones entre entidades son herramientas de gran potencia para usarlas en la Web Semántica [29].

2.2.4 Etiquetado semántico

Este tipo de etiquetado se usa como metadatos para agregar etiquetas a imágenes, son usadas en estructuras de contenidos HTML con la finalidad de enriquecer el proceso de búsqueda de la información. La base de la Web Semántica es añadir metadatos semánticos a las páginas, estos metadatos adicionales que describen el contenido y el significado de los datos se deben proporcionar de manera formal, para que así sea posible evaluarlas automáticamente por programas; para ello existen páginas como *schema.org,* donde se definen estos metadatos. Existen diferentes formas de codificar los metadatos, en concreto, Google admite tres: JSON-LD, micro formatos y RDFA [30].

2.2.5 Lógica

La representación del conocimiento y el razonamiento es un área que se aplica en la inteligencia artificial cuyo objetivo fundamental es representar el conocimiento de una

manera que facilite crear inferencias por medio de lógica a partir del conocimiento. La lógica analiza cómo pensar formalmente, cómo usar un sistema de símbolos para representar un dominio de discurso (aquello de lo que se puede hablar), junto con funciones que permitan inferir (realizar un razonamiento formal) sobre los objetos [31]. Generalmente, se usa algún tipo de lógica para proveer una semántica formal y se aplica a los símbolos del dominio del discurso, además de proveer cuantificadores, operadores modales, etc. esto, junto a una teoría de interpretación, da significado a las frases en la lógica [32].

2.2.5.1 Inferencias

Una interpretación para un sistema de lógica proposicional es una asignación de valores de verdad para cada variable proposicional, sumada a la asignación usual de significados para los operadores lógicos. A cada variable proposicional se le asigna uno de dos posibles valores de verdad: V (verdadero) o F (falso). Esto quiere decir que, si hay *n* variables proposicionales en el sistema, el número de interpretaciones distintas es de *2n* [33].

2.3 Bancos de datos

Existen diversos bancos de imágenes para trabajar con investigaciones relacionadas a semántica en imágenes.

• Caltech 101 [34].

• Event Dataset [35].

• Places the Scene Recognition Data Base [36].

• UCI Machine Learning Repository [37].

• Pascal VOC [38].

• Kodak Lossless True Color Image Suites [39].

• ADE20K Datase [40].

• Cityscapes dataset [41].

• Open Image [42].

2.4 Métricas

Las métricas permiten monitorizar un producto para medir la calidad y certeza, entre otros parámetros. Sin embargo, en el ámbito computacional su principal función es medir los resultados y proporcionar un grado o porcentaje de certeza con respecto a los resultados obtenidos. En el caso de problemas numéricos existen diversas métricas para medir el desempeño, certeza, precisión etc. En cambio, en el ámbito de las descripciones semánticas se vuelve una tarea complicada al intentar medir un trabajo cualitativo de manera cuantitativa además de que existen palabras totalmente diferentes pero que significan lo mismo, tales como los sinónimos.

Se revisaron diversas métricas enfocadas a texto, oraciones y lenguaje natural con la finalidad de encontrar la o las métricas adecuadas para medir los resultados de este trabajo. Inicialmente se revisaron métricas de similitud de texto, en específico medidas léxicas. En la Tabla 2.1, se muestran métricas de texto para léxico.

Tabla 2.1 Métricas basadas en léxico

Métrica	Descripción
Subsecuencia común más larga (LCS)	Es un algoritmo que calcula la similitud de dos cadenas basado en la longitud de la cadena continua de caracteres más larga que exista entre ambas cadenas [43].
Algoritmo de Levenshtein	El resultado de este algoritmo dinámico es el número mínimo de operaciones requeridas para transformar una palabra en otra. Se entiende por operaciones, una inserción, eliminación o la sustitución de un caracter [44].
Jaro-Winkler	Esta medida utiliza el número de caracteres que comparten ambas palabras, tomando en cuenta los caracteres que están en la misma posición y los que están transpuestos [45].

En la Tabla 2.2, se muestran métricas basadas en semántica.

Tabla 2.2 Métricas basadas en semántica

Métrica	Descripción
Hiperespacio análogo al lenguaje (HAL)	Crea un espacio semántico a través de las coocurrencias de las palabras en un corpus. Palabra por palabra se construye una matriz ***M*** donde ***Mij*** representa qué tanto coocurrió la palabra ***i*** con la palabra ***j*** en la colección de datos. El valor en cada posición de la matriz ***M*** se calcula de forma acumulativa. Así, mientras un par de palabras obtenga un valor alto en la matriz, es evidencia de que se refieren a lo mismo dentro del contexto [46].
Análisis semántico latente (LSA)	Es posible que sea la técnica basada en corpus más popular. Esta técnica asume que las palabras que semánticamente son similares coocurrirán en pequeños pedazos del texto. Para capturar eso, se construye una matriz donde los renglones representan los párrafos del texto y las columnas representan solo palabras. Posteriormente, como esta matriz regularmente es muy grande, se utiliza una técnica matemática llamada descomposición en valores singulares y es usada para reducir las dimensiones de la matriz tratando de mantener la similitud original. Al final las palabras se comparan de forma vectorial utilizando el coseno del ángulo que forman los vectores a comparar [47].
Información mutua - recuperación de la información	Esta técnica utiliza métodos avanzados de búsqueda de series para calcular la probabilidad de que una palabra sea similar a otra. Esta probabilidad se calcula midiendo qué tan a menudo concurren palabras cerca de otras en una colección de páginas web. El valor más alto es el resultado de este método [48].

En las métricas presentadas anteriormente se mide la similitud de cadenas, incluso el cuerpo en general de textos, lo cual podría ser de utilidad para medir la similitud entre descripciones de imágenes. La problemática con las oraciones es que no todas las personas utilizan palabras iguales para expresar lo mismo, por ello, aunque se diga la misma frase es posible que la concordancia y similitud calculada por este tipo de métricas sea muy baja, o incluso proporcione un veredicto con nula similitud. En la Tabla 2.3, se presentan métricas basadas en conocimiento.

Tabla 2.3 Métricas basadas en conocimiento

Métrica	Descripción
Resnik	Devuelve un valor continuo que denota qué tan similares son dos palabras, indicando que entre mayor sea el valor devuelto mayor es la similitud entre las palabras. Este cálculo está basado en el contenido de información en la red, es decir, a partir del nodo en común más cercano a las palabras que se están comparando. Esta función es en su totalidad dependiente de cómo se construyó la red de conocimiento por lo que cualquier cambio en la colección hace que el resultado de similitud entre dos palabras pueda cambiar de forma abrupta [49, 50].
Lin	Retorna un valor continuo denotando la similitud entre un par de palabras y el ancestro más cercano que sea común entre ellas [51, 52].
Jiang y Conrath	Retorna un valor continúo denotando la similitud entre un par de palabras. Este método está basado en las palabras que se van a comparar y el ancestro más cercano a las dos palabras [53, 54].
Leacock y Chodorow	De este método se obtiene un valor que denota el grado de similitud entre dos palabras que se están comparando. Esta medida está basada en el cálculo de la ruta más corta entre las palabras en la red, pero tomando en cuenta la profundidad máxima de las palabras. Para esta medida se toma en cuenta qué tan lejos se encuentra el ancestro común más cercano de las palabras comparadas [55, 56].
Wu y Palmer	Muy similar a la medida anterior, también se calcula la ruta más corta entre las palabras comparadas. La diferencia recae en el uso de uno de los ancestros de las palabras en la red. En la medida anterior se utiliza el ancestro en común más cercano; en esta medida se utilizará el ancestro más cercano a alguna de las dos palabras [57, 58].
Path length	Devuelve un valor continuo que denota qué tan similares son dos palabras, con base en la trayectoria más corta que conecta las palabras en el árbol obtenido únicamente a través de relaciones del tipo es-un. El resultado está acotado en el rango de cero a uno, donde cero significa nula similitud y uno significa máxima similitud [59, 60].

Finalmente se optó por utilizar la métrica *"similitud de oraciones basado en fragmentos y contenido"* la cual considera la semántica y las oraciones como un todo mediante un análisis de trozos y utiliza la ecuación 1

$$sim(X,Y) = \frac{2*\sum_i w_i(x,y)}{|X|+|Y|} \quad (1)$$

Donde X es el conjunto de trozos de la primera oración e Y es el conjunto de trozos de la segunda oración. Por tanto, los valores de w (x, y) son puntuaciones de similitud calculadas entre los fragmentos de X y los de Y. Los pesos se asignan con la métrica palabra a palabra TF-IDF. La métrica TF-IDF son las siglas en inglés de *Term frequency – Inverse document frequency* que traducido al español será "Frecuencia de términos – Frecuencia inversa del documento".

CAPÍTULO 3

ESTADO DEL ARTE

La descripción semántica de las imágenes es una tarea compleja debido a que no existe un modelo que proporcione el protocolo a seguir para realizar dicha tarea; además una misma imagen puede tener diversos niveles de descripción, algunos más extensos que otros y no existe algún algoritmo, regla o norma que dicte cuál es el nivel de detalle de descripción correcta. Para describir imágenes es necesaria una explicación de los elementos de estas. 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.

Para realizar una interpretación o recuperación de las imágenes mediante sistemas computacionales es necesario tener un elemento de descripción y uno de consultas.

El elemento de descripción tiene como misión representar numérica o textualmente las características o propiedades de las imágenes que ingresan para formar el banco de datos; estas pueden ser propiedades intrínsecas y extrínsecas.

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 se les denomina descripciones de bajo nivel [10].

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, fecha, formato o propiedades de carácter subjetivo denominadas semánticas extraídas a raíz de la contemplación de la imagen y dando una interpretación de ellas, las cuales suelen incorporarse en el apartado de descripciones o notas en los metadatos [11].

Algunos trabajos se enfocan en describir la semántica de las imágenes considerando los aspectos que podrían ser relevantes para los autores. En esta investigación se consideró importante la categorización, jerarquías, distribuciones y relaciones para con todo ello crear un mecanismo de inferencias para describir imágenes de manera semántica por medio de oraciones. En la Figura 3.1, se muestra el diagrama del proceso del modelo propuesto.

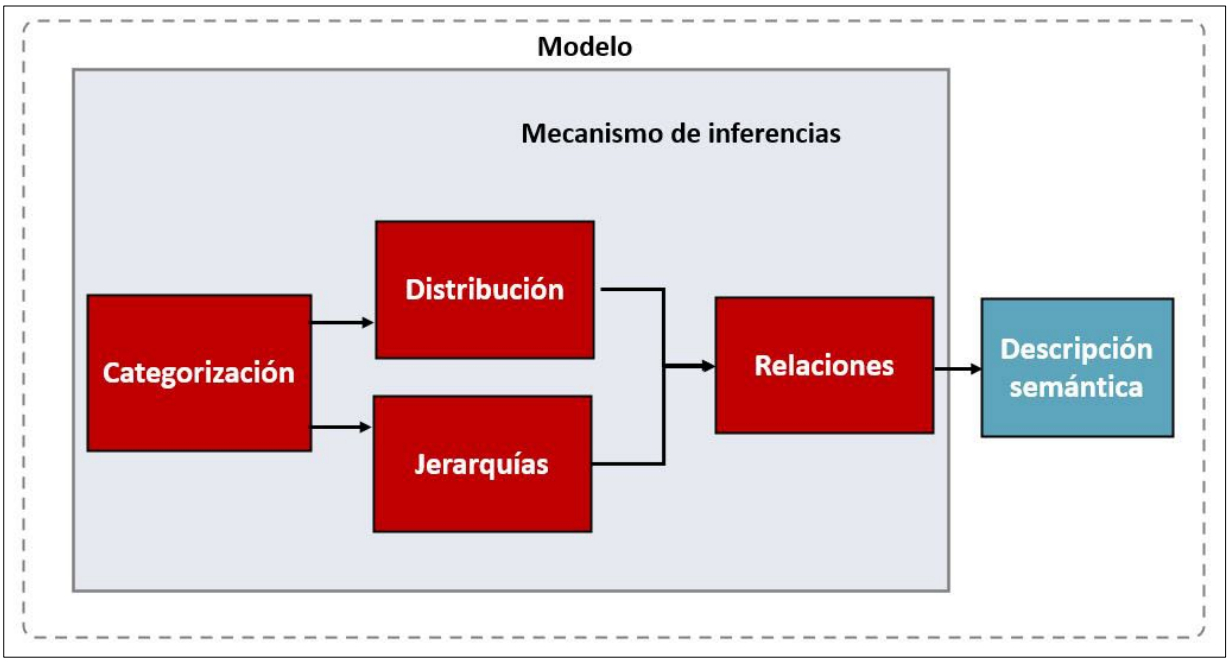


Figura 3.1 Diagrama del proceso del modelo propuesto para la descripción semántica de imágenes

Una vez comprobado por diversos investigadores a lo largo de los años que la representación de propiedades extrínsecas de alto nivel mediante atributos perceptuales de bajo nivel es una tarea extremadamente compleja, parece razonable considerar el texto como un medio eficaz de expresar significados abstractos o de alto nivel presentes en una imagen. En consecuencia, los sistemas de recuperación de imagen de la tercera generación emplean la anotación de imágenes mediante descriptores textuales para representar propiedades extrínsecas de carácter subjetivo o semántico.

La semántica estudia el significado de una palabra cuando su acción describe la función de un objeto y este es parte de un contexto visible y comprensible. Se puede enunciar que la semántica comprende un complejo sistema de variables que son usadas para determinar el sentido de una oración o un texto [10].

Dentro del reconocimiento de imágenes por contenido semántico, se tiene la problemática donde varias imágenes podrían tener contenido similar, hasta cierto nivel, es decir: dos o más imágenes pueden tener varios elementos iguales y eso le dará cierto grado de similitud.

3.1 Estado del arte y trabajos relacionados

Un trabajo pionero fue publicado por Chang en 1984, el autor presentó un enfoque de indexación y abstracción de imágenes para la recuperación de la base de datos pictórica. La base de datos pictórica consiste en objetos y relaciones de imágenes, para construir índices de imágenes [61].

Durante bastante tiempo se trabajó la recuperación semántica tomando en cuenta únicamente textura y color, sin embargo, tras varias investigaciones y trabajos se llegó a la conclusión que la distribución era un factor importante, dado que la ubicación espacial también es útil en la clasificación de la región; por ejemplo, cielo y mar podrían tener características de color y textura similares, pero sus ubicaciones espaciales son diferentes; el cielo generalmente aparece en la parte superior de una imagen, mientras que el mar en la parte inferior. Fue así como la distribución espacial comenzó a tomarse en cuenta en las descripciones semánticas de imágenes [62, 63].

La distribución espacial absoluta no es suficiente, la distribución relativa es más importante que la distribución espacial absoluta en la derivación de características semánticas y sus variantes son la estructura más común utilizada para representar relaciones direccionales entre objetos tales como izquierda, derecha, abajo, arriba. Sin embargo, tales relaciones direccionales por sí solas no son suficientes para representar el contenido semántico de las imágenes que ignoran las relaciones [64].

3.1.1 Descripción semántica de imágenes reduciendo la brecha semántica

En el trabajo Propuesta de recuperación de imágenes basada en contenido con semántica de alto nivel [65]. Identificaron cinco categorías principales de las técnicas más avanzadas para reducir la brecha semántica; donde toman en cuenta para describir las imágenes semánticamente, color, posición, tamaño y forma.

Pueden plantearse diversas maneras de representar el conocimiento para, posteriormente lograr descripciones semánticas, mediante ontologías, lógica de primer

orden, lógica difusa, etc. [17]. Casi todas las tareas que puede realizar un ser humano que se considera que requieren inteligencia, también, en ocasiones se basan en una gran cantidad de conocimiento. Por ejemplo, la mayoría de las declaraciones humanas son ambiguas, y esta ambigüedad es una característica esencial, no sólo del lenguaje, sino también de los procesos de clasificación, del establecimiento de taxonomías y jerarquías, y de los procesos de razonamiento en sí mismos [18].

El tema de las descripciones semánticas se ha abordado con diversos enfoques, sin embargo, en el trabajo siguiente utilizan cinco técnicas para reducir la *"brecha semántica":* ontologías de objetos para definir conceptos de alto nivel, métodos de aprendizaje automático para asociar funciones de bajo nivel con conceptos de consulta, comentarios relevantes para conocer la intención de los usuarios, generación de plantilla semántica para soportar la recuperación de imágenes de alto nivel y fusión de las evidencias del texto *Hyper Text Markup Language* (HTML) y el contenido visual de imágenes para la recuperación de imágenes en *World Wide Web* (WWW) [67].

En el siguiente trabajo [68] proponen un paradigma para evaluar la imagen considerando descripciones que utilizan el consenso humano, dicho paradigma consta de tres partes principales: un nuevo método basado en la recopilación de anotaciones humanas para medir el consenso, una nueva técnica automatizada que captura el consenso, y dos nuevos conjuntos de datos: PASCAL-50S y ABSTRACT-50S que contienen 50 oraciones que describen cada imagen. Argumentan que la técnica es simple y captura el juicio humano del consenso mejor que las técnicas existentes en las oraciones generadas por varias fuentes.

La mayoría de los prototipos y productos de bases de datos de imágenes se centran en búsquedas de similitud sobre características sintácticas de imágenes. DISIMA es un sistema de gestión de bases de datos de imágenes distribuidas, en desarrollo por investigadores de la Universidad de Alberta y tiene como objetivo proporcionar consultas tanto sintácticas como de las características semánticas de las imágenes. El contenido de una imagen se ve como un conjunto de objetos y regiones de interés. Los objetos destacados se organizan en dos niveles: objetos que se almacenan por sus

características sintácticas y lógicas, y objetos destacados que dan la semántica. DISIMA integra un lenguaje de consulta declarativa mediante *multimedia query language* (MOQL) y *visual multimedia query language* VisualMOQL [68].

3.1.2 Descripción semántica considerando jerarquías

Los datos de imagen y video se vuelven cada vez más populares para diversas aplicaciones. Se ha producido un volumen considerable de datos multimedia y es necesario desarrollar herramientas potentes y fáciles de usar para poder recuperar eficientemente dicha información. Un gran desafío es desarrollar sistemas que admitan múltiples búsquedas de información bajo diferentes criterios. Para hacer eso, en el trabajo *"Semantic structuration of image annotations: A data mining approach",* se realiza una estructuración de metadatos adjuntos a datos de imagen o video y proponen construir semiautomáticamente una jerarquía conceptual a partir de anotaciones manuales y descriptores visuales. El resultado es una jerarquía de conceptos que un usuario puede recorrer de forma no lineal, solo se usan datos de imagen y de video anotados con un enfoque "de abajo hacia arriba" que se desarrolla en dos pasos. Primero, de individuos anotando imágenes, usando el operador *Most Specific Concept* (MSC) en una descripción conceptual considerando no sólo objetos abstractos presentes en la imagen, sino también objetos de *dominio concreto* como el color, forma y textura. El siguiente paso compara los conceptos de descripción calculados entre sí según su similitud mediante inferencias. La idea principal de este trabajo es utilizar la noción de dominios concretos para mantener un enlace entre la jerarquía de conceptos y metadatos según la semántica de las imágenes y los descriptores visuales extraídos de la imagen (color, textura, forma) [69].

3.1.3 Descripción semántica mediante anotaciones

Otra de las maneras en que se están realizando las descripciones semánticas, es mediante un enfoque para aprender la semántica de las imágenes que permita anotar automáticamente palabras clave en una imagen, para posteriormente, recuperar imágenes basadas en consultas de texto [70]. En este trabajo, se modela la generación de descripción de imágenes mediante anotaciones, bajo el supuesto de que cada

imagen se divide en regiones, y que cada una está descrita por un vector de características de valor continuo. En el trabajo *"A Model for Learning the Semantics of Pictures",* utilizan un conjunto de imágenes de entrenamiento con anotaciones, para calcular un conjunto de características de imagen y palabras mediante un modelo probabilístico, que permite predecir la probabilidad de generar una palabra dadas las regiones de imagen.

Correlacionar la similitud semántica y visual de una imagen es una tarea difícil. Las posibilidades para la descripción de objetos en el mundo real son desafíos para el aprendizaje de técnicas. Algunos autores proponen las jerarquías de categorías, como una solución a la representación del conocimiento para ser descrito de manera semántica. En este trabajo, se propone un sistema para la asignación automática de semántica a imágenes usando una combinación adaptativa de múltiples características visuales. El "Algoritmo de selección de rama" selecciona sólo unos pocos subárboles para buscar desde esta base de datos de imágenes. Los algoritmos de poda reducen aún más este espacio de búsqueda. La correlación de las similitudes semánticas y visuales también se explora para comprender la superposición de la semántica. La eficacia de los algoritmos propuestos analizados en forma jerárquica y no jerárquica, las bases de datos muestran que el sistema es capaz de asignar semánticas generales y específicas a imágenes [71].

En el trabajo *"Semantic Image Retrieval via Active Grounding of Visual Situations",* proponen un enfoque para la recuperación de imágenes semánticas, donde, los resultados han demostrado la capacidad de utilizar la información obtenida para realizar búsquedas enfocadas y localizar objetos difíciles de detectar (por ejemplo, correas apenas visibles, perros pequeños, objetos parcialmente ocluidos) [72].

Por otra parte, se tiene el siguiente artículo que considera la importancia del conocimiento de expertos, es decir, para poder realizar una descripción automática les es importante tener una buena base de conocimiento, y se enfocan en la descripción semántica y la anotación de imágenes en el área cultural y de humanidades. Debido a la falta de un esquema de descripción jerárquica, en este trabajo se enfocaron en

descripciones del área cultural, específicamente de los frescos de Dunhuang mediante anotaciones basada en características visuales de bajo nivel y la anotación humana basada de expertos. Para resolver este problema, proponen un marco de descripción semántica para la descripción del contenido, basado en las necesidades de información y la teoría de recuperación. Este trabajo combina la descripción semántica, con información de sistemas de representación de conocimiento y se describe la relación entre los niveles semánticos en un marco descriptivo, se realizan pruebas preliminares específicamente en el campo del patrimonio cultural, utilizando imágenes digitales de los frescos de Dunhuang [73].

3.1.4 Descripciones del contenido en imágenes mediante relaciones espaciales

En el trabajo *"Visual Semantic Search: Retrieving Videos via Complex Textual Queries",* abordan el problema de la recuperación de videos, utilizando consultas de lenguaje natural para la obtención de imágenes por contenido semántico. Primero analizan las descripciones en un grafo, y se exploran las características de los datos, tales como: movimiento y las relaciones espaciales entre los objetos. Este trabajo se enfoca en la búsqueda semántica en el contexto de conducción autónoma, donde suelen existir escenas dinámicas y con muchos objetos. En este enfoque construye grafos semánticos para predecir sucesos en los videos de conducción automática, tales como colisiones [74].

3.1.5 Clasificación y descripción semántica

En el trabajo "*Prototypicality effects in global semantic description of objects",* utilizan un modelo de descripción basado en prototipos, el cual, codifica y almacena el significado semántico de un objeto, mientras describe sus características mediante clasificación con Redes Neuronales. El método utiliza prototipos semánticos para crear firmas descriptivas discriminativas que describen un objeto destacando sus características más distintivas dentro de una categoría, diferenciando objetos que podrían pertenecer a una misma categoría como razas de perros, marcas de coches, entre otros [75].

La descripción de habitaciones interiores también es un tema de investigación actual, en el trabajo *"A Robust Indoor Scene Recognition Method based on Sparse Representation",* realizan la interpretación semántica de imágenes para asignar tipos de habitaciones tales como: oficina, teatro, recámara, baño, etc. La asignación la realizan considerando los elementos que aparecen en una imagen y de esta manera se asigna la interpretación utilizando redes neuronales convolucionales (CNN) y la configuración de codificación dispersa mediante la creación de una nueva representación de escenas interiores para capturar características de objetos comunes de una escena determinada [76].

Entre los trabajos revisados se encontró el etiquetado de imágenes, en el trabajo "*Deep Visual-Semantic Alignments for Generating Image Descriptions";* se presenta un modelo que genera descripciones de imágenes y sus regiones. Este enfoque aprovecha los conjuntos de datos de imágenes y las descripciones de oraciones para aprender sobre las correspondencias intermodales entre el lenguaje y los datos visuales, el cual se basa en una combinación de redes neuronales convolucionales sobre regiones de imagen para clasificar los objetos, y redes neuronales recurrentes bidireccionales sobre oraciones y un objetivo estructurado que alinea las dos modalidades a través de una incrustación multimodal. De esta manera, se genera una alineación entre el objeto clasificado con la red neuronal convolucional dada su ubicación y una serie de oraciones, para generar la descripción [77].

3.1.6 Descripción mediante bolsa de palabras

Adicional se detectaron otras investigaciones del área que trabajan con descripción semántica a bajo y medio nivel. En "*Estado del arte la visión artificial, un nuevo aliado para el análisis de imágenes artísticas"* [78], se evalúa el rendimiento de los métodos de Bag-of-Visualterms (BOV) para la clasificación automática de imágenes digitales de la base de datos del artista Miquel Planas. Estas imágenes intervienen en la ideación y diseño de su producción escultórica. Constituye un interesante desafío dada la dificultad de la categorización de escenas cuando éstas difieren más por los contenidos semánticos que por los objetos que contienen. Se ha empleado un método de

reconocimiento basado en Kernels introducido por Lazebnik, Schmid y Ponce en 2006. Los resultados son prometedores, en promedio, la puntuación del rendimiento es del 70%. Los experimentos sugieren que la categorización automática de imágenes basada en métodos de visión artificial puede proporcionar principios objetivos en la catalogación de imágenes y que los resultados obtenidos pueden ser aplicados en diferentes campos de la creación artística. La metodología utilizada en esta investigación está basada en determinar las características locales que producen una representación de la imagen versátil y sólida capaz de mostrar el contenido global y local al mismo tiempo, y a la vez hacen robusta la descripción ante la oclusión parcial de objetos contenidos y la transformación de la propia imagen. En la Figura 3.2, se muestran las etapas del sistema.

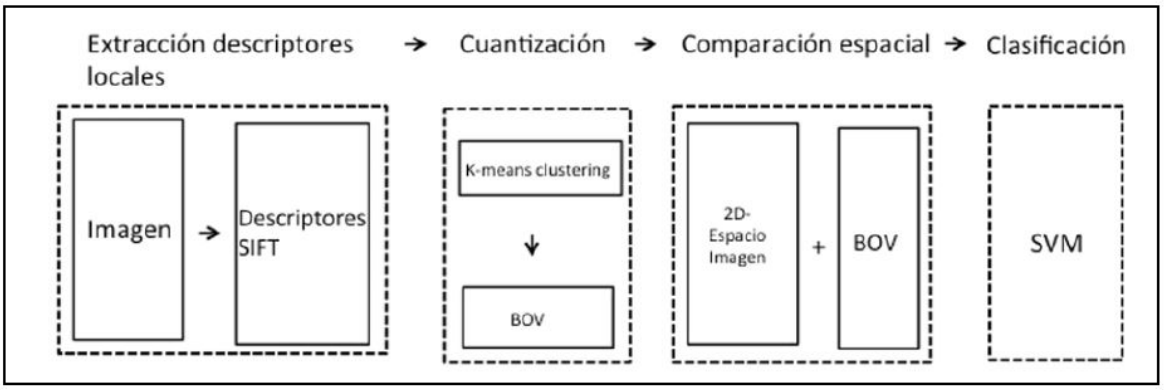


Figura 3.2 Etapas del sistema propuesto en [78]

En la etapa de extracción de descriptores locales se utiliza el descriptor de extracción de características locales (SIFT *Descriptors*). El descriptor *Scale Invariant Feature Transform* (SIFT) fue desarrollado como un algoritmo capaz de detectar puntos característicos (*keypoints*) estables en una imagen. Estos puntos son invariantes frente a diferentes transformaciones como traslación, escala, rotación, iluminación y transformaciones afines. Originalmente fue desarrollado para el reconocimiento de objetos en general y para realizar la alineación de imágenes. El algoritmo SIFT, se compone principalmente de cuatro etapas:

1. Detección de extremos en el Espacio de Escala: la primera etapa del algoritmo realiza una búsqueda sobre las diferentes escalas y dimensiones de la imagen identificando los candidatos a *keypoints*.

2. Localización de los *keypoints*: se seleccionan los *keypoints* a partir del conjunto de candidatos encontrados, aplicando una medida de estabilidad sobre todos ellos para descartar los que no sean adecuados. Un punto quedará seleccionado como *keypoint* sólo si es mayor que sus vecinos o menor que todos ellos.

3. Asignación de la orientación: se asignan una o más orientaciones a cada *keypoint* basándose en las direcciones locales presentes en la imagen gradiente. Se definió una región de 16x16 píxeles alrededor del punto donde se determina a orientación, y a cada uno de los píxeles se le calcula su gradiente. Éste viene determinado por su elemento e inclinación, ambos parámetros se calculan utilizando deferencias entre pixeles.

4. Descriptor del *keypoint*: la última etapa hace referencia a la representación de los *keypoints* como una medida de los gradientes locales de la imagen en las proximidades de dichos puntos clave y respecto de una determinada escala. Cada punto de interés corresponde a un vector de características compuesto por 128 elementos.

En la etapa de cuantización, se agrupan los descriptores en *M* clústeres los cuales definirán un vocabulario visual de *M visualterms* utilizando el algoritmo *k-means*. En la etapa de comparación espacial, una vez se tiene definido el vocabulario los descriptores de cada imagen se asignan al *visualterm* más cercano. Finalmente, para obtener la representación BOV (*Bag-of-visualterms*) de una imagen dada, se calcula la frecuencia de cada *visualterms* en la imagen. Para superar las limitaciones del enfoque bag-of-*visualterms* (BOV) se ha implementado la metodología PHOW (*Pyramid histogram of visual words*), la pirámide de coincidencias trabaja mediante la colocación de una secuencia de cuadrículas cada vez más finas sobre el espacio de características obteniendo una suma ponderada de la cantidad de coincidencias que ocurren en cada nivel de resolución.

Los histogramas espaciales se pueden utilizar como descriptores de imagen y alimentar con ellos un clasificador automático como por ejemplo los SVM (*Support Vector Machine*) el cual fue utilizado en este trabajo.

La arquitectura central y siluetas son las categorías con una mayor proporción de la clasificación correcta, 79% y 85% respectivamente. Posteriormente, se encontró las categorías piedra con textura y piedra irregular con el 61% y el 57% de clasificación correcta.

3.1.7 Comparación de imágenes por similitud

En el trabajo "*comparación perceptual basada en métricas de similitud de imágenes*" [79], la operación de comparar imágenes es un componente integral de muchas rutinas visuales. Las tareas visuales clave, como la estimación de la profundidad estereoscópica y la extracción del flujo de movimiento, dependen de poder establecer correspondencia entre las diferentes regiones de imágenes a lo largo del espacio o el tiempo. La correspondencia, a su vez, depende críticamente de la comparación de las regiones de imágenes y la evaluación de su similitud mutua.

Las comparaciones de imágenes desempeñan un papel aún más evidente en tareas como el reconocimiento de objetos. Para poder construir modelos precisos de estas habilidades visuales, se necesitó utilizar métricas de similitud de imagen formalmente especificadas que imiten a sus homólogos humanos. La necesidad de elegir métricas de similitud de imagen apropiadas también puede motivarse desde una perspectiva más pragmática. La creciente preponderancia de las imágenes digitales en diversos aspectos de la vida cotidiana requiere métodos automáticos para su manipulación, almacenamiento y uso.

Las métricas de similitud de imagen ("funciones de distancia" o, de manera más general en la teoría de la información, "medidas de distorsión") que cuantifican cuán bien una imagen coincide con otra, son fundamentales para muchas operaciones en imágenes digitales. Tres amplias clases de aplicaciones que se basan en las métricas de similitud de imagen apropiadamente elegidas son:

• la búsqueda de imágenes

• la compresión de imágenes

• la evaluación de la calidad de imagen

Por lo tanto, las métricas de similitud son importantes tanto para comprender la visión humana como para mejorar el procesamiento de imágenes en entornos aplicados. De hecho, los dos objetivos son complementarios. Las métricas computacionales se pueden mejorar al aproximar los juicios de similitud humana, mientras que los juicios de similitud humana se pueden comprender mejor comparándolos con métricas de similitud computacional formalmente caracterizadas.

Las métricas de similitud de imagen implican realizar alguna operación sobre las diferencias entre píxeles correspondientes en dos imágenes, luego sumando sobre estas diferencias modificadas. Estas se denominan colectivamente como la familia de métricas de similitud, la primera métrica a probar fue métrica de Minkowski, la métrica dos fue la distancia Manhattan.

En este trabajo, se tomaron en cuenta personas para medir la distorsión, ambos experimentos utilizaron el mismo diseño de elección forzada de dos alternativas, en el cual se instruyó a los participantes a elegir cuál de las dos imágenes (sonda) se parecía más a una imagen de referencia (objetivo). Los participantes indicaron su preferencia presionando un botón del teclado. Cada experimento consistió en 384 ensayos. Los participantes se sentaron aproximadamente a 75 cm del monitor de la pantalla en una habitación con poca iluminación ambiental. En la Figura 3.3, se muestra un ejemplo de reconstrucción de imágenes por VQ.

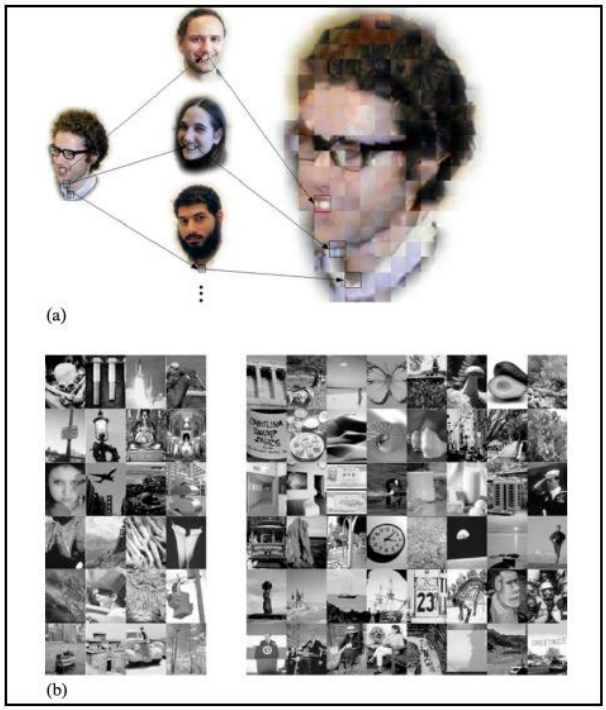


Figura 3.3 Ejemplo de reconstrucción de imágenes por VQ

En la Figura 3.3, la sub-imagen (A) combinan los fragmentos de las imágenes de la biblioteca usando una métrica de similitud de imagen con la imagen original. Estos fragmentos de las imágenes de la biblioteca luego reemplazan a los de la imagen de destino, creando una imagen reconstruida que está compuesta completamente por fragmentos de las imágenes de la biblioteca. En esta figura, la imagen a la izquierda es la imagen de destino, las imágenes centrales son el libro de códigos o biblioteca de imágenes, y la imagen a la derecha es una versión ampliada de la imagen reconstruida creada al colocar fragmentos de las imágenes de la biblioteca. Los fragmentos de la biblioteca se eligen para reemplazar los fragmentos objetivo y se aplica métrica de similitud de imagen. La mala calidad de la reconstrucción aquí se debe al muy reducido tamaño de las imágenes de biblioteca (5 imágenes). La sub-imagen (b) son las miniaturas de todas las imágenes utilizadas en el destino y en las imágenes de la biblioteca. Las 24 imágenes de la izquierda son las imágenes de destino, y las 48 imágenes de la derecha son las imágenes de la biblioteca.

Los participantes mostraron una pequeña preferencia, pero muy consistente, para las reconstrucciones. El 54% de todas las respuestas (promediadas entre tamaños de fragmentos y tamaños de bibliotecas) se realizaron para las reconstrucciones con la métrica de Minkowski. De los doce participantes, once eligieron la métrica de Minkowski en más del 50% de los ensayos y el sujeto restante eligió la métrica de distancia Manhattan en el 49% de los ensayos. Los participantes eligieron imágenes reconstruidas con la métrica de Minkowski significativamente más a menudo que las creadas con la métrica de distancia Manhattan. Por lo cual se concluye que usando la métrica Minkowski se podrían obtener mejores resultados en la categorización de escenas.

En el trabajo "*algoritmo de categorización de escena basado en codificación dispersa*" [80], se trabaja el reconocimiento de escena abstracta el cual tiene una amplia gama de aplicaciones, como el reconocimiento y la detección de objetos, la indexación y recuperación de imágenes basadas en el contenido, y la navegación inteligente de vehículos y robots. En particular, las imágenes de escenas naturales tienden a ser muy complejas y difíciles de analizar debido a los cambios de iluminación y rotación.

En este estudio, se investigó algunos modelos novedosos para aprender y reconocer escenas en la naturaleza mediante combinación de codificación dispersa limitada por la localidad (LCSP), agrupación de pirámides espaciales (SPP *Spatial Pyramid Pooling*) y máquinas de soporte vectorial (SVM) lineal en el modelo de extremo a extremo. Primero, los puntos interesantes para cada imagen en el conjunto de entrenamiento se caracterizan por una colección de rasgos locales, conocidas como palabras de código, que se obtienen usando un descriptor de SIFT denso. Cada palabra de código se representa como parte de un tema. Luego, se empleó el algoritmo LCSP para aprender la distribución de palabras clave de esas características locales a partir de las imágenes de entrenamiento.

Se utiliza un modelo de agrupamiento de pirámides espaciales modificado para codificar la distribución espacial de las características locales. Para la etapa final, se utiliza un SVM lineal para clasificar las características locales codificadas por SPP. Las

evaluaciones experimentales en varios puntos de referencia muestran bien la efectividad y robustez del método propuesto en comparación con varios descriptores visuales de vanguardia.

Dadas las imágenes de la escena, primero se caracterizó su información de gradiente local usando densos descriptores SIFT. Después de eso, se empleó una representación dispersa para la tarea de clasificación. Al ver que la representación escasa estándar no conserva la característica de localización de datos durante su proceso de codificación, se desarrolló un enfoque de aprendizaje del diccionario que emplea un término de regularización de localidad. Evaluaron su enfoque en tres conjuntos de datos: evento deportivo de 8 clases. El conjunto de datos de eventos deportivos de 8 clases [35] contiene 1,579 imágenes de ocho categorías de eventos deportivos, es decir, bádminton, petanca, croquet, polo, escalada en roca, remo, vela y tabla de nieve. El número de imágenes en cada categoría varía de 137 a 250 El conjunto de datos Caltech-101 [34] contiene 9,144 imágenes de 101 categorías de escenas naturales (incluidos animales, flores, rostros humanos, vehículos, etc.), con el número de imágenes por categoría de 31 a 800.

El conjunto de datos de la escena interior MIT 67 consiste en 15,620 imágenes de escenas interiores en 67 categorías, con al menos 100 imágenes por categoría. Se encontró que las técnicas aplicadas supera a otros enfoques con una precisión promedio de 84.40% para el conjunto eventos deportivos de 8 clases. 73.50% para el conjunto de datos Caltech-101 y 38.90%, respectivamente para el conjunto MIT escena interior.

En el trabajo *“Propuesta de recuperación de imágenes basada en contenido con semántica de alto nivel”* [81], identificaron cinco categorías principales de las técnicas más avanzadas para reducir la brecha semántica; donde toman en cuenta para describir las imágenes semánticamente, color, posición, tamaño y forma.

En la Figura 3.4, se muestra las características y sub-características tomadas en cuenta para realizar este trabajo.

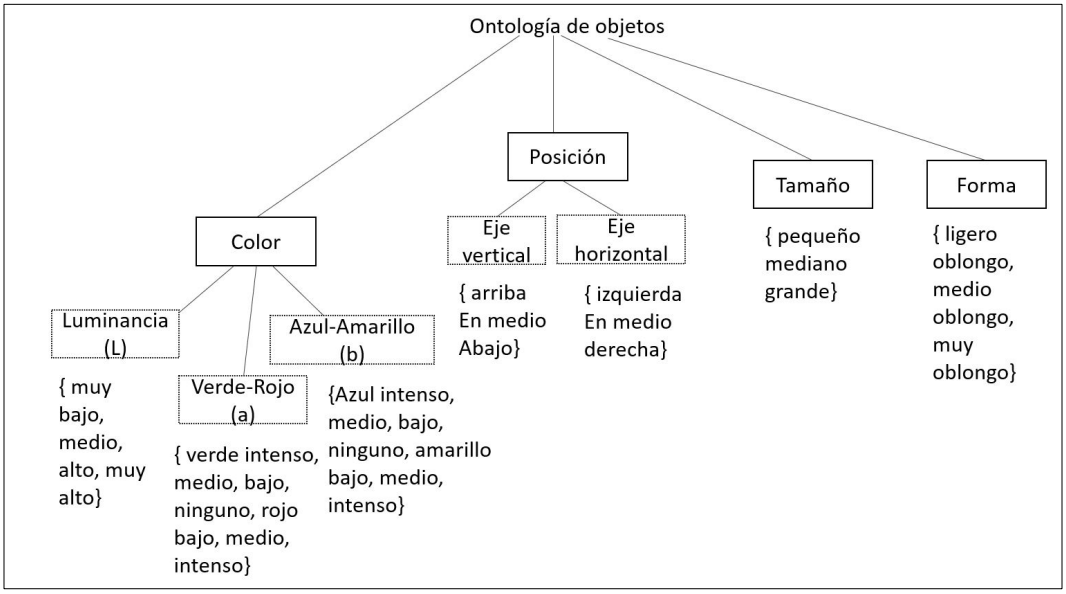


Figura 3.4 Características extraídas para la recuperación por contenido semántico en imágenes [81]

Como se muestra en la figura 3.4, se trata de un árbol de ontologías de objetos usado en este trabajo, los autores consideran que tomar en cuenta el color, la posición, el tamaño y la forma es una manera eficiente de realizar interpretación del contenido en imágenes desde una perspectiva semántica.

En un sistema típico de recuperación de imágenes basada en el contenido (CBIR), los resultados de la consulta son un conjunto de imágenes ordenadas por similitudes de características con respecto a la consulta; sin embargo, las imágenes con características muy similares pueden ser muy diferentes en términos de semántica. Esto se conoce como la brecha semántica.

3.1.8 Descripción mediante combinación de técnicas

Un nuevo esquema de recuperación de imágenes, basado en agrupamiento, se presenta en el trabajo "*Recuperación de imágenes basada en contenido mediante clustering"* [82], presenta CLUE (*CLUster-based rEtrieval*), que aborda el problema de brecha semántica basado en una hipótesis: donde las imágenes semánticamente similares tienden a agruparse en algún espacio, dadas sus características. CLUE intenta capturar conceptos, aprendiendo semánticas similares y recuperando grupos de

imágenes con características similares en lugar de un conjunto de imágenes ordenadas.

La agrupación en CLUE es dinámica, en particular, los conglomerados formados dependen de qué imágenes se recuperan en respuesta a la consulta. Por lo tanto, los clústeres le dan al algoritmo, así como a los usuarios, pistas semánticas relevantes. En la Figura 3.5, se muestra un diagrama de recuperación de imágenes incluyendo CLUE en un sistema.

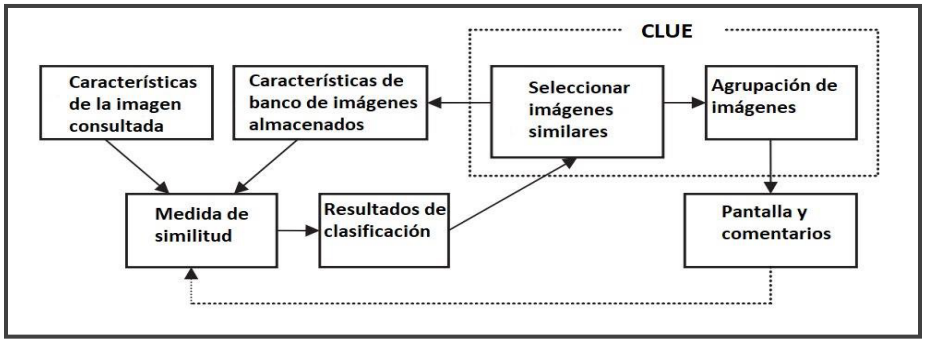


Figura 3.5 se muestra un diagrama de recuperación de imágenes incluyendo CLUE [82]

Al tomar agrupaciones de imágenes que comparten características, como se realiza con CLUE se pueden tener categorías más amplias. En la Figura 3.6, se muestra la categoría “Australia” tomada del banco de imágenes *Corel image data base* [66].

Figura 3.6 Categoría “Australia” del banco de imágenes Corel image data base

En la figura anterior se puede apreciar que la categoría “Australia” es muy amplia, desde imágenes de naturaleza, fauna, hasta ciudad. Las características entre ambas no tienen elementos en común, pero ambas son imágenes de paisajes o fauna de la

zona. Al agrupar por categorías que contienen semántica similar se puede conseguir este tipo de resultados; sin embargo, encontrar un vocabulario que represente la rica semántica de las imágenes no es una tarea fácil.

El trabajo *"Capturando la semántica de imágenes con descriptores de bajo nivel"* [83], realizó experimentos psicofísicos para obtener información sobre las categorías semánticas que guían la percepción humana de la similitud de la imagen. Al analizar los datos perceptuales, se analizan los datos perceptivos utilizando escalamiento multidimensional y agrupamiento jerárquico.

Con base en este análisis, se establecieron las categorías semánticas más importantes en la percepción de la similitud de la imagen. Luego se usaron estos datos para describir las características de bajo nivel que mejor describen cada categoría.

Al tomar en cuenta las jerarquías dentro de las categorías podría obtenerse una descripción más precisa con respecto a las interpretaciones que no las contemplan, por ejemplo, si se retomara el caso de la Figura 3.6, al aplicar jerarquías, la categoría "Australia" podría dividirse en: naturaleza, que a su vez podría dividirse en flora, fauna etc. de esta manera se tendrían descripciones detalladas de las categorías.

Un enfoque conjunto para entender la escena, es decir, la comprensión de la escena es la combinación de segmentación, reconocimiento de objetos y clasificación de escenas. Estas tareas son altamente codependientes; por un lado, las claves más importantes para la clasificación de escenas son los objetos contenidos en la escena. Por otro lado, los resultados de la clasificación de escenas ayudan a determinar los objetos que ocurren dentro de la escena, si se sabe que se está viendo una escena natural, es probable que haya césped y cielo, pero los sillones serían un objeto fuera del lugar dentro de la escena.

Finalmente, los resultados de segmentación se pueden mejorar mediante los resultados de reconocimiento de los objetos que conforman una imagen, ya que los modelos de color y forma típicos se pueden asociar con los objetos. En lugar de resolver todas las tareas por separado o secuencialmente, En el trabajo "*Optimización*

convexa para interpretar la escena [84], se adopta un enfoque holístico para la comprensión al resolver todas las tareas simultáneamente, similar a la forma en que los humanos razonan sobre el mundo que los rodea.

En la visión humana y la comprensión del mundo, especialmente las jerarquías de objetos son un concepto común, se pueden encontrar en un nivel de escena más grande que caracteriza los objetos que aparecen en un contexto específico. Los "automóviles" y las "señales viales" aparecen en "contextos callejeros", mientras que una "vaca" y una "oveja" suelen aparecer en "contextos naturales" y no en la "cocina" o al lado de una "computadora". Pero las jerarquías también se pueden encontrar en un nivel de pequeña escala que describe objetos individuales que se componen de diferentes partes, por ejemplo, una "bicicleta" consiste en "manubrios" y "neumáticos". En ambos contextos se caracterizan por relaciones semánticas específicas entre objetos o partes de objetos. Por lo tanto, la integración de la información jerárquica relacionada con el contexto en el nivel de escena es importante para obtener resultados altamente precisos [85]. En la Figura 3.7, se muestra una imagen con su *Ground Truth* y su árbol de descripción jerárquica.

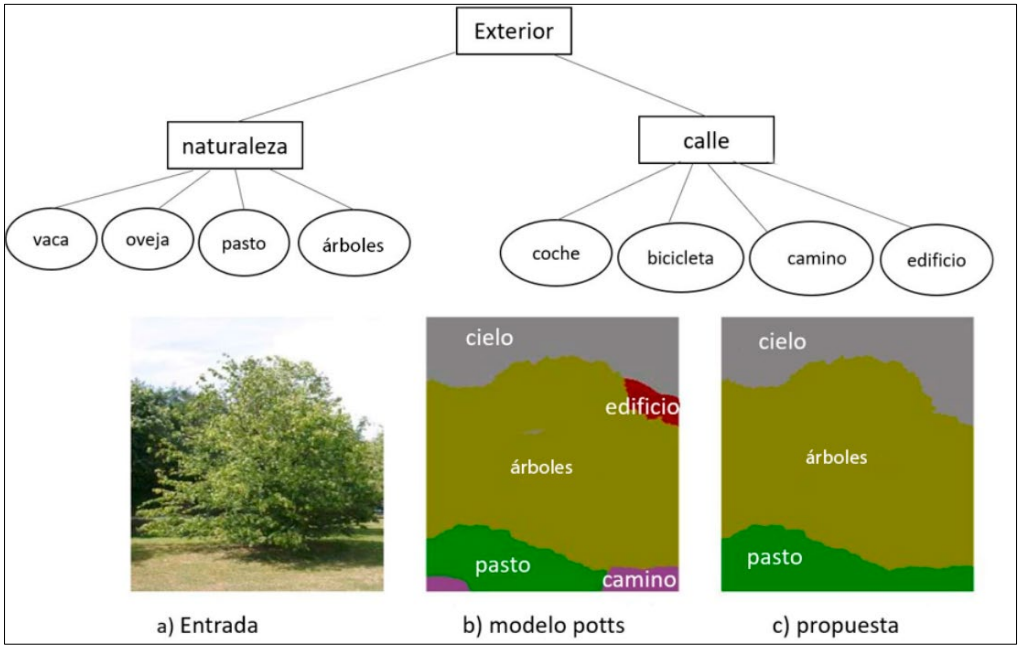


Figura 3.7 Imagen de naturaleza y su árbol de jerarquías

En general, esta tarea de comprensión de escenas se puede formular como un problema de etiqueta múltiple. Existen dos paradigmas populares para resolver tales problemas de optimización energética: enfoques discretos basados en *Markov Random Field* (MRF) y enfoques de optimización continua [86].

Un enfoque de comprensión holística de escenas se presenta en el trabajo "*Descripción de la escena como un todo: detección conjunta de objetos, clasificación de escenas y segmentación semántica"* [86], en el cual se interpreta una imagen como un todo, tomando en cuenta regiones, ubicación, clase y distribución espacial y relativa de los objetos, la presencia de una clase en la imagen, así como el tipo de escena.

Es muy común trabajar con los *Ground Truth* de imágenes para enfocarse en las tareas de interpretación semántica específicamente ya que es una tarea compleja. En este trabajo se propone un modelo jerárquico probabilístico para la detección y el reconocimiento de objetos en escenas naturales desordenadas. El modelo se basa en un conjunto de partes que describen el aspecto y la posición esperada de los objetos dada la lógica de donde podrían encontrarse coherentemente dentro de una imagen, además de proporcionar la ubicación de los objetos que conforman la imagen. En la Figura 3.8, se tiene una imagen dónde se pueden apreciar coches, calle; básicamente una escena de exteriores.

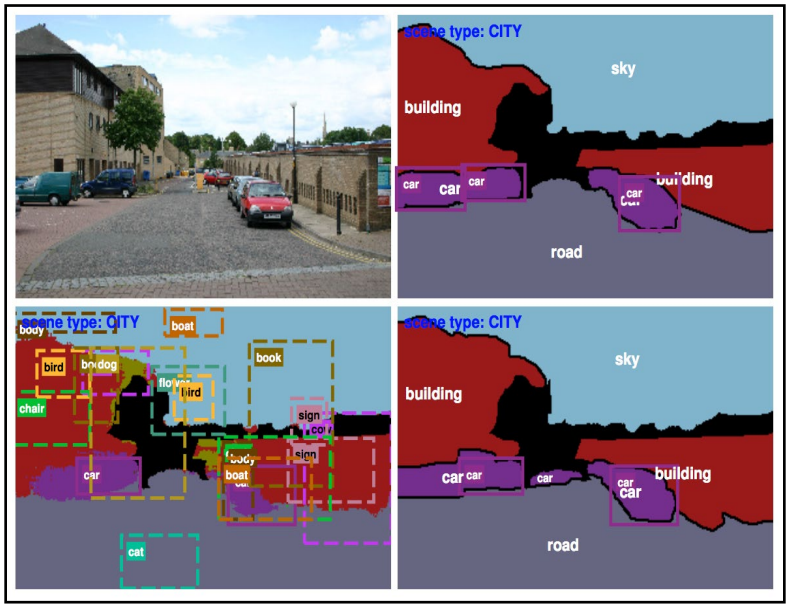


Figura 3.8 Imagen original, *Ground Truth*, salida de elementos individuales [87]

En algunas imágenes se tienen objetos que pertenecen a categorías distintas, lo cual es muy común; por ello es importante medir la similitud de las imágenes. En el trabajo "*Enfoques relacionales para la clasificación de objetos comunes y la medición de similitudes de escenas en entornos interiores"* [88], se habla sobre la creación de una estructura cualitativa de los objetos y su distribución espacial, en gran medida, definen una escena de entorno.

A pesar de tener esta gran variedad en categorías de objetos, formas, poses, texturas, etc. [89] El tipo de escena interior es definido por un subconjunto de objetos que se puede esperar que se vean en la escena dado que son elementos básicos que caracteriza el sitio y también se puede inferir un cierto posicionamiento relativo de los objetos; por ejemplo, en una escena de escritorio en oficina se podría esperar ver un monitor, un teclado y un *mouse* en determinada posición. En la Figura 3.9, se muestran algunas imágenes de interiores con diversos objetos que en conjunto definen los escenarios de interiores de una casa.

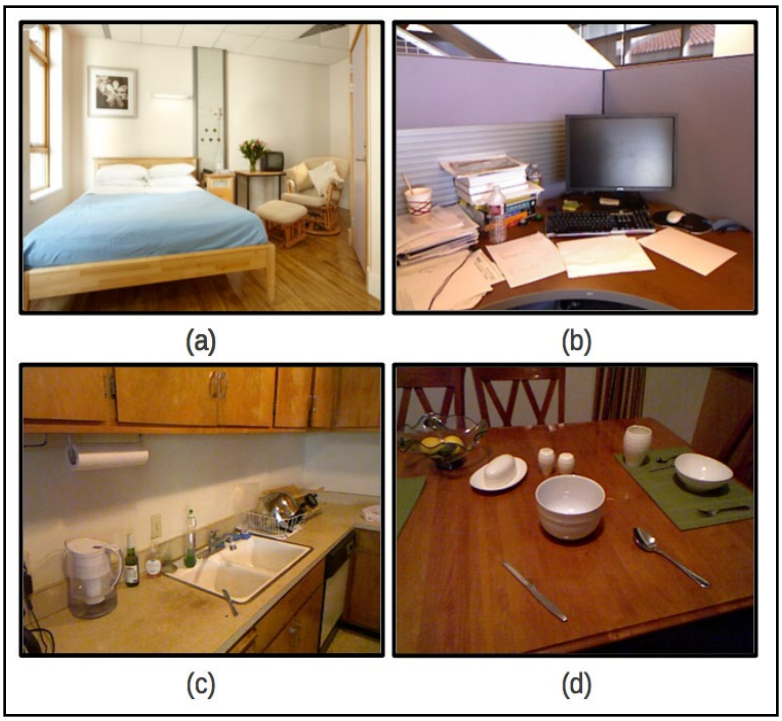


Figura 3.9 Ejemplos de escenas en interiores: a) recámara, b) oficina, c) cocina, d) comedor.

Este trabajo presenta un enfoque para la predicción conjunta de objetos y la medición de la similitud de escenas en ambientes interiores, en función de las relaciones espaciales relativas de los objetos.

Una relación espacial especifica cómo se ubica un objeto en el espacio en relación con un objeto de referencia. La información de la escena se representa como una nube de puntos 3D. Los puntos correspondientes a objetos separados se segmentan mediante una anotación manual donde se usan recuadros de delimitación 3D para definir los objetos. Las características geométricas 3D se calculan a partir de los objetos segmentados.

3.2 Discusión

Existen diferentes aspectos de la investigación en esta área que no están resueltas por completo desde la extracción de características de imagen de bajo nivel, la medición de la similitud entre imágenes, una descripción precisa de la misma y la relación entre objetos de una escena entre otras. Se han identificado 3 áreas con mayor aplicación de comparación e interpretación de imágenes por contenido semántico: observación terrestre, imágenes médicas, video vigilancia; en los trabajos relacionados se hace énfasis en que la recuperación por contenido semántico es apropiada para el trabajo con grandes cantidades de información. Tras la revisión de trabajos similares se han detectado áreas de oportunidad para este tema, en la etapa de la segmentación por regiones ha sido poco trabajada con respecto a la segmentación local adicionalmente, aunque la extracción de características de imagen de bajo nivel es la base de los sistemas de reconocimiento por contenido, la representación de las imágenes a nivel de la región es similar a la percepción humana y aún queda mucho trabajo para reducir la brecha entre los niveles en cómo recuperar imágenes por medio de contenido semántico, descripciones y etiquetado.

También mediante descriptores del estándar MPEG-7 queda trabajo por hacer, los descriptores de forma y color son los utilizados para el tema de contenido semántico en imágenes. Finalmente se detectó que un aspecto importante a considerar en la

interpretación de imágenes, un factor importante para una correcta interpretación es la distribución espacial de los objetos; existen escenarios que contienen los mismos objetos, pero su distribución espacial y relación entre los objetos es lo que las hace diferentes.

Como se puede apreciar en la revisión del problema y estado del arte, las descripciones de contenido semántico siguen siendo un área de investigación, algunos autores se enfocan en describir escenarios interiores, identificando los objetos que la componen para lograr identificar si se trata de un bar, restaurante, recámara, cocina, etc. Otros autores se enfocan en describir a un solo objeto, pero de manera semántica y finalmente otros trabajos se enfocan en describir las imágenes como un todo, tal como lo propuesto en este trabajo.

Un factor importante de mencionar es que, aunque las descripciones semánticas salen del rango cuantitativo, se sigue trabajando con técnicas para cuantificar la similitud semántica entre oraciones, lo que a su vez ha llevado a una amplia gama de enfoques para medir la similitud semántica.

Otro factor importante de observar es cómo los trabajos relacionados muestran algunos ejemplos donde se describen imágenes, sin embargo, no muestran resultados de los enunciados generados de descripciones, por lo que las comparativas se han limitado a pocos trabajos y fue necesario crear RIDeCS.

CAPÍTULO 4

MODELO PARA LA DESCRIPCIÓN DEL CONTENIDO SEMÁNTICO EN IMÁGENES

En la presente tesis, se propone un modelo para la descripción del contenido semántico de imágenes. El modelo propone estructurar el conocimiento para describir imágenes semánticamente mediante la categorización, jerarquización, distribución espacial y relaciones entre los objetos que aparecen en una imagen; cada uno de estos aspectos se considera como un elemento distinto, sin embargo, en conjunto proporcionan las descripciones semánticas. Estos elementos están formados por estructuras flexibles, donde la información es almacenada de manera que puede tenerse acceso jerárquicamente a ella, gracias a la estructuración del conocimiento. En la Figura 4.1, se muestra el proceso del sistema implementado para las pruebas del modelo.

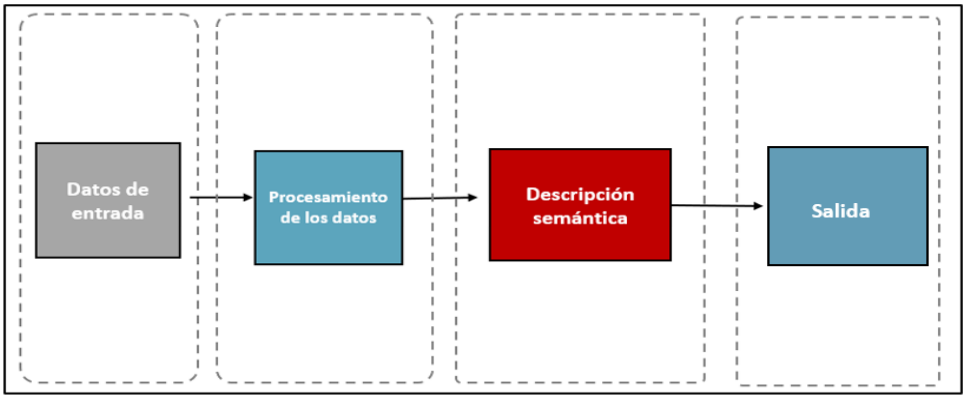


Figura 4.1 Diagrama del sistema utilizado para el modelo propuesto

En las siguientes subsecciones se describen las cuatro etapas que conforman al sistema: entrada, procesamiento de los datos, modelo de descripción semántica y salida.

4.1 Sistema de descripción semántica de imágenes

Se realizó un sistema computacional para realizar pruebas del modelo propuesto, en el punto 4.1.1, se habla de los datos de entrada, es decir, cómo se recibe la información. En el 4.1.2, se habla de cómo se procesan los datos de entrada. En el punto 4.1.3 de la descripción semántica, en ese punto se encuentra el aporte de esta investigación y finalmente en el 4.1.4 la salida.

4.1.1 Datos de entrada

Datos de entrada, como su nombre lo dice son los datos que entran al sistema, pueden ser texto en formato *JSON, txt, Ground truth y sistema de visión YOLOv3* [90], estos datos son los que alimentan al modelo. Cabe mencionar que el dato de entrada no está limitado a estas opciones, el usuario puede ingresar datos generados por su propio clasificador, sin embargo, para fines prácticos y demostrativos del modelo se utilizaron estas entradas. Entre mayor cantidad de información se reciba, mayor detalle tendrán las descripciones con respecto a los casos que cuenten con pocos datos. Los únicos datos que entran al sistema son las categorías y su posición. En la Figura 4.2, se muestra un caso genérico de entrada de los datos y cómo son estructurados.

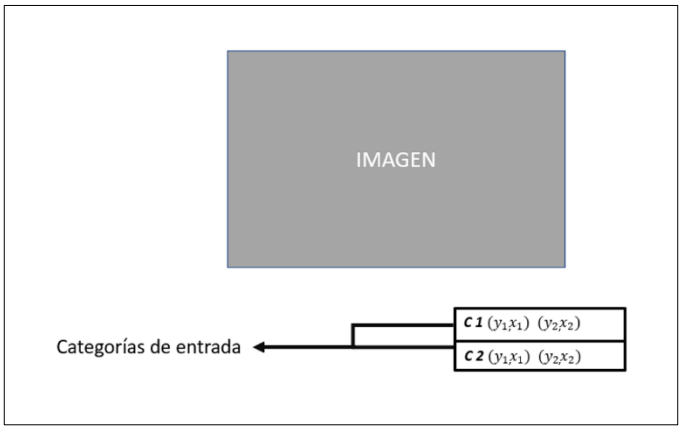


Figura 4.2 Entrada de datos universal

4.1.2 Procesamiento de los datos

Una vez que los datos han sido recibidos, es necesario comprobar que las categorías pertenecen al universo de datos con el que se está trabajando. En este caso se está trabajando con 702 categorías las cuales detecta YOLOv3, organizadas en seis super categorías: personas, animales, transporte, alimentos, objetos y entorno. Los datos que entran al modelo llevan un proceso de verificación sintáctica, se evalúa si la categoría recibida se encuentra entre las categorías en español con las que se está trabajando, si no hay coincidencia se revisa el diccionario de sinónimos e idiomas, por el momento los idiomas con los que se está trabajando son inglés y español tomados del diccionario de la lengua española (RAE) y para el caso de idioma inglés mediante diccionario de idioma. Si alguna categoría coincide se carga el árbol de dicha categoría.

4.1.3 Descripción semántica

En esta etapa es donde se encuentra el aporte de este trabajo mediante la creación de un modelo para describir imágenes, el cual consta de categorización, distribución, jerarquías, relaciones y un mecanismo de inferencias.

4.1.4 Datos de salida

En esta sección, la información es enviada a la interfaz para que pueda ser visualizada por el usuario o escuchada mediante el sintetizador de voz.

4.2 Modelo propuesto

El modelo está conformado por cinco elementos, *categorías, jerarquías, distribuciones, relaciones y mecanismo de inferencia*. En la Figura 4.3, se muestra en colores la sección que corresponde al aporte de esta investigación.

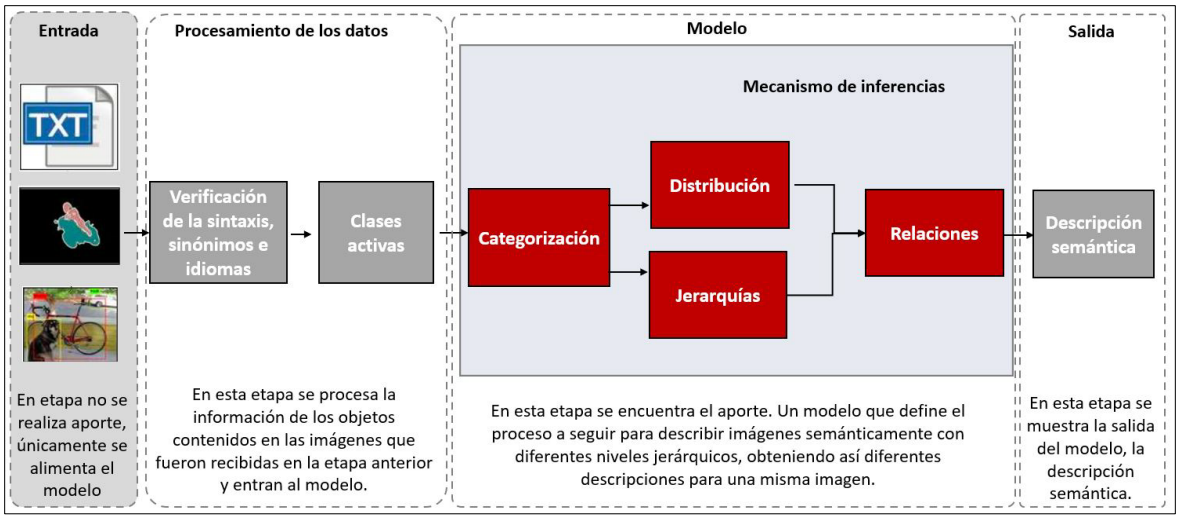


Figura 4.3 Modelo propuesto

En la Figura 4.4, se muestra el autómata finito determinista para modelar las transiciones de los estados para llegar al último paso la descripción semántica. Donde q0 entrada de datos, q1 obtener categorías, q2 obtener jerarquías, q3 obtener distribución, q4 obtener relaciones, q5 estado final.

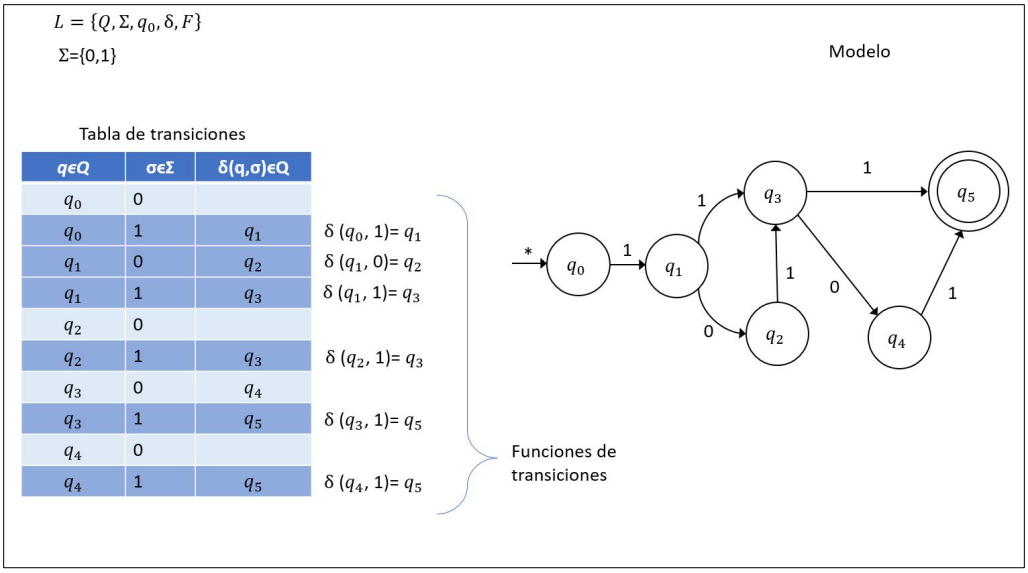


Figura 4.4 Autómata del modelo

4.2.1 Categorización

Una vez que los datos han sido recibidos, es necesario comprobar que las categorías pertenecen al universo de datos con el que se está trabajando. La información es almacenada en estructuras *hash* en la base de datos *Redis* [91], este tipo de estructuras almacenan datos que tienen campo-valor donde campo contiene el nombre de una categoría y valor hace referencia a la ruta del nodo dentro de los árboles jerárquicos.

Como se muestra en la Figura 4.5, la clave es el identificador único para acceder a la estructura de datos y a la estructura *hash* se le asocia un campo-valor (uno a uno), los anteriores parámetros corresponden a: campo a la categoría recibida y valor al nodo de la categoría.

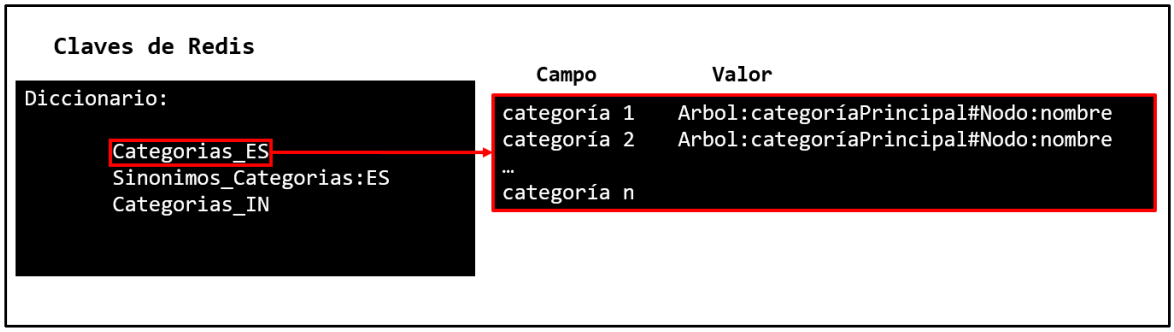


Figura 4.5 Procesamiento de los datos de entrada para un caso genérico

Estas etapas (categorización, distribución) están conformados por *Categorias_ES*, la cual es la lista de categorías válidas para el sistema en idioma español. *Sinónimos_CategoriasES* corresponde a la lista de sinónimos y *Categorias_IN* corresponde a la lista de categorías en inglés. Estas tres estructuras *hash* están compuestas por el campo *valor* que corresponde al nombre de la categoría y la clave del nombre del nodo en el que se encuentra dentro del árbol jerárquico.

En la Figura 4.6, se muestra el autómata finito determinista correspondiente al elemento categoría del modelo donde se observa cada uno de los estados por los que cruza la información para las categorías, donde, así como su tabla de transiciones.

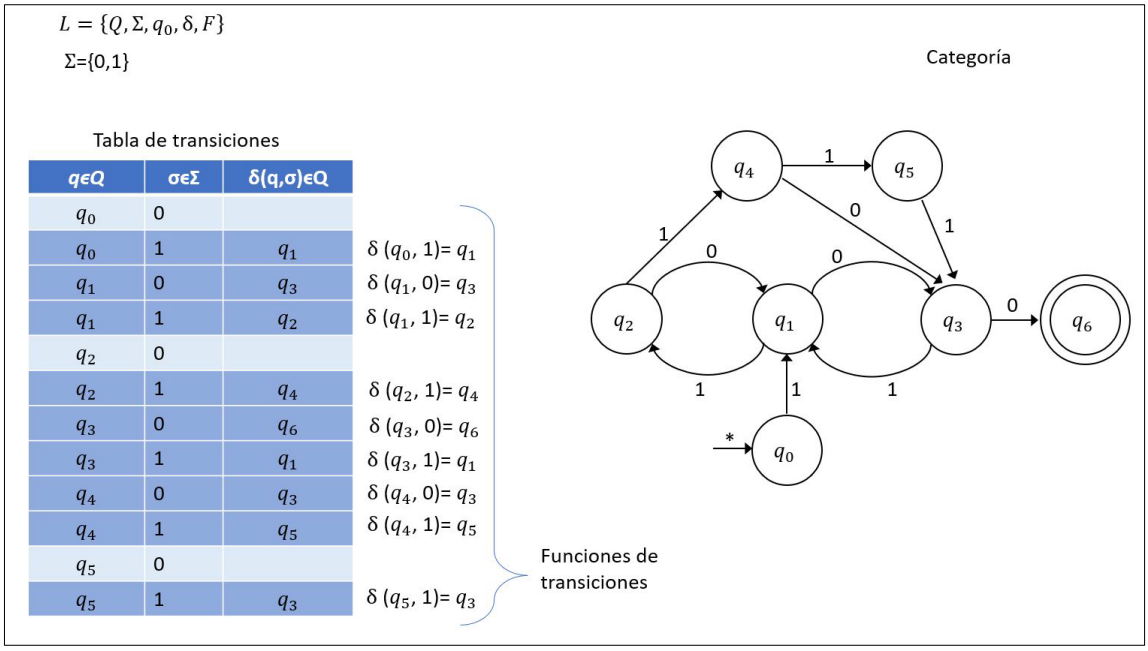


Figura 4.6 Autómata del elemento categoría del modelo

Donde q0 entrada de datos, q1 evalúa si la categoría existe, q2 activa la categoría, q3 verifica si hay más categorías a evaluar, q4 verifica si la categoría ha sido registrada con anterioridad, q5 pluraliza categoría y q6 estado final.

4.2.2 Jerarquías

En *jerarquías,* se encuentra el campo árbol, el cual contiene la información del nodo recibido, después *nodo:nombre* corresponde al nombre de la categoría entrante. Por ejemplo: *Arbol:animalesl#Nodo:vaca.* En la Figura 4.7, se muestra la estructuración de las jerarquías.

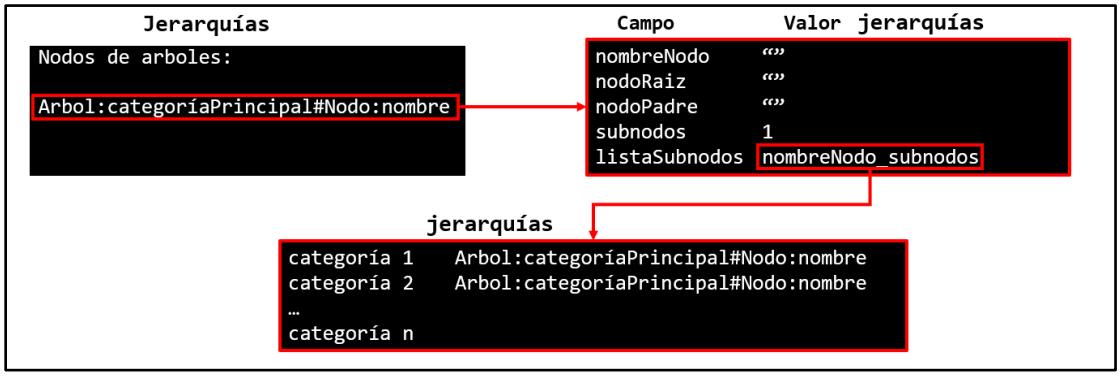


Figura 4.7 Estructuración de categorización y jerarquías en Redis

En la Figura 4.8, se muestra el autómata finito determinista correspondiente al elemento categoría del modelo donde se observa cada uno de los estados por los que cruza la información para las jerarquías, así como su tabla de transiciones.

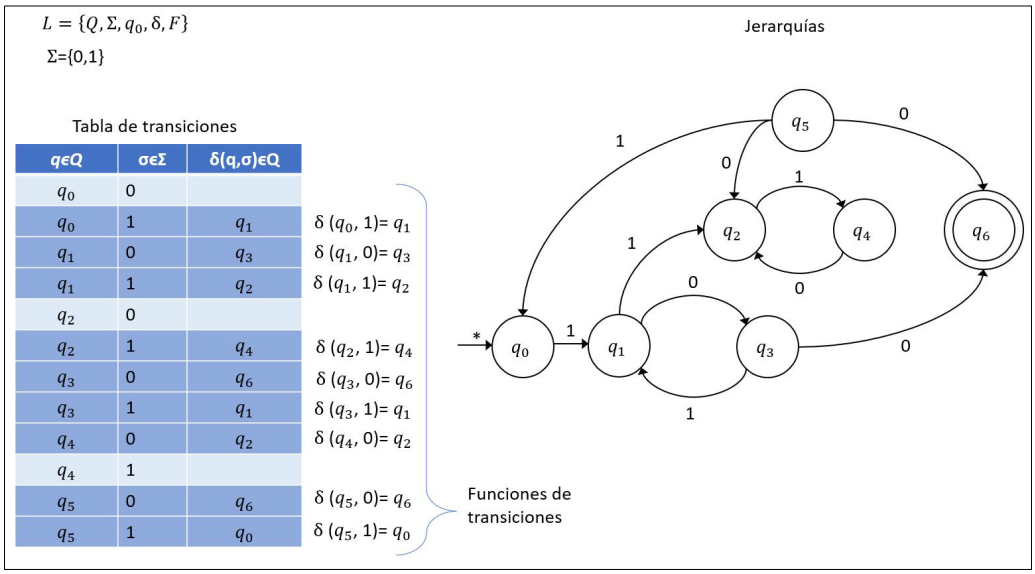


Figura 4.8 Autómata del elemento jerarquías del modelo

Donde q0 entrada de datos, q1 carga información de categoría, q2 verifica si es nodo raíz, q3 verifica si hay más categorías a evaluar, q4 recorre un nivel del árbol, q5 verifica si hay más categorías a evaluar y q6 estado final.

En la Figura 4.9, se muestra un árbol para ejemplificar el proceso de jerarquías, las jerarquías pueden tener un nivel de profundidad de varios niveles, hasta donde lo requiera la especificación de los datos.

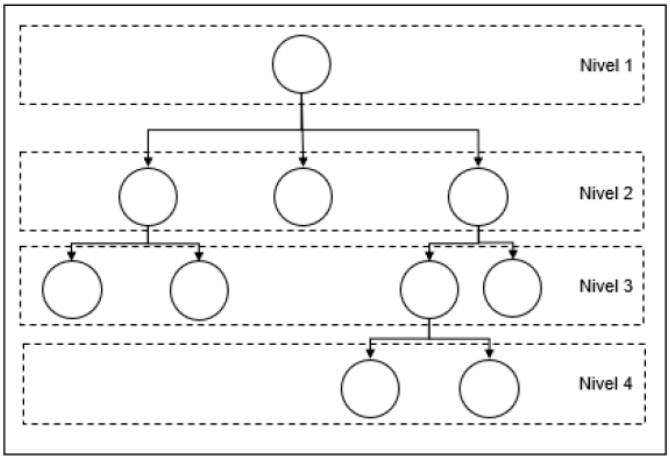


Figura 4.9 Árbol descriptivo de las jerarquías

Cada nodo tiene su estructura y está conformado como se muestra en la Tabla 4.1.

Tabla 4.1 Estructura de los nodos

Campo	Valor
nombreNodo	El nombre del nodo actual
nodoRaiz	El nodo raíz del árbol de la categoría
nodoPadre	El nodo inmediato antecesor al actual
Subnodos	1 (significa que este nodo tiene hijos) 0 (significa que este es un nodo final)
listaSubnodos	Estructura que contiene los nodos sucesores

4.2.3 Distribuciones

Una vez obtenidas las categorías y sus jerarquías se recuperan las distribuciones y son almacenadas y los parámetros que se reciben son los siguientes: (y_1, x_1) (y_2, x_2) los cuales corresponde a la distribución espacial de la categoría y en caso de no contener centroide se calcula utilizando las ecuaciones 2 y 3.

$$\bar{Y} = ((y_2 - y_1)/2) + y_1 \qquad (2)$$

$$\bar{X} = ((x_2 - x_1)/2) + x_1 \qquad (3)$$

En la Figura 4.10, se encuentra el autómata finito determinista correspondiente al elemento distribuciones del modelo donde se observa cada uno de los estados por los que cruza la información para las jerarquías, así como su tabla de transiciones.

En la Figura 4.11, se muestra un árbol para ejemplificar el proceso de distribuciones, en este caso, se agregan datos de ubicación para un nodo hijo, sin embargo, la ubicación puede ser de cualquier nodo del árbol.

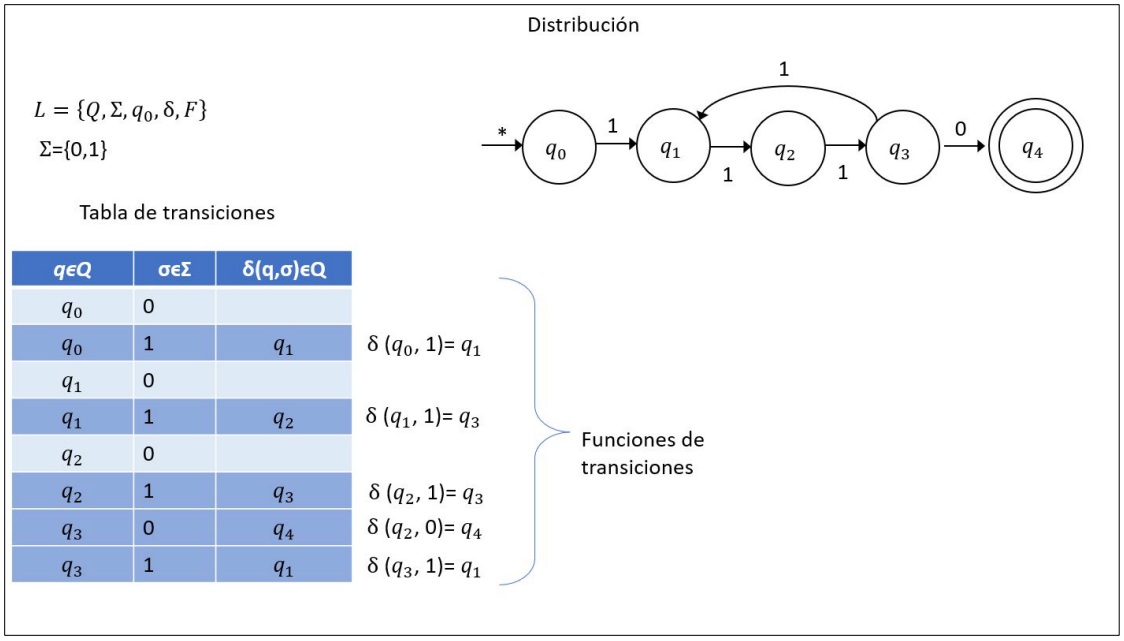


Figura 4.10 Autómata del elemento distribuciones del modelo

Donde q0 entrada de datos, q1 selecciona el objeto/categoría, q2 ubicación del objeto/categoría, q3 verifica si hay más categorías a evaluar, q4 estado final.

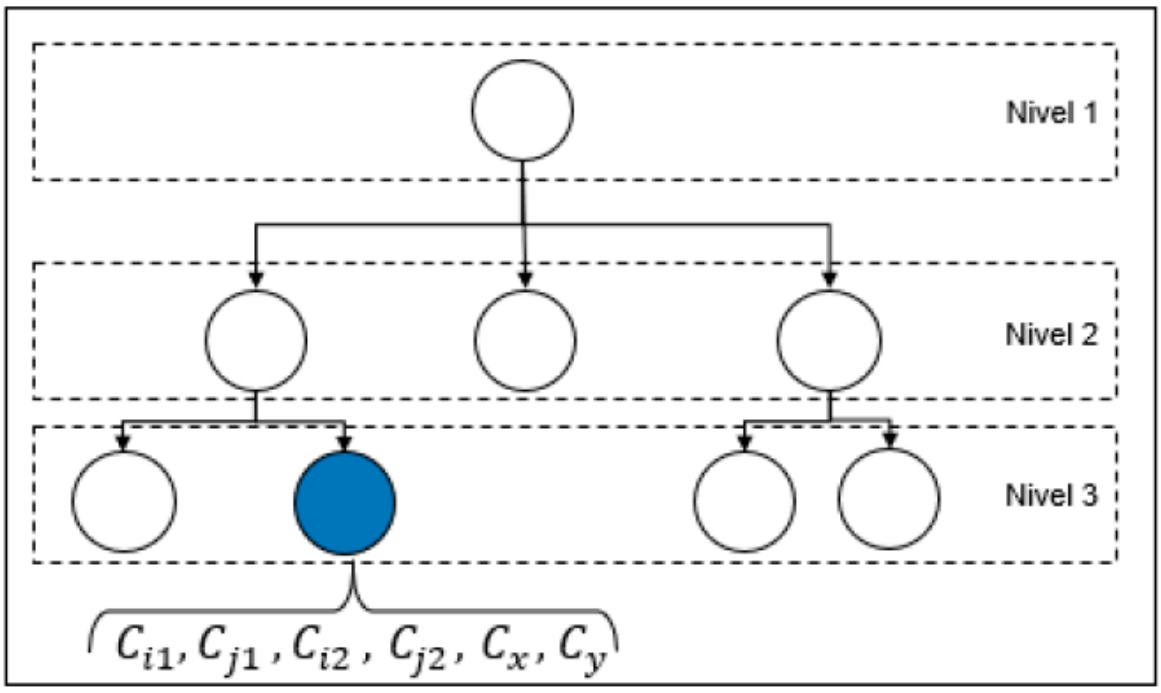


Figura 4.11 Árbol descriptivo de las distribuciones

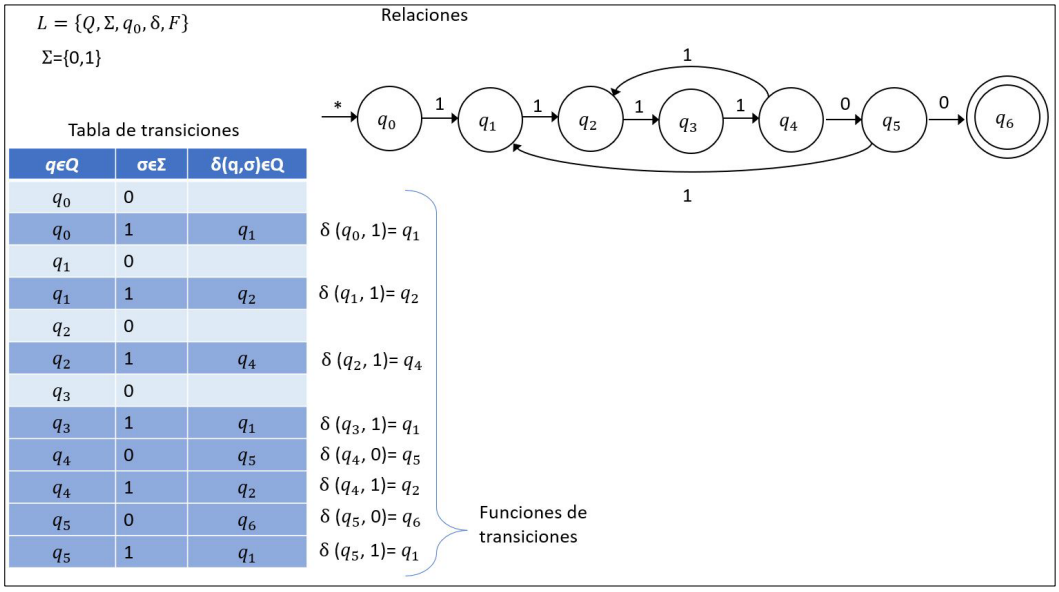
4.2.4 Relaciones

En la sección de *relaciones* se calcula la distancia Euclidiana con la ecuación 4 para relacionar los objetos que se traslapen, tengan cercanía y/o estén relacionados.

$$d(\mathbf{p},\mathbf{q}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2} \tag{4}$$

La etapa de relaciones es la encargada de que cada una de las categorías, jerarquías y la distribución espacial de cada uno de los objetos existentes en la imagen se relacionen, recibe información y la estructura para ser usada por el mecanismo de inferencias.

En la Figura 4.12, se aprecia el autómata finito determinista correspondiente al elemento distribuciones del modelo donde se puede ver cada uno de los estados por los que cruza la información para las relaciones, así como su tabla de transiciones.



qϵQ	σϵΣ	δ(q,σ)ϵQ	
q_0	0		
q_0	1	q_1	$\delta(q_0, 1) = q_1$
q_1	0		
q_1	1	q_2	$\delta(q_1, 1) = q_2$
q_2	0		
q_2	1	q_4	$\delta(q_2, 1) = q_4$
q_3	0		
q_3	1	q_1	$\delta(q_3, 1) = q_1$
q_4	0	q_5	$\delta(q_4, 0) = q_5$
q_4	1	q_2	$\delta(q_4, 1) = q_2$
q_5	0	q_6	$\delta(q_5, 0) = q_6$
q_5	1	q_1	$\delta(q_5, 1) = q_1$

Figura 4.12 Autómata del elemento relaciones del modelo

Donde q0 entrada de datos, q1 carga una categoría (1), q2 carga otra categoría (2), q3 se obtiene datos de distancia, q4 verifica si hay más categorías a evaluar (2), q5 verifica si hay más categorías a evaluar (1) y q6 estado final. En la Figura 4.13, se muestra un árbol para ejemplificar el proceso de relaciones, en este caso, se

relacionan dos nodos hoja, sin embargo, la relación se puede dar en cualquier nivel jerárquico.

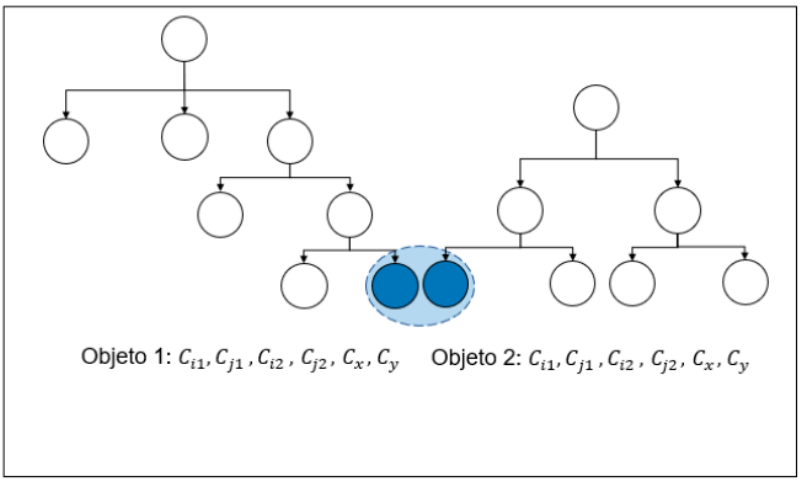


Figura 4.13 Árbol descriptivo de las relaciones

Una vez que todas las categorías han sido activadas, que las jerarquías se han desplegado en el máximo posible, así como sus distribuciones dentro de la imagen se relacionan entre ellos, para finalmente con toda la información estructurada y relacionada entre ellos, puedan realizarse descripciones semánticas, esto mediante un mecanismo de inferencias, además en esa última etapa se realizan sustituciones a los enunciados para dar una descripción más natural, por ejemplo, en lugar de describir que hay una persona sobre un caballo se puede inferir que se está montando o paseando.

En la siguiente sección, se describe el mecanismo de inferencias. Dentro del mecanismo de inferencias se utilizan reglas para mejorar las descripciones, por ejemplo, en lugar de describir “hay una persona sobre un caballo” dadas las características de las oraciones y los objetos involucrados, se puede decir, “hay una persona montando a caballo o persona paseando a caballo” y así para diversas frases y que contengan ciertos objetos donde pueda aplicarse este tipo de descripción, logrando que la descripción parezca más natural. También en esta sección la información es enviada a la interfaz para que pueda ser visualizada por el usuario o escuchada mediante el sintetizador de voz, el resultado es una descripción semántica obtenida a partir de los datos recibidos por el modelo.

4.2.5 Mecanismo de inferencia

En esta sección, se muestra el *mecanismo de inferencia* basado en reglas de producción, el cual permite realizar inferencias sobre los datos de las imágenes. En la Tabla 4.2, se muestran 55 de 224 reglas proposicional utilizadas para crear las descripciones semánticas. Cabe mencionar que estas reglas de producción son genéricas. Finalmente, la información puede ser visualizada o escuchada mediante el sintetizador de voz.

Tabla 4.2 Reglas para la descripción semántica mediante inferencias

Regla	Descripción
$\forall x\ \exists a$	Para cada categoría existe una *a*
$a = arriba$	*a* toma el valor constante *arriba*
$\forall x\ \exists b$	Para cada categoría existe una *b*
$b = abajo$	*b* toma el valor constante abajo
$\forall x\ \exists c$	Para cada categoría existe una *c*
$c = derecha$	*c* toma el valor constante derecha
$\forall x\ \exists d$	Para cada categoría existe una *d*
$d = izquierda$	*d* toma el valor constante izquierda
$\forall x\ \exists e$	Para cada categoría existe una *e*
$e = centroide$	*e* toma el valor constante centroide
$\forall e\ \exists\ i$	Para cada centroide existe la coordenada *i*
$i = valor\ de\ i\ en\ la\ imagen$	Valor de coordenada *i* de centroide
$\forall e\ \exists\ j$	Para cada centroide existe la coordenada *j*
$j = valor\ de\ j\ en\ la\ imagen$	Valor de coordenada *j* de centroide
$A = arriba\ de\ x$	Evaluación de relación de pares de objetos
$B = abajo\ de\ x$	Evaluación de relación de pares de objetos
$C = a\ la\ derecha\ de\ x$	Evaluación de relación de pares de objetos
$D = a\ la\ izquierda\ de\ x$	Evaluación de relación de pares de objetos
$\forall entorno\ \exists a, b, c, d, e$	Para cada entorno existe *a, b, c, d, e*
$\neg\ entorno\ \rightarrow\ r$	Si no existe entorno entonces es interior
$entorno_1 \rightarrow t$	Si existe entorno entonces es exterior
$entorno_1 \wedge entorno_2 \rightarrow t$	Si hay dos entornos entonces es exterior

Tabla 4.2 Reglas para la descripción semántica mediante inferencias *(continuación)*

Regla	Descripción
$entorno_1 \wedge entorno_2 \wedge entorno_3 \rightarrow t$	Si hay tres entornos entonces es exterior
$entorno_1 \wedge entorno_2 \vee entorno_3 \wedge entorno_4 \rightarrow t$	Si existe dos de cuatro entornos es exterior
$entorno_1 \vee entorno_2 \rightarrow t$	Si hay uno de los dos entornos es exterior
$entorno_1 \vee entorno_2 \vee entorno_3 \rightarrow t$	Si hay uno de los tres entornos es exterior
$p_1 \rightarrow q$	Premisa para una categoría
$p_1 \wedge p_2 \rightarrow q$	Evaluación para dos categorías
$p_1 \wedge p_2 \wedge p_3 \rightarrow q$	Evaluación para tres categorías
$p_1 \wedge p_2 \rightarrow q$	Evaluación para dos categorías y la ausencia de entorno
$p_1 \wedge p_2 \wedge p_3 \rightarrow q \wedge r$	Evaluación para tres categorías y la ausencia de entorno
$p_1 \wedge p_2 \wedge p_3 \wedge entorno \rightarrow q$	Evaluación para dos categorías y un entorno
$p_1 \wedge p_2 \vee p_3 \wedge\ p_3 \rightarrow q$	Evaluación para cuatro categorías
$p_1 \wedge p_2 \vee p_3 \wedge\ p_3 \rightarrow\ \ q \wedge r$	Evaluación para cuatro categorías y ausencia de entorno
$p_1 \wedge p_2 \wedge (j^{p_1} < j^{p_2}) \rightarrow p_1\ A(p_2)$	Si existen dos categorías y la coordenada *j* de la primera categoría es menor entonces la primera categoría está arriba de la segunda
$p_1 \wedge p_2 \wedge (j^{p_1} < j^{p_2}) \rightarrow p_1\ A(p_2) \wedge r$	Si existen dos categorías y la coordenada *j* de la primera categoría es menor entonces la primera categoría está arriba de la segunda y ausencia de entorno
$p_1 \wedge p_2 \wedge (j^{p_1} < j^{p_2}) \wedge entorno \rightarrow p_1\ A(p_2) \wedge entorno$	Si existen dos categorías y la coordenada *j* de la primera categoría es menor entonces la primera categoría está arriba de la segunda en un entorno
$p_1 \wedge p_2 \wedge (j^{p_1} > j^{p_2}) \rightarrow p_1\ B(p_2)$	Si existen dos categorías y la coordenada *j* de la primera categoría es mayor entonces la primera categoría está abajo de la segunda
$p_1 \wedge p_2 \wedge (j^{p_1} > j^{p_2}) \rightarrow p_1\ B(p_2) \wedge r$	Si existen dos categorías y la coordenada *j* de la primera categoría es mayor entonces la primera categoría está abajo de la segunda y ausencia de entorno
$p_1 \wedge p_2 \wedge (j^{p_1} > j^{p_2}) \wedge entorno \rightarrow p_1\ B(p_2) \wedge entorno$	Si existen dos categorías y la coordenada *j* de la primera categoría es mayor entonces la primera categoría está abajo de la segunda en un entorno

Tabla 4.2 Reglas para la descripción semántica mediante inferencias *(continuación)*

Regla	Descripción
$p_1 \wedge p_2 \wedge (i^{p_1} < i^{p_2}) \rightarrow p_1\ C(p_2) \wedge r$	Si existen dos categorías y la coordenada *i* de la primera categoría es menor entonces la primera categoría está a la derecha de la segunda y ausencia de entorno
$p_1 \wedge p_2 \wedge (j^{p_1} > j^{p_2}) \wedge entorno \rightarrow p_1\ C(p_2) \wedge entorno$	Si existen dos categorías y la coordenada *i* de la primera categoría es menor entonces la primera categoría está a la derecha de la segunda
$p_1 \wedge p_2 \wedge (i^{p_1} > i^{p_2}) \rightarrow p_1\ D(p_2) \wedge r$	Si existen dos categorías y la coordenada *i* de la primera categoría es mayor entonces la primera categoría está a la izquierda de la segunda y ausencia de entorno
$p_1 \wedge p_2 \wedge (j^{p_1} < j^{p_2}) \wedge p_3 \wedge entorno \rightarrow p_1\ A(p_2) \wedge p_3 \wedge entorno$	Si existen dos categorías y la coordenada *j* de la primera categoría es menor entonces la primera categoría está arriba de la segunda y una tercera categoría en un entorno
$p_1 \wedge p_2 \wedge (j^{p_1} > j^{p_2}) \wedge p_3 \rightarrow p_1\ B(p_2) \wedge p_3$	Si existen dos categorías y la coordenada *j* de la primera categoría es mayor entonces la primera categoría está abajo de la segunda y una tercera categoría
$p_1 \wedge p_2 \wedge (j^{p_1} > j^{p_2}) \wedge p_3 \rightarrow p_1\ B(p_2) \wedge p_3 \wedge r$	Si existen dos categorías y la coordenada *j* de la primera categoría es mayor entonces la primera categoría está abajo de la segunda y una tercera categoría en ausencia de entorno
$p_1 \wedge p_2 \wedge (j^{p_1} > j^{p_2}) \wedge p_3 \wedge entorno \rightarrow p_1\ B(p_2) \wedge p_3 \wedge entorno$	Si existen dos categorías y la coordenada *j* de la primera categoría es mayor entonces la primera categoría está abajo de la segunda y una tercera categoría en un entorno
$p_1 \wedge p_2 \wedge (i^{p_1} < i^{p_2}) \wedge p_3 \rightarrow p_1\ C(p_2) \wedge p_3$	Si existen dos categorías y la coordenada i de la primera categoría es menor entonces la primera categoría está a la derecha de la segunda y una tercera categoría
$p_1 \wedge p_2 \wedge (i^{p_1} < i^{p_2}) \wedge p_3 \rightarrow p_1\ C(p_2) \wedge p_3 \wedge r$	Si existen dos categorías y la coordenada *i* de la primera categoría es menor entonces la primera categoría está a la derecha de la segunda y una tercera categoría en ausencia de entorno
$p_1 \wedge p_2 \wedge (i^{p_1} < i^{p_2}) \wedge p_3 \wedge entorno \rightarrow p_1\ C(p_2) \wedge p_3 \wedge entorno$	Si existen dos categorías y la coordenada *i* de la primera categoría es menor entonces la primera categoría está a la derecha de la segunda y una tercera categoría en un entorno
$p_1 \wedge p_2 \wedge (i^{p_1} > i^{p_2}) \wedge p_3 \rightarrow p_1\ D(p_2) \wedge p_3$	Si existen dos categorías y la coordenada *j* de la primera categoría es mayor entonces la primera categoría está a la izquierda de la segunda y una tercera categoría
$p_1 \wedge p_2 \wedge (i^{p_1} > i^{p_2}) \wedge p_3 \rightarrow p_1\ D(p_2) \wedge p_3 \wedge r$	Si existen dos categorías y la coordenada *i* de la primera categoría es mayor entonces la primera categoría está a la izquierda de la segunda y una tercera categoría en ausencia de entorno
$p_1 \wedge p_2 \wedge (i^{p_1} > i^{p_2}) \wedge p_3 \wedge entorno \rightarrow p_1\ D(p_2) \wedge p_3 \wedge entorno$	*i* existen dos categorías y la coordenada *i* de la primera categoría es mayor entonces la primera categoría está a la izquierda de la segunda y una tercera categoría en un entorno

CAPÍTULO 5

EXPERIMENTACIÓN Y RESULTADOS

En esta seccion, se muestran los resultados obtenidos a lo largo de este trabajo, comenzando por los más antiguos y divididos en cinco secciones. La primera sección corresponde a la experimentación con el elemento *categorías.* La segunda sección corresponde a los elementos *categorías* y *jerarquías.* La tercera sección corresponde a pruebas con los elementos *categorías, jerarquías* y *distribuciones.* La cuarta sección corresponde a los experimentos con los elementos *categorías, jerarquías, distribuciones* y *relaciones.* Finalmente la quinta sección corresponde a la experimentación con el modelo completo aplicando el *mecanismo de inferencias.*

5.1 Ambiente de pruebas

Las pruebas se realizaron en un equipo con sistema operativo Windows con procesador i7, 16gb de memoria RAM. Las imágenes utilizadas fueron tomadas del banco de imágenes *Pascal Voc* [38]*, Open Image* [42].

En esta sección, se muestran visualmente las clases trabajadas en esta investigación mediante redes semánticas. En la Figura 5.1, se muestra la red semántica de la clase transporte.

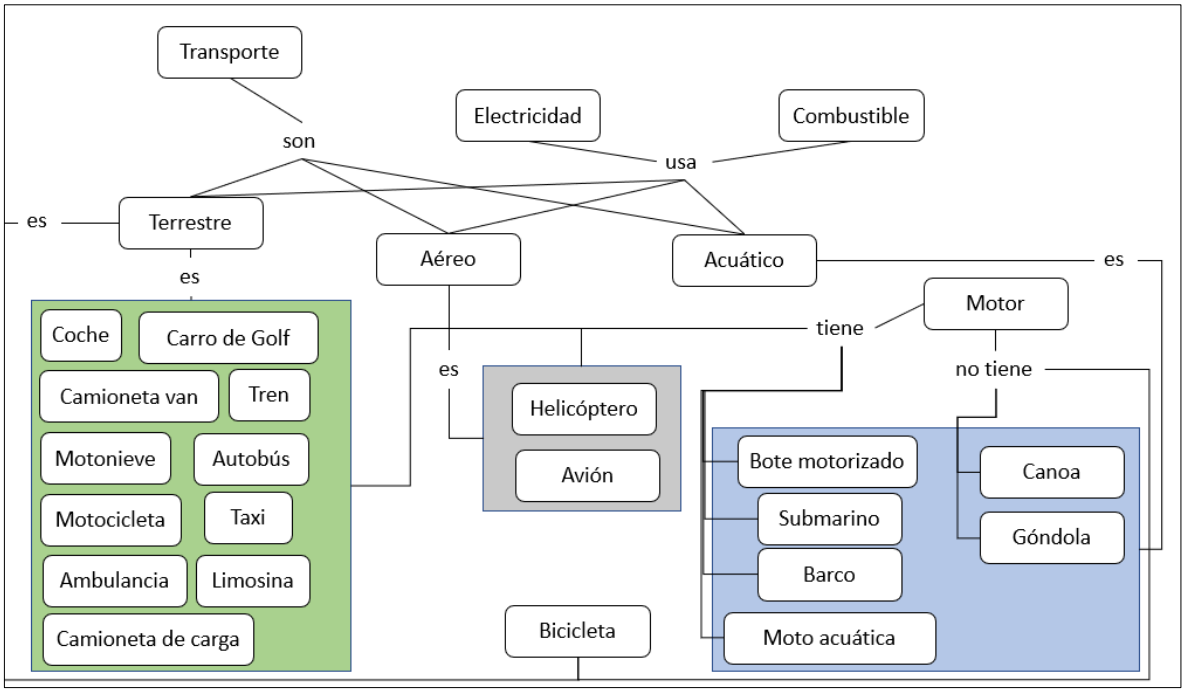


Figura 5.1 Red semántica de la clase transporte

Se consideran transportes acuáticos, terrestres y aéreos tanto motorizados como no motorizados. En total para las pruebas del modelo propuesto se trabajó con 21 transportes. Para la clase de entorno se consideró construcciones, cielo, caminos, objetos de entorno, vegetación. En total se trabajó con 40 clases de entorno, desde flores hasta hidrante, el entorno está dividido en seis súper clases.

En la Figura 5.2, se muestra la red semántica de la clase entorno.

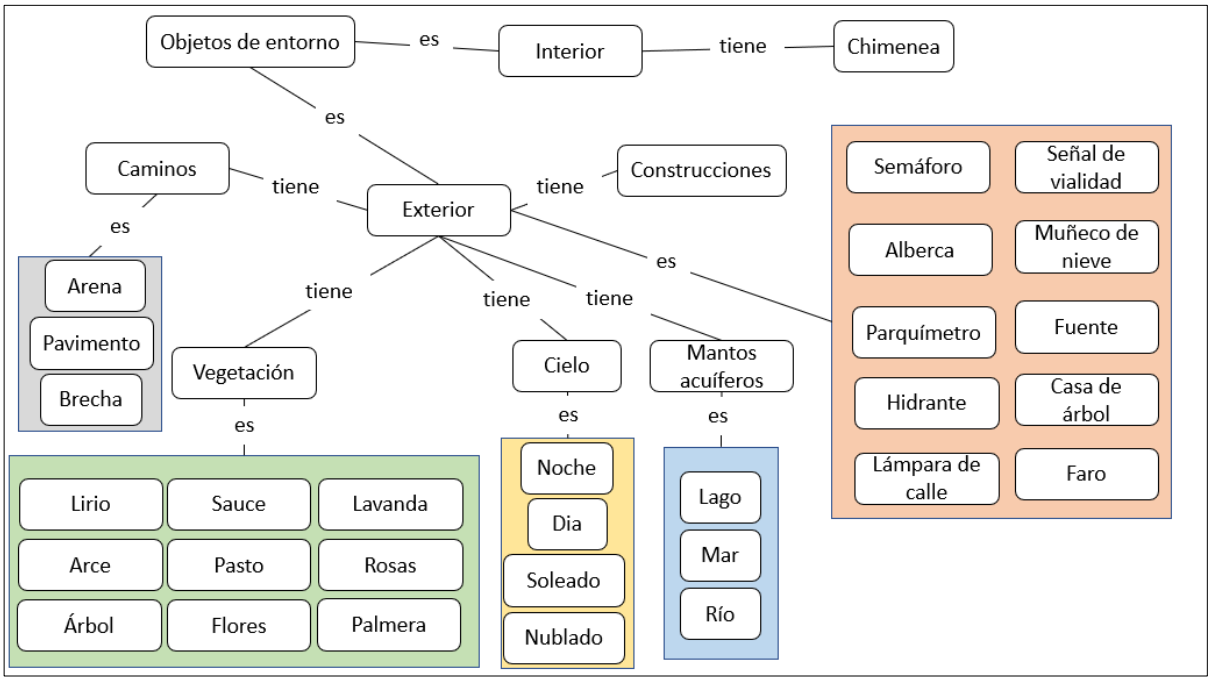


Figura 5.2 Red semántica de la clase entorno

Como se puede apreciar en la Figura 5.2, se tienen entornos en donde se contempla interiores y exteriores y en exteriores diversos tipos de escenarios donde se considera el cielo, la vegetación, construcciones, mantos acuíferos.

En la Figura 5.3, se muestra la red semántica de alimentos y se consideraron siete súper clases, frutas donde se incluye mango, uva, pepino. Verduras entre ellas calabacín, brócoli. Bebidas. Platillos de alimentos. Comida rápida, es decir pizza, hamburguesas. Meriendas y mariscos, con un total de 65 clases.

En la Figura 5.4, se muestra la red semántica de animales la cual cuenta con 91 clases, los animales se dividen entre mamíferos, aves, animales acuáticos.

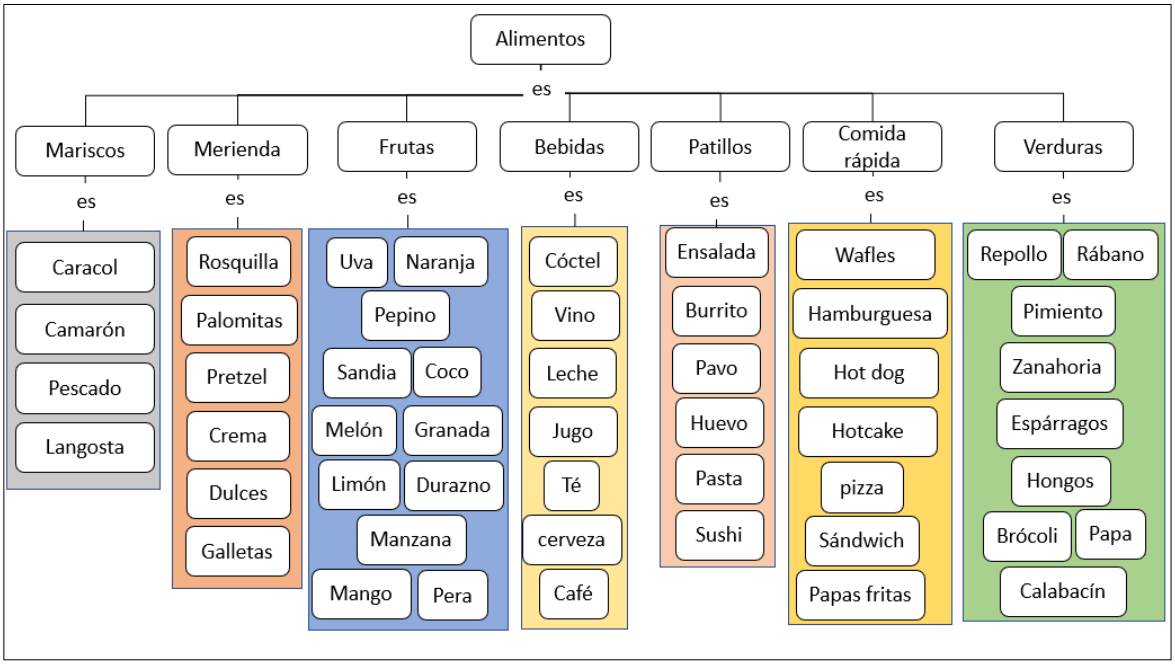


Figura 5.3 Red semántica de la clase alimentos

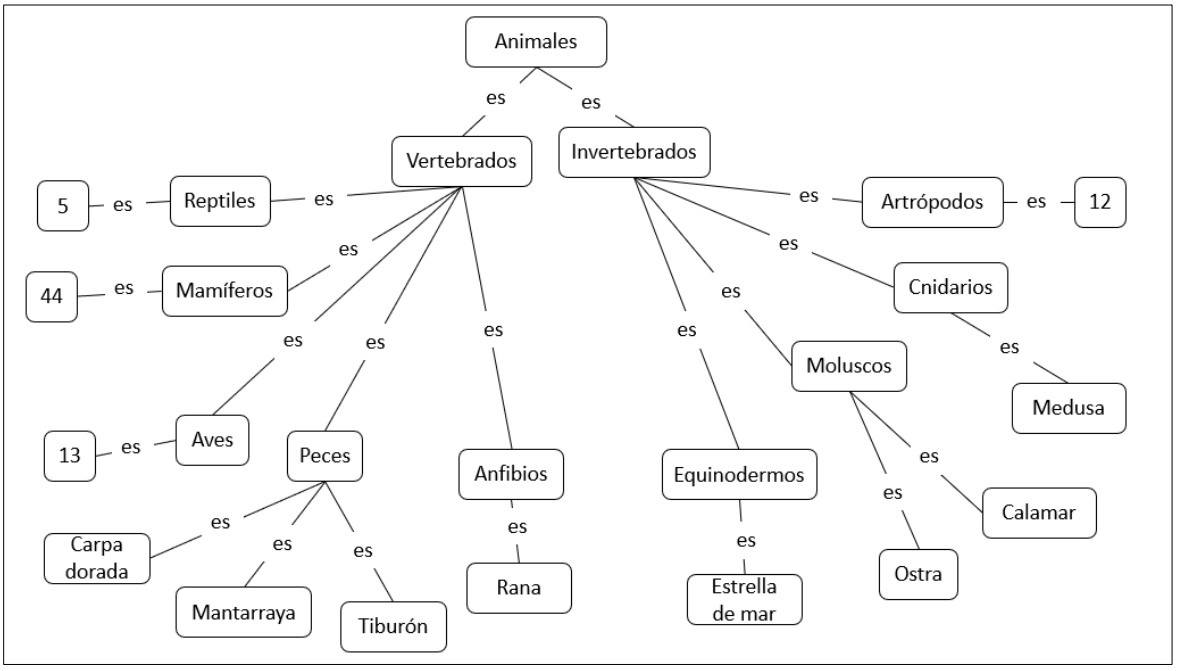


Figura 5.4 Red semántica de la clase animales

En la Figura 5.4, se muestra la red semántica de la clase animales y en el anexo A la tabla con la información completa.

En la Figura 5.5, se muestra la red semántica de la clase objetos, en este caso se tienen 348 clases y únicamente se agregó el número ya que no es posible agregar la gran cantidad de información en la imagen y que sea legible, se muestra únicamente las clases de nodos superiores y en el anexo B la tabla con la información completa.

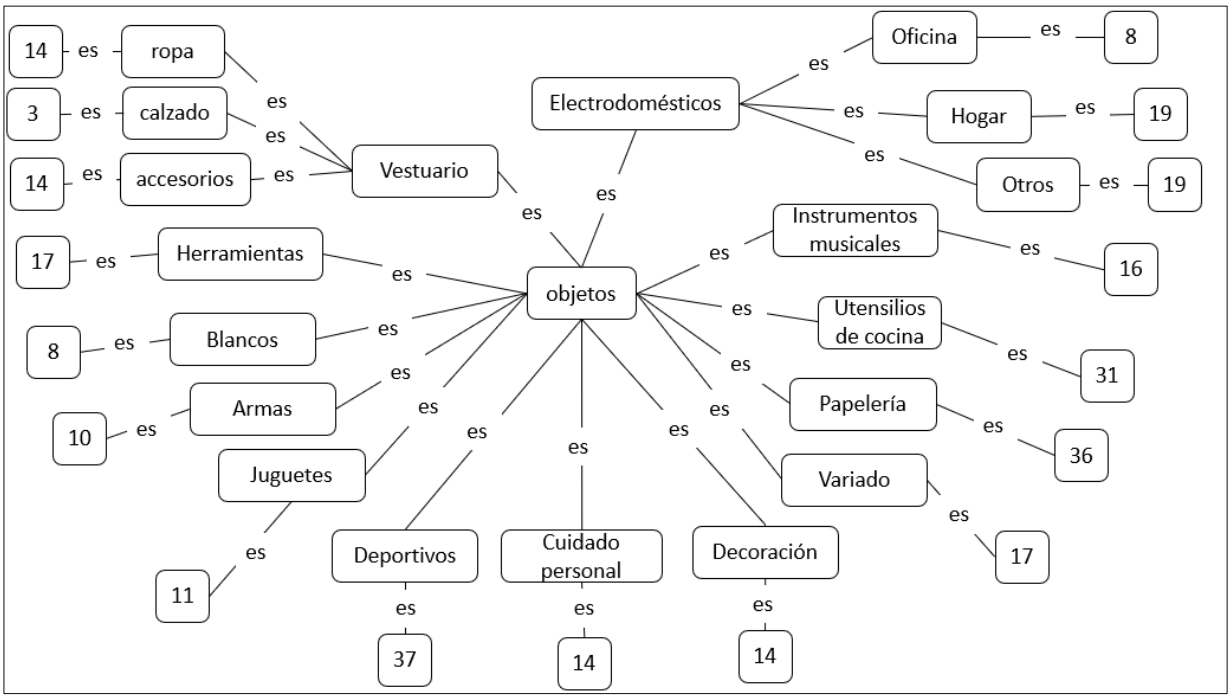


Figura 5.5 Red semántica de la clase objetos

En este trabajo, el entorno de las imágenes es considerado, ya que proporciona el escenario de los elementos y, por ende, la descripción se dará con mayor naturalidad, es importante mencionar que entre mayor cantidad de datos de entrada se le proporcionen al modelo, mayor detalle tendrán las descripciones. Otro elemento considerado fue el orden de prioridad de los elementos que aparecen en la imagen, es decir, al momento de describir se considera cual objeto es el que tiene mayor relevancia.

En la Figura 5.6, se muestra el diagrama de prioridad de clase para relacionar y generar descripciones.

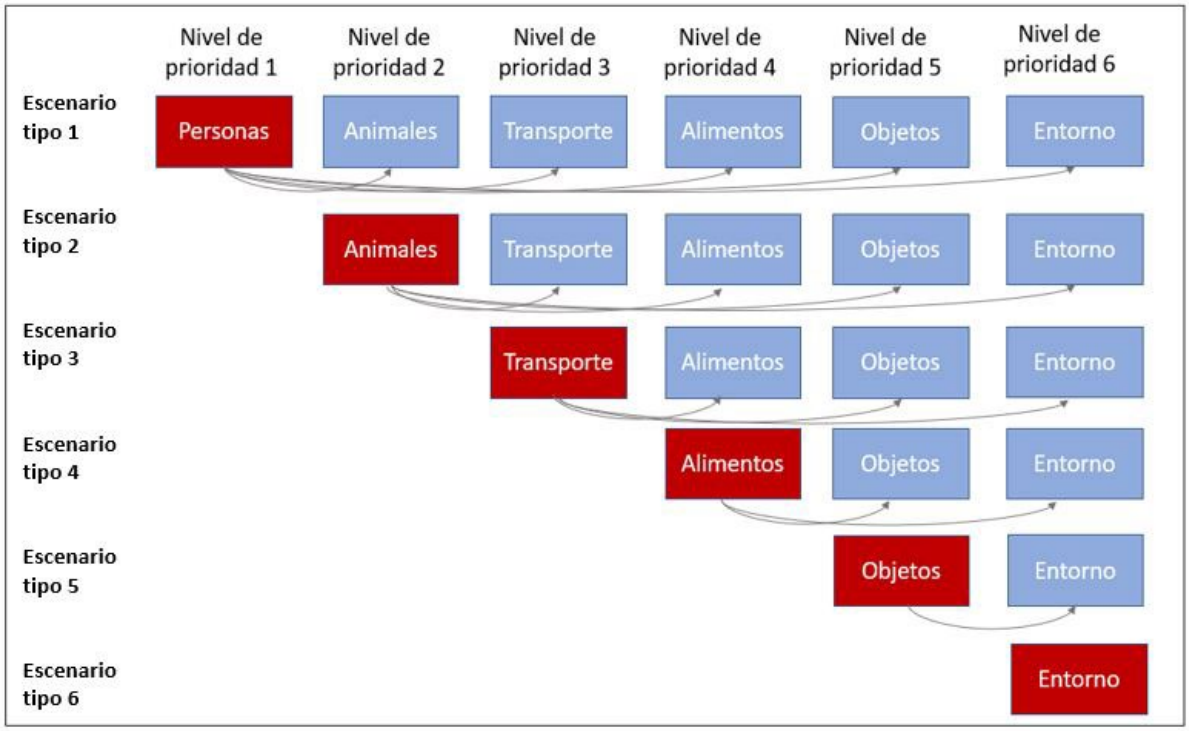


Figura 5.6 Diagrama de prioridad de clases y escenario

Como se puede apreciar en la Figura 5.6, se describe cómo está organizada la prioridad de clases, esta prioridad fue asignada con base a un análisis en el banco de representación del conocimiento RIDeCS, en el cual, se consideró muestreo estadístico para la fiabilidad de la información de un total de 480,000,000 de personas en el mundo que hablan español la muestra representativa es de 385 considerando un margen de error del 5% y un nivel de confianza del 95%.

Las relaciones no requieren estar ligadas todas entre sí, en algunos casos las imágenes solo contienen personas, en otros animales, transporte, alimentos, objetos o únicamente entorno en el último caso aplica para imágenes de paisajes donde lo que destaca es la vegetación, objetos de la naturaleza.

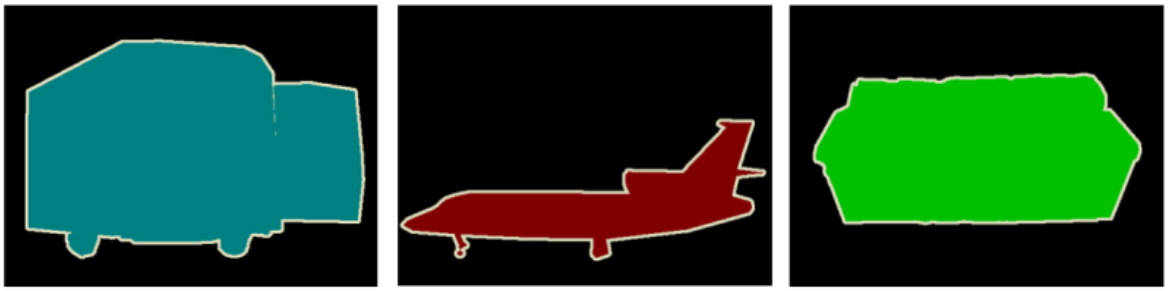
RIDeCS Es un repositorio de imágenes que se creó como complemento a esta tesis con la finalidad de formar un banco de conocimiento a partir de las descripciones de humanos, las cuales fueron brindadas por personas de entre 5 a 65 años que colaboraron describiendo lo que observaban en las imágenes. Estas descripciones fueron usadas en las pruebas del modelo propuesto en este trabajo de investigación.

Fue importante conocer la edad de las personas que describieron el contenido de las imágenes, ya que el razonamiento entre un adulto y un niño no es el mismo. De esta manera se logró medir el desempeño de los algoritmos computacionales y conocer el rango de edad al que pertenecen las descripciones.

Dicho repositorio cuenta actualmente con dos versiones, la primera está conformada por cien imágenes del repositorio y se usaron para comparar las respuestas del modelo con las respuestas de personas describiendo la misma imagen, la comparativa se realizó mediante métricas de descripción semánticas. En la segunda versión, estas respuestas se utilizaron para ajustar el mecanismo de inferencias, proporcionando al modelo la capacidad de realizar descripciones más naturales. En la versión dos, se ingresaron nuevas imágenes, las cuales se solicitó fueran descritas por personas nuevamente y de esta manera comparar los resultados de las descripciones de las personas, con las del modelo ya ajustado con la etapa anterior.

5.1.1 Experimentación con la categorización

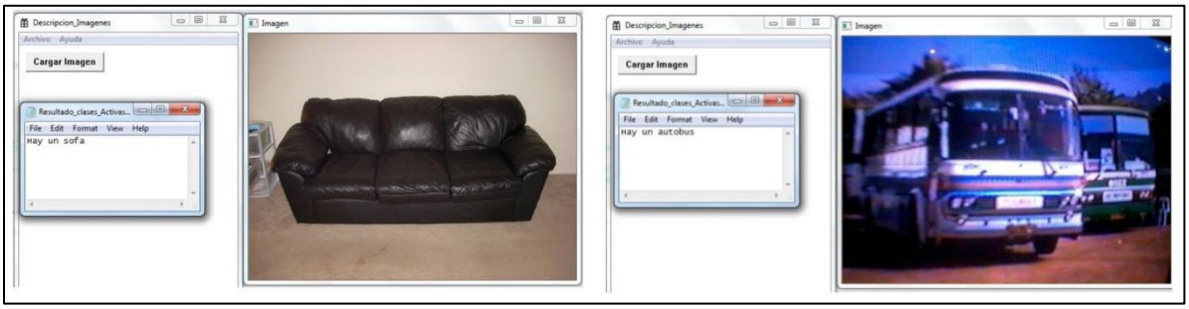
En esta sección del documento, se muestran algunos experimentos y resultados involucrando unicamente el elemento *categorías*. Esto para comenzar con la entrada de datos al modelo, el banco de imágenes utilizado fue Pascal VOC 2012 [38]. A continuación se documentan tres experimentos. En la Figura 5.7, se muestra un *ground truth* correspondiente a los objetos de la Figura 5.8.



a) autobús **b) avión** **c) sofá**

Figura 5.7 *Ground Truth* del banco de imágenes pascal VOC 2012

Después se cargó la categoría recibida en una interfaz y su etiqueta de contenido; en este punto del desarrollo sólo era posible cargar la información de una sola categoría y la información de la clase se guardaba en un archivo de texto. En la Figura 5.8, se muestra la clase de entrada.



a) sofá **c) autobús**

Figura 5.8 Clases reconocidas del banco de imágenes Pascal VOC 2012

En la Figura 5.9, se muestra un *Ground truth* en formato *Java Script Object Notation (JSON)*, el cual alimentará al modelo con sus datos. El formato válido para el ingreso

de datos al modelo es mediante el estándar JSON [92] el cual está basado en texto estándar para representar datos estructurados en la sintaxis de objetos de Java Script. Tienen otro formato de texto con la información organizada en estructuras con extensión JSON. Un archivo JSON se puede crear con cualquier editor de texto plano. Una vez que se tienen los datos que alimentarán al modelo, se valida cuáles de las clases están activas.

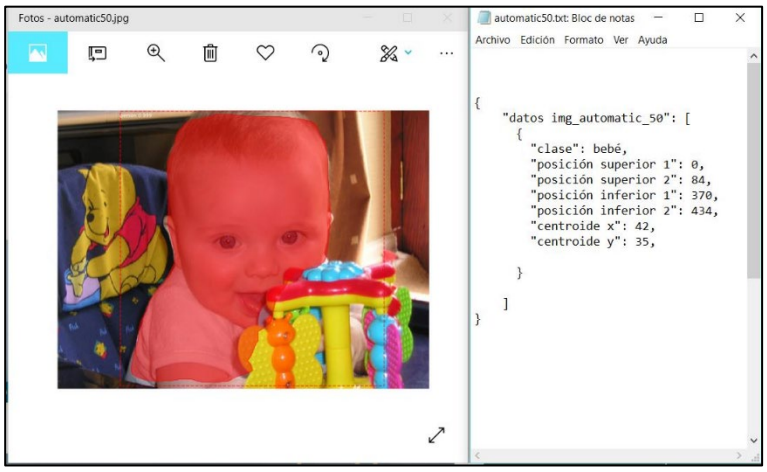


Figura 5.9 Formato de JSON usado como anotaciones de texto para la imagen de su costado

En la Figura 5.10, se muestra la verificación sintáctica de la entrada al modelo mediante un generador de analizadores léxicos rápidos haciendo uso de *LEX - FLEX* [93].

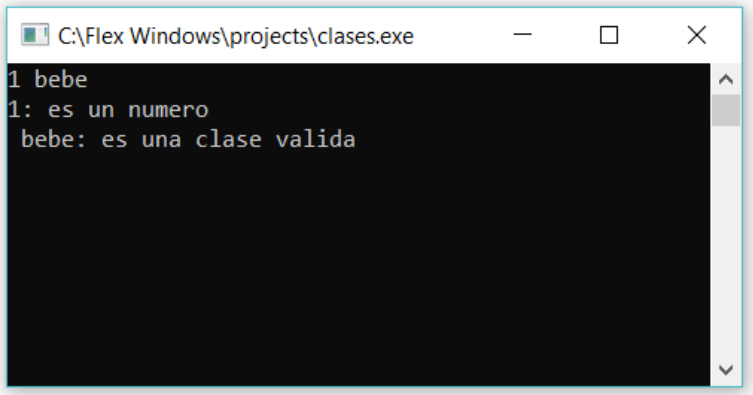


Figura 5.10 Análisis sintáctico de la entrada al sistema

Como se puede apreciar en la figura anterior en el primer renglón dice: 1 bebé, esta línea es el *Ground truth* que viene del sistema que está alimentando al modelo. En el segundo renglón analiza que 1 es un número. Finalmente, la clase bebé la cual es una clase aceptada, aparece con la leyenda “es una clase válida”. Cabe mencionar que en la salida del sistema *LEX - FLEX* no se permite el uso de acentos.

Una vez que son validadas qué clases sí pertenecen al lenguaje definido para trabajar en el modelo se cargan los datos de cada objeto. En la Figura 5.11, se muestra el resultado de la información que ingresó al sistema, sin embargo, al tener un único elemento no es posible generar una descripción semántica como tal.

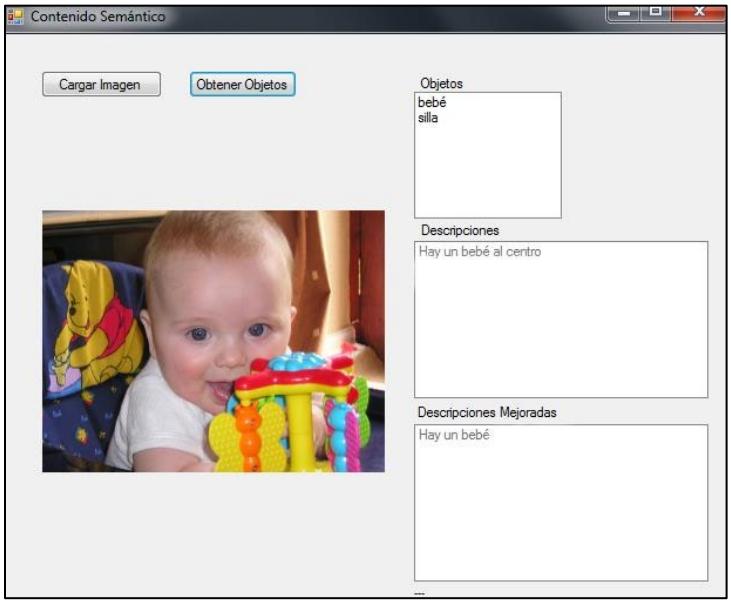


Figura 5.11 Descripción de los elementos contenidos en la imagen, descripción y descripción mejorada sin éxito debido a la escasa información

Posteriormente se dejó de lado *LEX - FLEX* para las verificaciones sintácticas y se construyó el elemento correspondiente a estas tareas.

5.1.2 Experimentación con la distribución espacial

En esta sección del documento, se muestran algunos experimentos y resultados involucrando los elementos categorías y distribuciones. Esto para identificar en qué

region de la imagen se encuentra cada categoría. A continuación se documentan tres experimentos relacionados con los elementos categorización y distribuciones.

En la Figura 5.12, se muestra la imagen de un gato del banco de imágenes Pascal VOC 2012 [38], en la cual la distribución espacial de los objetos ha sido almacenada en el elemento distribuciones. Se puede apreciar que hay un gato, una silla, una mesa y estos elementos estarían conformando un comedor y el gato sobre ella.

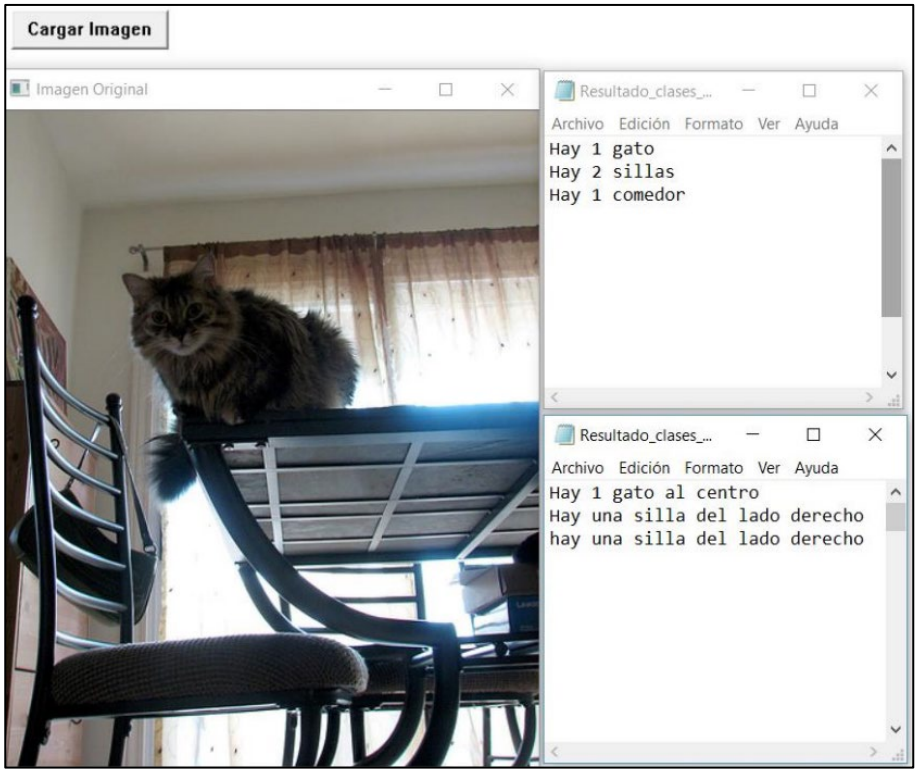


Figura 5.12 Descripción de los elementos de las categorías contenidas en la imagen de un gato y su distribución

En la Figura 5.13, se muestra la imagen de tres personas en la calle junto a un coche, en este caso son personas adultas y están en la *pose* parado.

Figura 5.13 Descripción de los elementos y su distribución para una imagen de ciudad

En la Figura 5.14, se observa el resultado de la integración de las variables de entorno en el proceso de descripción semántica y se muestra una imagen de exterior con descripciones y distribución de elementos.

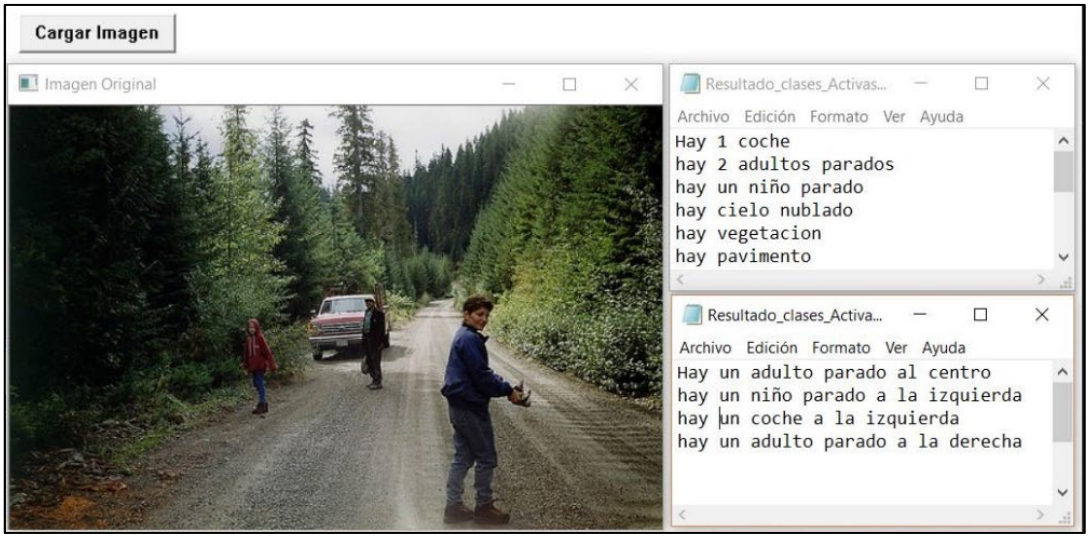


Figura 5.14 Descripción de los elementos de un paisaje de campo y su distribución

5.1.3 Experimentación con jerarquías

En esta sección del documento, se muestran algunos experimentos y resultados involucrando los elementos *categorías, distribuciones y jerarquías*. El banco de imágenes utilizado fue Pascal VOC [38]. Estas pruebas fueron realizadas para obtener descripciones semánticas con diferentes niveles jerárquicos. A continuación se documentan tres experimentos relacionados con los elementos *categorizacion, distribuciones y jerarquías*.

La experimentación se llevó a cabo con diversas imágenes con mayor y menor riqueza para mostrar la importancia de los datos para describir semánticamente, no todas las imágenes se pueden describir de esta manera. Habrá casos donde se tienen pocos objetos y no se pueda inferir o describir mucho, como se muestra en la Figura 5.15.

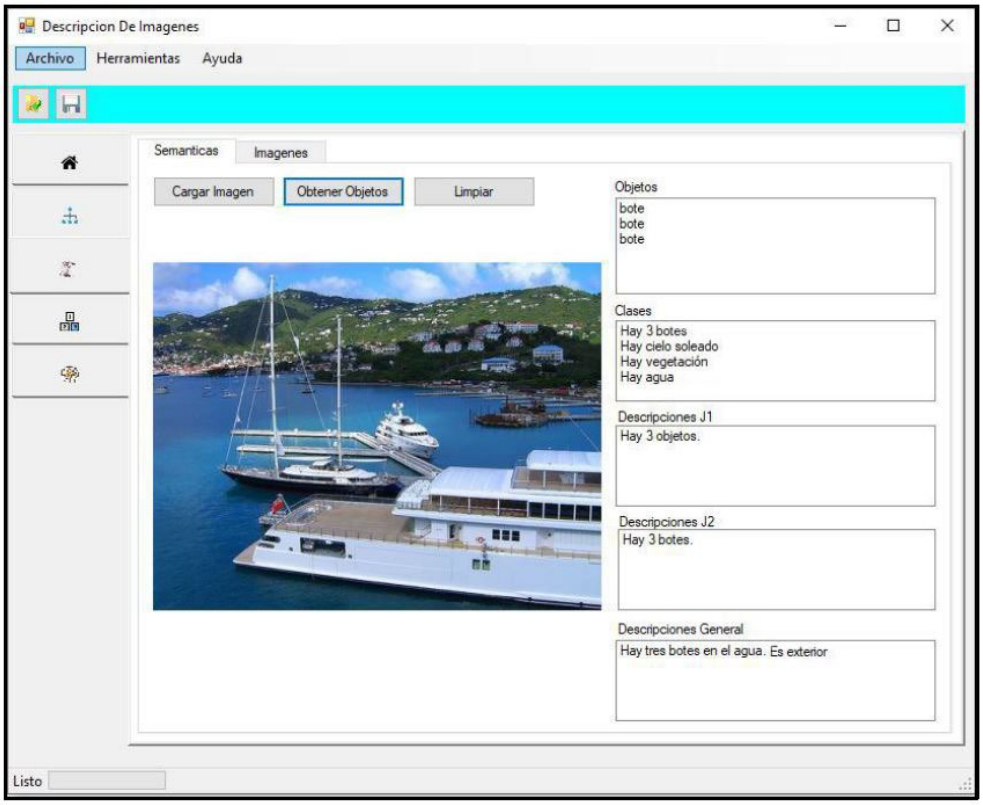


Figura 5.15 Imagen de prueba con escasa información para una descripción semántica

Sin embargo, en otros casos las imágenes cuentan con varios objetos y se encuentran relacionados entre ellos como se muestra en la Figura 5.16.

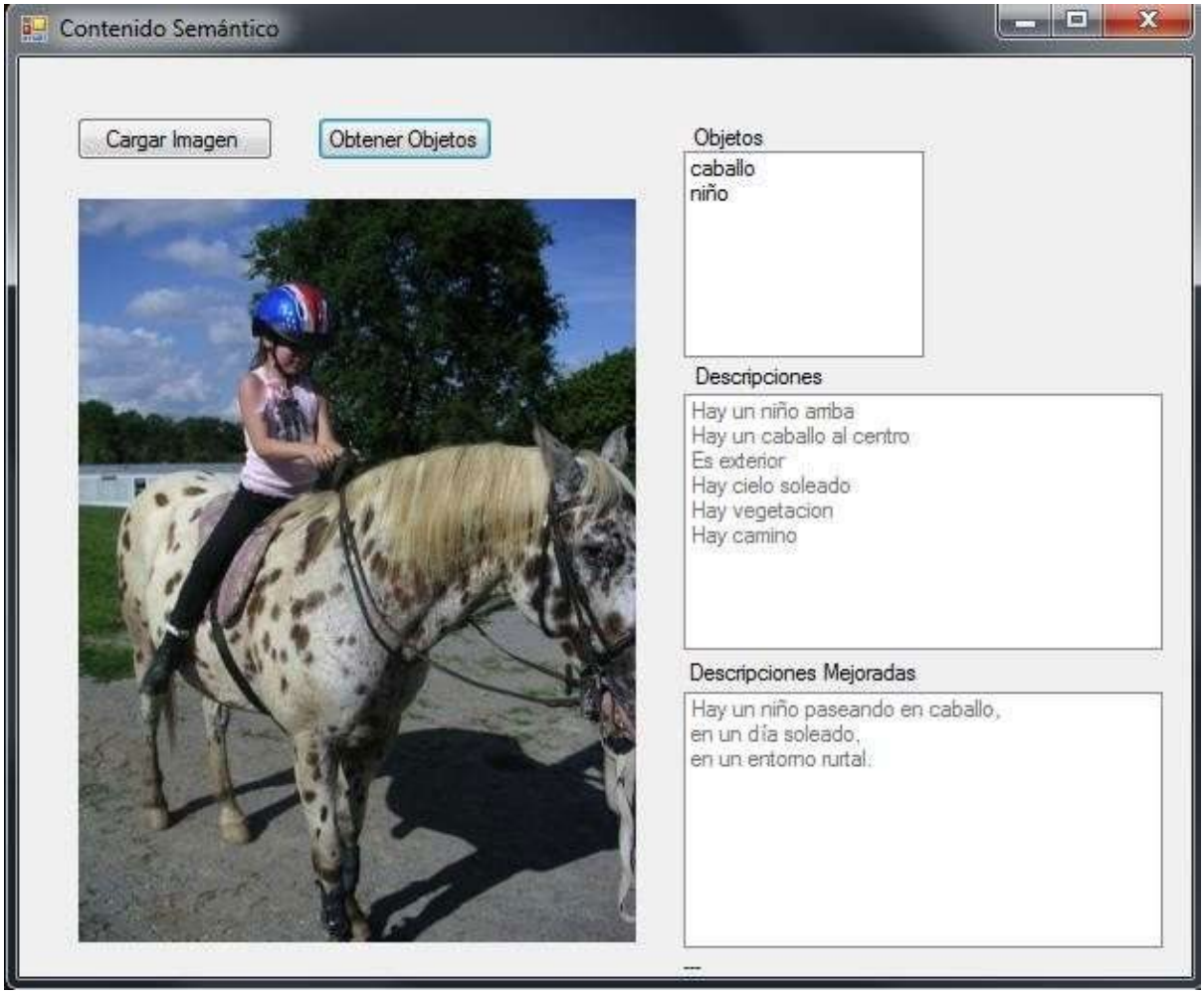


Figura 5.16 Imagen de prueba donde una niña se encuentra paseando a caballo

Como se puede apreciar en la Figura 5.16, al tener mayor riqueza de categorías es posible describir mejor una imagen. En este caso se incluyó la categoría de 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 5.17, se muestra el resultado de la descripción de otra imagen con riqueza de categorías.

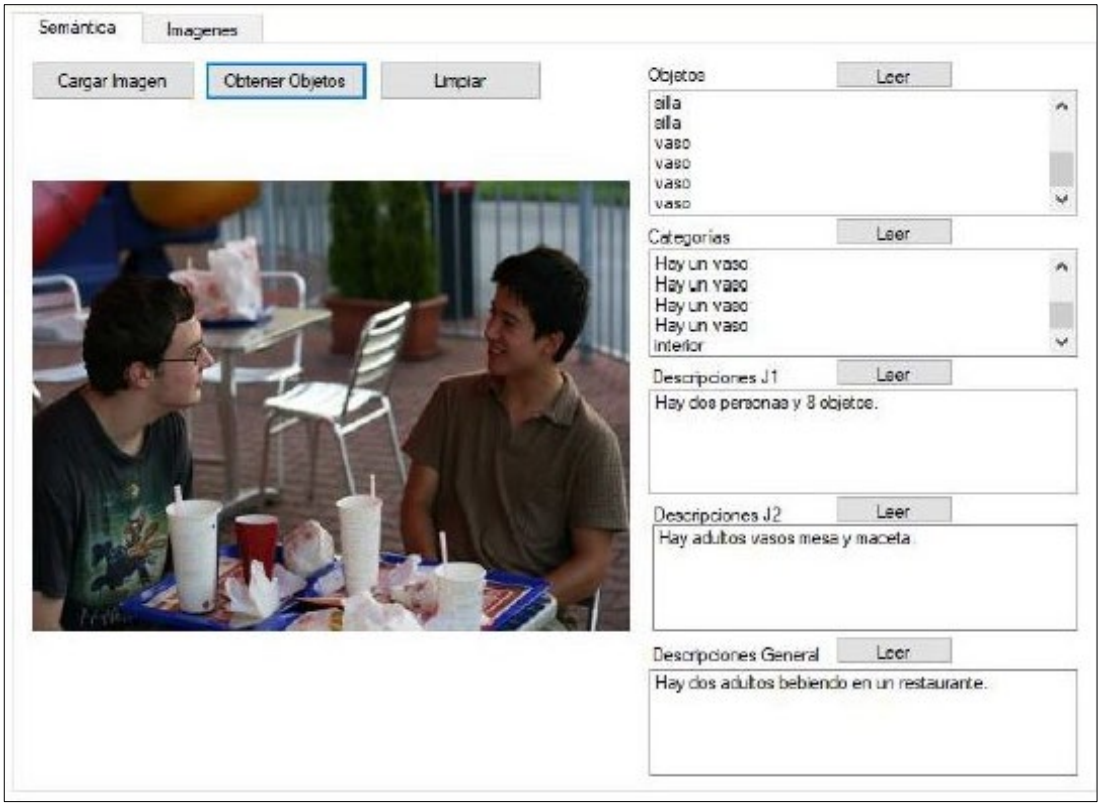


Figura 5.17 Imagen de prueba con información de entorno y diversas categorías para una descripción semántica

Como se puede apreciar en la Figura 5.17, 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.

5.1.4 Experimentación con relaciones

En esta sección del documento, se muestran algunos experimentos y resultados involucrando los elementos *categorías, distribuciones, jerarquías* y *relaciones*. Esto para unir todos los elementos del modelo propuesto. A continuación se documentan tres experimentos relacionados con los elementos *categorización, distribuciones, jerarquías,* y *relaciones*.

En la Figura 5.18, se tienen dos personas paseando en un bote, como objetos reconoce 2 adultos y un bote, como clases de entorno reconoce manto acuífero,

vegetación, y cielo. Con estas variantes en la imagen y cantidad de objetos es posible realizar las descripciones mejoradas, generando 3 oraciones, donde se toman en cuenta las jerarquías de personas y sus relaciones para lograr inferir que se trata de personas paseando en un bote, y no personas a un lado del bote, cargando un bote, etc.

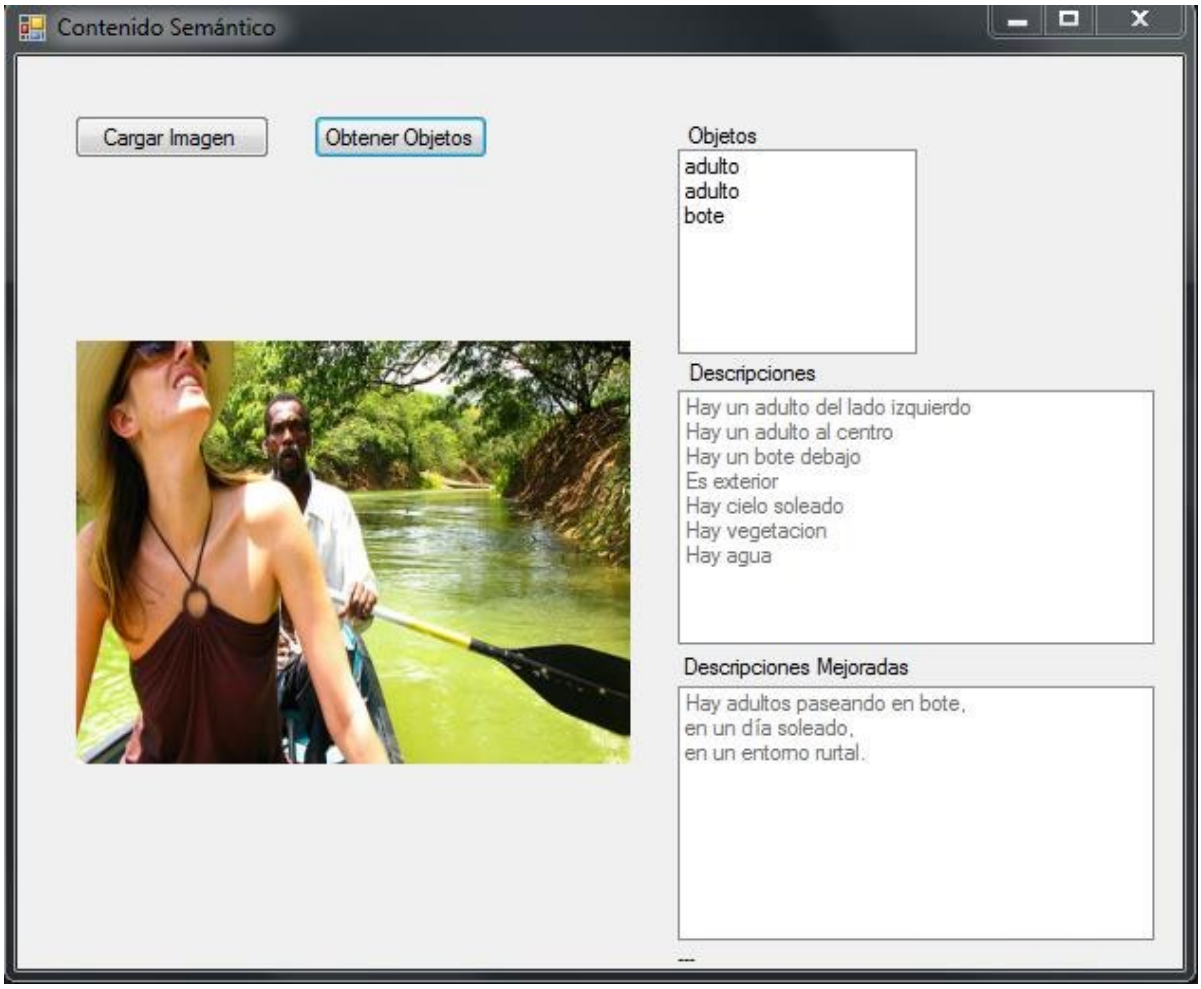


Figura 5.18 Descripción semántica de imagen de personas paseando en bote

En la Figura 5.18, se muestra la descripción semántica de una imagen perteneciente al banco de imágenes de Pascal Voc 2012, [38].

En la Figura 5.19, se tienen 2 personas comiendo juntas, no se sabe si es de mañana o de tarde dada la iluminación, sin embargo, se descarta que se trate de una cena. En objetos identificados están dos adultos, dos sillas y una mesa. Como variables de entorno el cielo y la vegetación, debido a que existen varios elementos en la imagen, fue posible relacionarlos para realizar inferencias al respecto de la imagen como una escena completa.

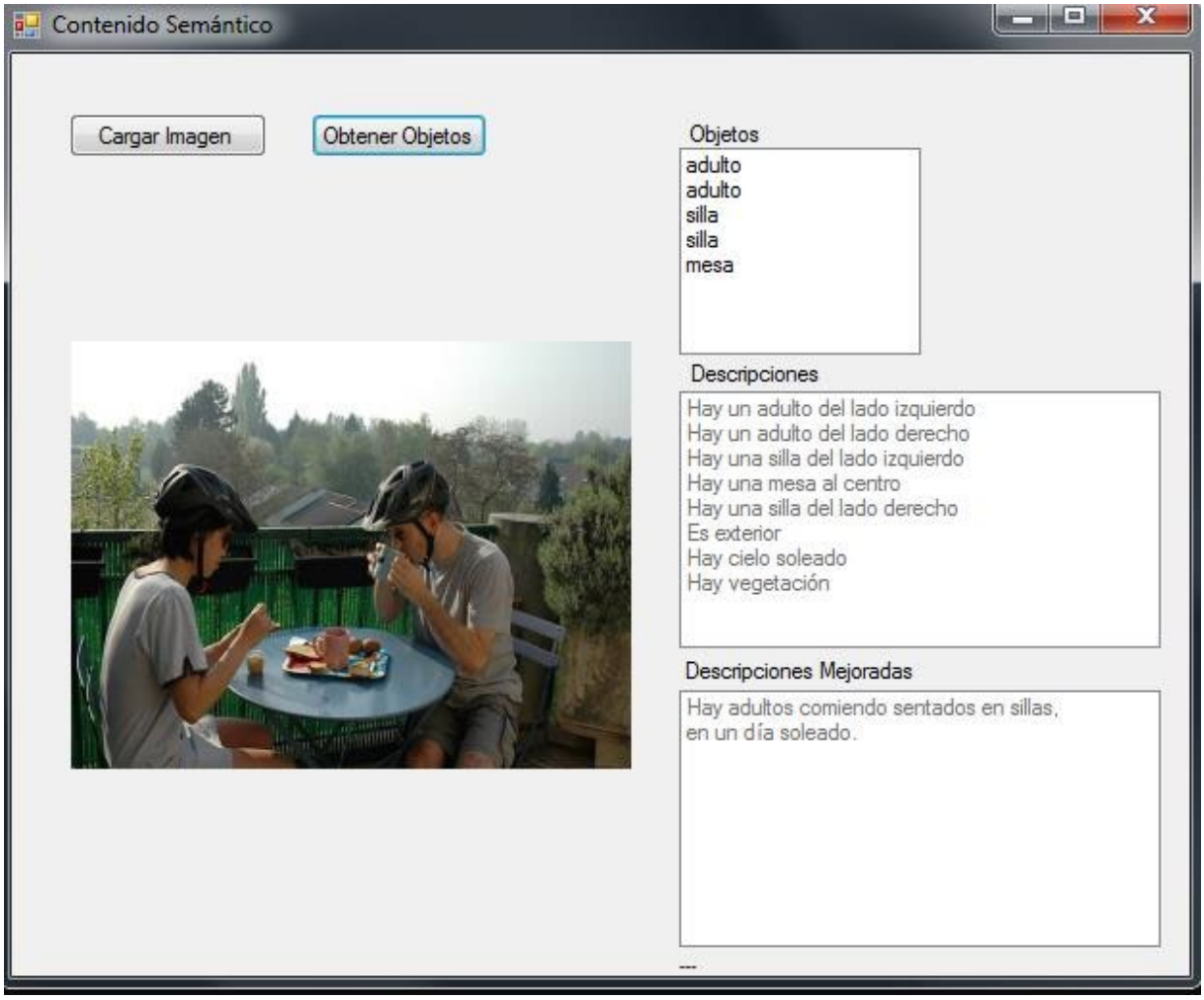


Figura 5.19 Descripción semántica de imagen de personas comiendo al aire libre

5.1.5 Experimentación con el mecanismo de inferencias

En esta sección del documento, se muestra algunos experimentos y resultados involucrando los elementos *categorías, distribuciones, jerarquías, relaciones* y aplicando *el mecanismo de inferencias* documentado en la Tabla 4.2. A continuación

se muestran tres experimentos aplicando el mecanismo de inferencias. Para las pruebas con el modelo completo se utilizaron 2 bancos de imágenes, Pascal VOC 2012 [38] y OpenImage [42]. En la Figura 5.20, se tiene a dos personas viajando en motocicleta.

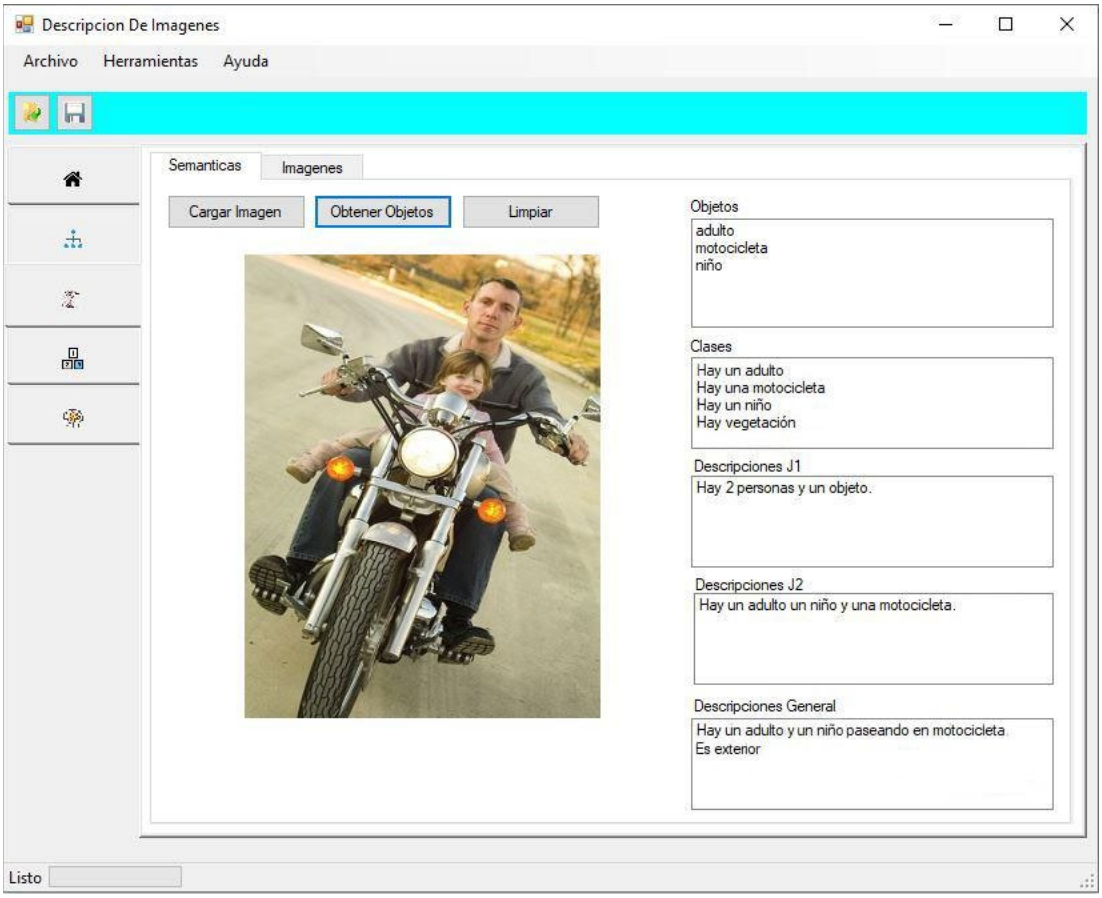


Figura 5.20 Descripción semántica considerando jerarquías para una imagen de personas paseando en motocicleta

En la interfaz se muestran cinco ventanas de descripción:

1.- Objetos: corresponde a los objetos detectados.

2.- Clases: corresponde a las categorías y se realiza un conteo de su aparición (1 adulto, 1niño, motocicleta).

3.- Descripciones J1: corresponde a la descripción considerando sólo 1 nivel de profundidad en cuanto a las jerarquías (2 personas y 1 objeto).

4.- Descripciones J2: corresponde a la descripción con 2 niveles de profundidad respecto a jerarquías (1 adulto y 1 niño).

5.- Descripción general: corresponde a la descripción general aplicando el mecanismo de inferencias, donde se describe que hay un adulto y un niño paseando en motocicleta.

En la Figura 5.21, se muestra una imagen de una persona con macetas, no se logra identificar si las está sembrando o sólo está rodeada de ellas.

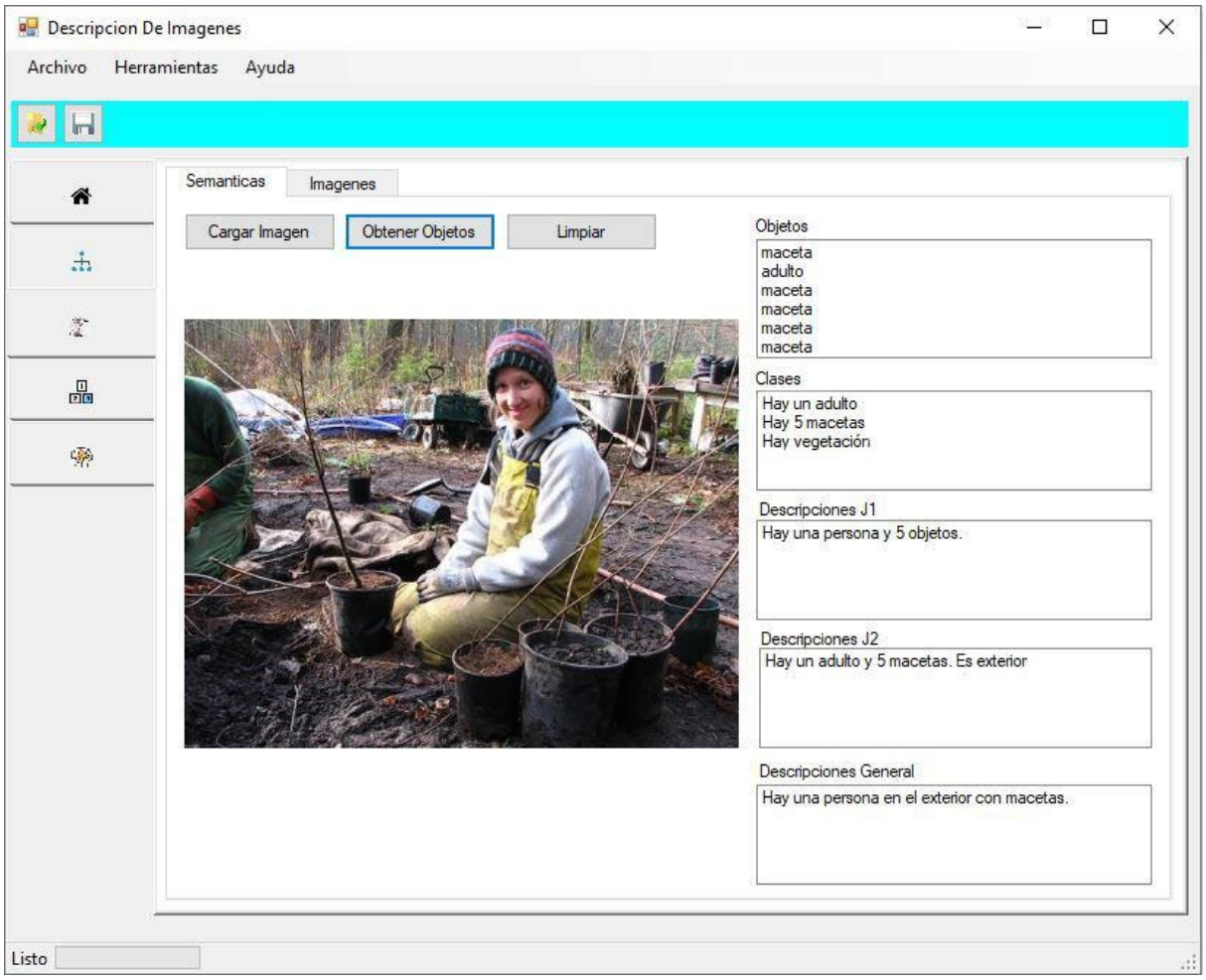


Figura 5.21 Descripción semántica considerando jerarquías para una imagen de personas con macetas

En la interfaz se muestran cinco ventanas de descripción:

1.- Objetos: Corresponde a los objetos detectados.

2.- Clases: Corresponde a las categorías y se realiza un conteo de su aparición (1 adulto, 5 macetas y variables de entorno).

3.- Descripciones J1: Corresponde a la descripción considerando sólo 1 nivel de profundidad en cuanto a las jerarquías (objetos y personas).

4.- Descripciones J2: Corresponde a la descripción con 2 niveles de profundidad respecto a jerarquías (adulto y maceta).

5.- Descripción general: Corresponde a la descripción general aplicando el mecanismo de inferencias, donde se describe que hay una persona en el exterior con macetas.

A partir de las siguientes pruebas se trabajó con el modelo completo, donde, la verificación sintáctica e inferencias son creadas para el modelo. El universo de los datos está formado por seis súper clases:

1. Personas: esta súper clase engloba todo lo referente a personas, las jerarquías fueron basadas en etapas de la vida (bebé, niño, joven, adulto, anciano).

2. Animales: en esta súper clase están considerados todos los animales y fueron jerarquizados acorde a taxonomías de animales de la comisión internacional de la nomenclatura zoológica [94].

3. Transportes: en esta súper clase se agregan todo tipo de transportes de los tres tipos posibles, aéreo, terrestre, y acuático.

4. Objetos: en esta súper clase se encuentran todos los objetos jerarquizados por tipos, tales como: muebles, trastes, ropa, etc.

5. Alimentos: en esta súper clase se tienen los alimentos y bebidas.

6. Variables de entorno: en esta súper clase se almacenan las variables de entorno, tales como: vegetación, cielos, semáforos, etc.
Cada uno de estos árboles jerárquicos corresponden a estructuras de datos expandibles y ajustables.

En la Figura 5.22, se muestran las categorías con las que se estará trabajando, perro y persona.

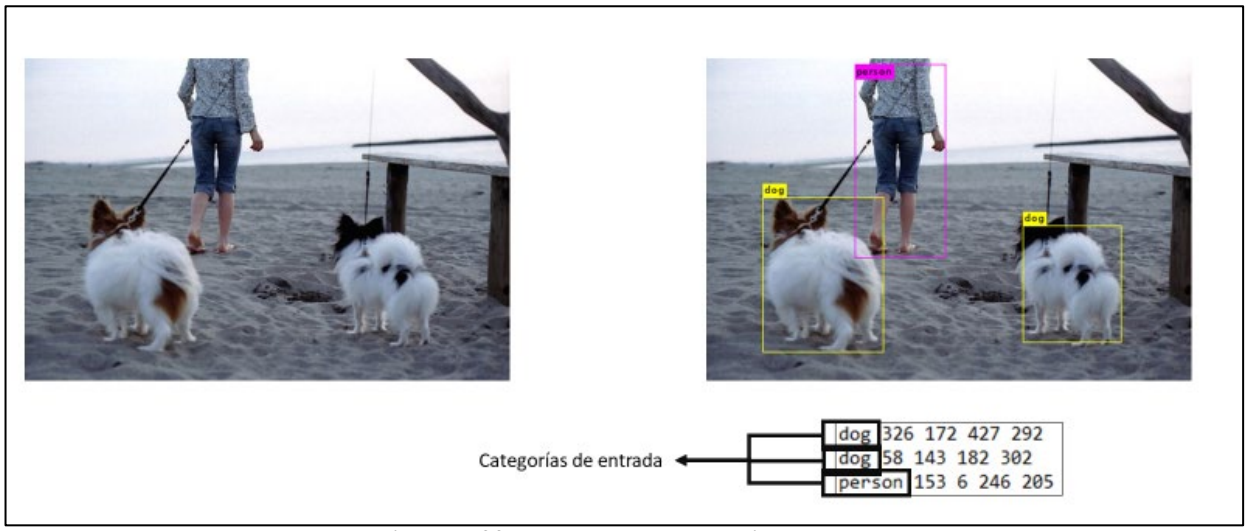


Figura 5.22 Datos de entrada al sistema

Tomando en cuenta el modelo y su representación mediante el diagrama de bloques se realiza el siguiente proceso.

1.- Datos: En esta parte son tomados los datos proporcionados por el sistema de visión, (clases y coordenadas).

2.- Procesamiento de los datos: consta de dos pasos, verificación sintáctica, diccionario de sinónimos e idiomas, y dos, la activación de las clases existentes. En la Figura 5.23, se muestran las clases existentes dentro de la base de datos, y se recupera su campo y valor, los cuales son tomados por el modelo y usados en el elemento de categorías.

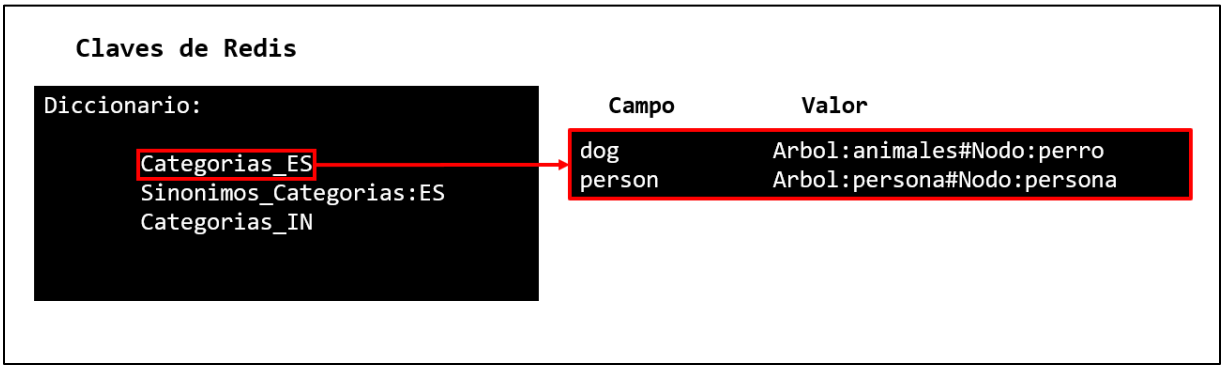


Figura 5.23 Datos de entrada al sistema y categorías a describir

3. Modelo: se divide en cinco elementos en la imagen anterior se mostró el primero de ellos *categorización*. En la Figura 5.24, se muestran las *jerarquías* de las categorías para la Figura 5.22.

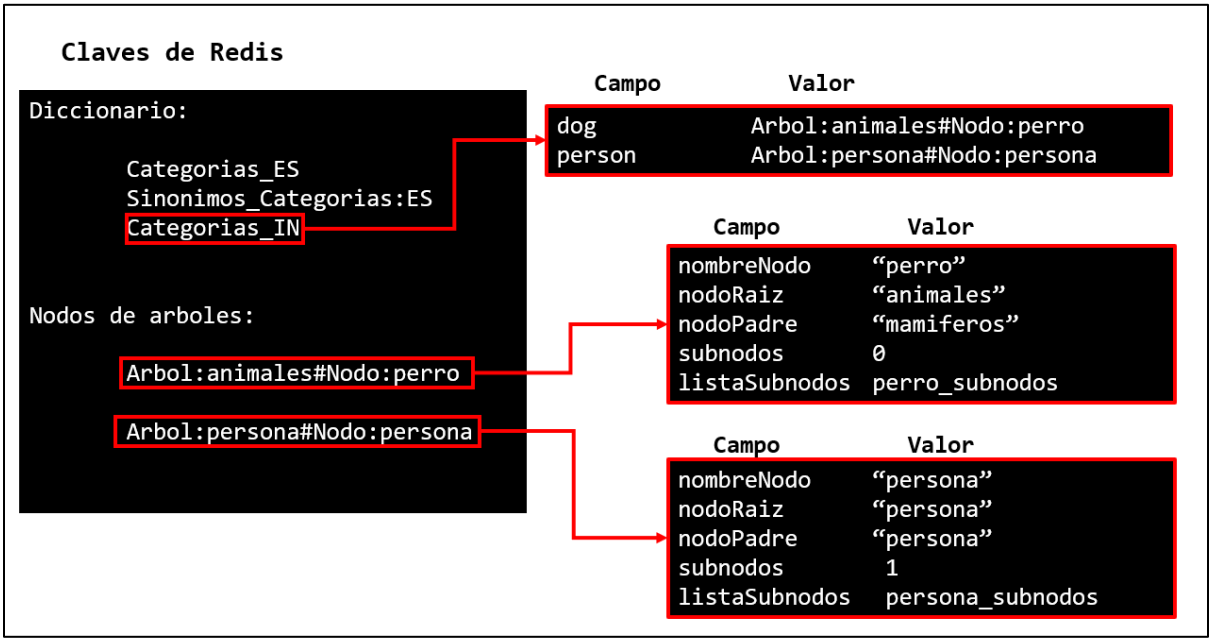


Figura 5.24 Jerarquías aplicadas a las categorías de la Figura 5.22

Una vez que los datos terminaron en jerarquías se obtienen los nodos activos dentro de los árboles de clase independientemente del nivel jerárquico en que se encuentren.

Posteriormente en la etapa de distribución, se recuperan las coordenadas de las categorías y se calcula en centroide de cada región asociada a una categoría. En la Figura 5.25, se muestran los datos del elemento *distribuciones* para la Figura 5.22.

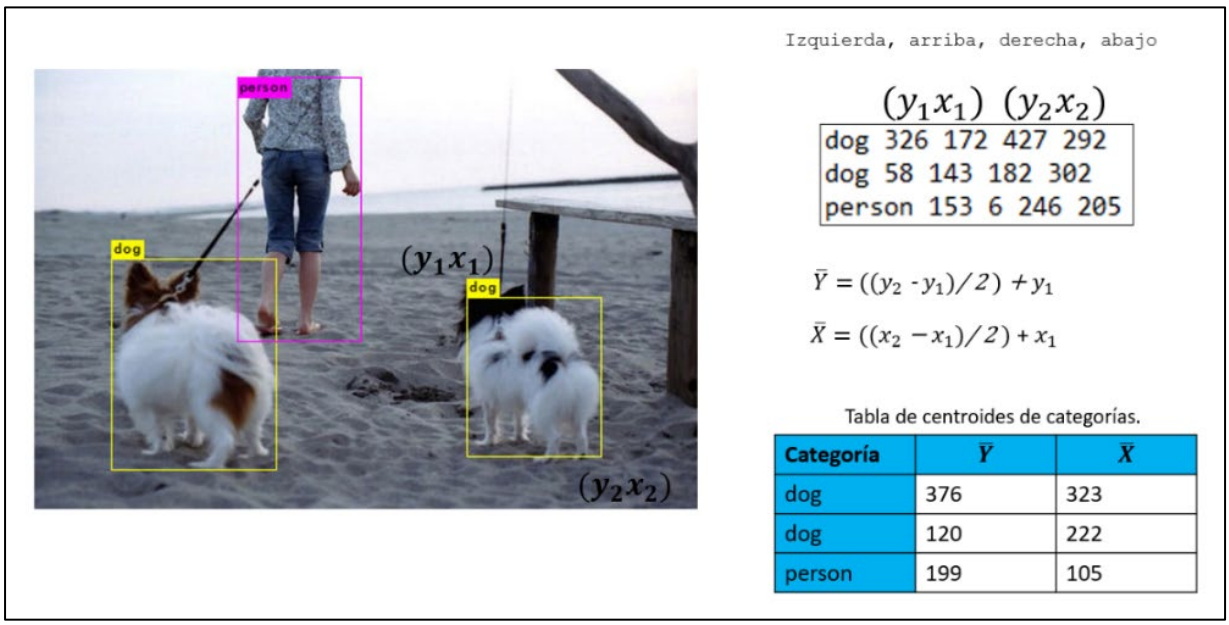


Figura 5.25 Salida del elemento *distribuciones*

4. Relaciones: en este elemento del modelo se relacionan las categorías presentes. En el caso de la Figura 5.25, los objetos a relacionar son dos perros y una persona, a cada uno de los objetos se le calcula la distancia con el otro para verificar si están interactuando o únicamente son objetos aislados que forman parte de la escena. En este caso se encuentra un perro del lado izquierdo y uno del lado derecho y al centro de la imagen se tiene una persona, la cual lleva ambos perros con correa en una escena de playa. Para calcular la distancia entre los objetos se utilizó la distancia euclidiana, Ecuación 3.

En la Figura 5.26, se muestra el resultado del cálculo de relaciones de la Figura 5.25 misma donde se muestran dos perros con una persona paseando y del lado derecho las categorías, así como su ubicación.

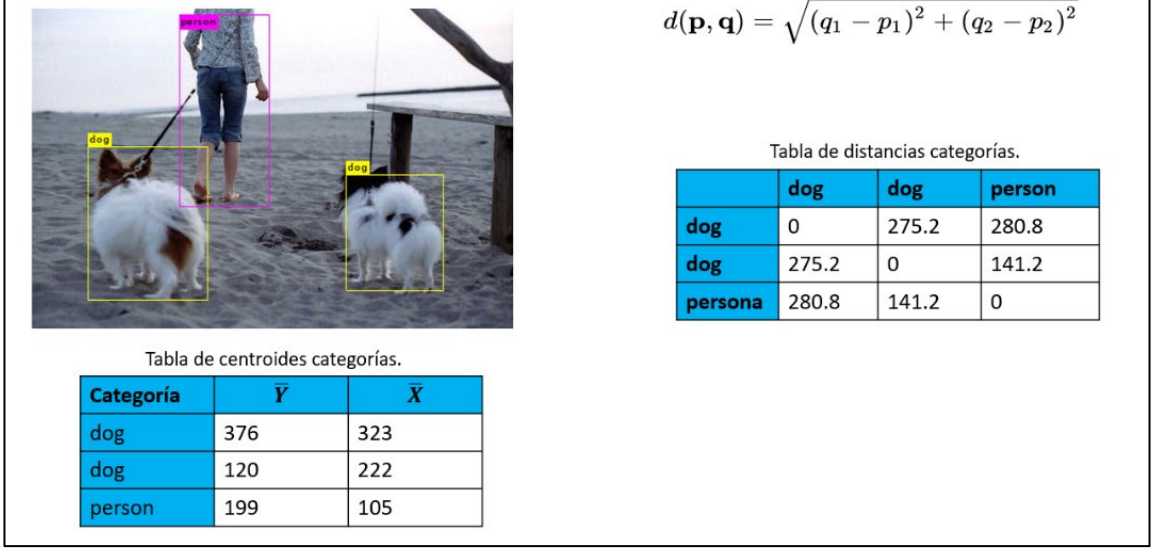


Figura 5.26 Salida del elemento *relaciones*

Finalmente entran los datos al mecanismo de inferencias, en la Tabla 5.1, se muestran las reglas de inferencias, aplicadas al experimento con la Figura 5.22.

Tabla 5.1 Reglas aplicadas a Figura 5.22

Regla	Inferencia
$entorno_1 \wedge entorno_2 \wedge entorno_3 \rightarrow q$	*arena* $\wedge$ *agua* $\wedge$ *cielo nublado* $\rightarrow$ *exterior*
$p_1 \wedge p_2 \wedge p_3 \wedge entorno \rightarrow q$	*Hay un adulto* $\wedge$ *Hay un perro* $\wedge$ *Hay un perro* $\wedge$ *exterior* $\rightarrow$*Hay un adulto paseando con dos perros al aire libre.*

En la Figura 5.27, se muestra en una interfaz visual, la descripción semántica de la Figura 5.22. Las descripciones son dadas en diversos niveles de categoría, J1, J2, y una descripción general. También se enlistan los objetos que aparecen en la imagen y finalmente se describe el entorno.

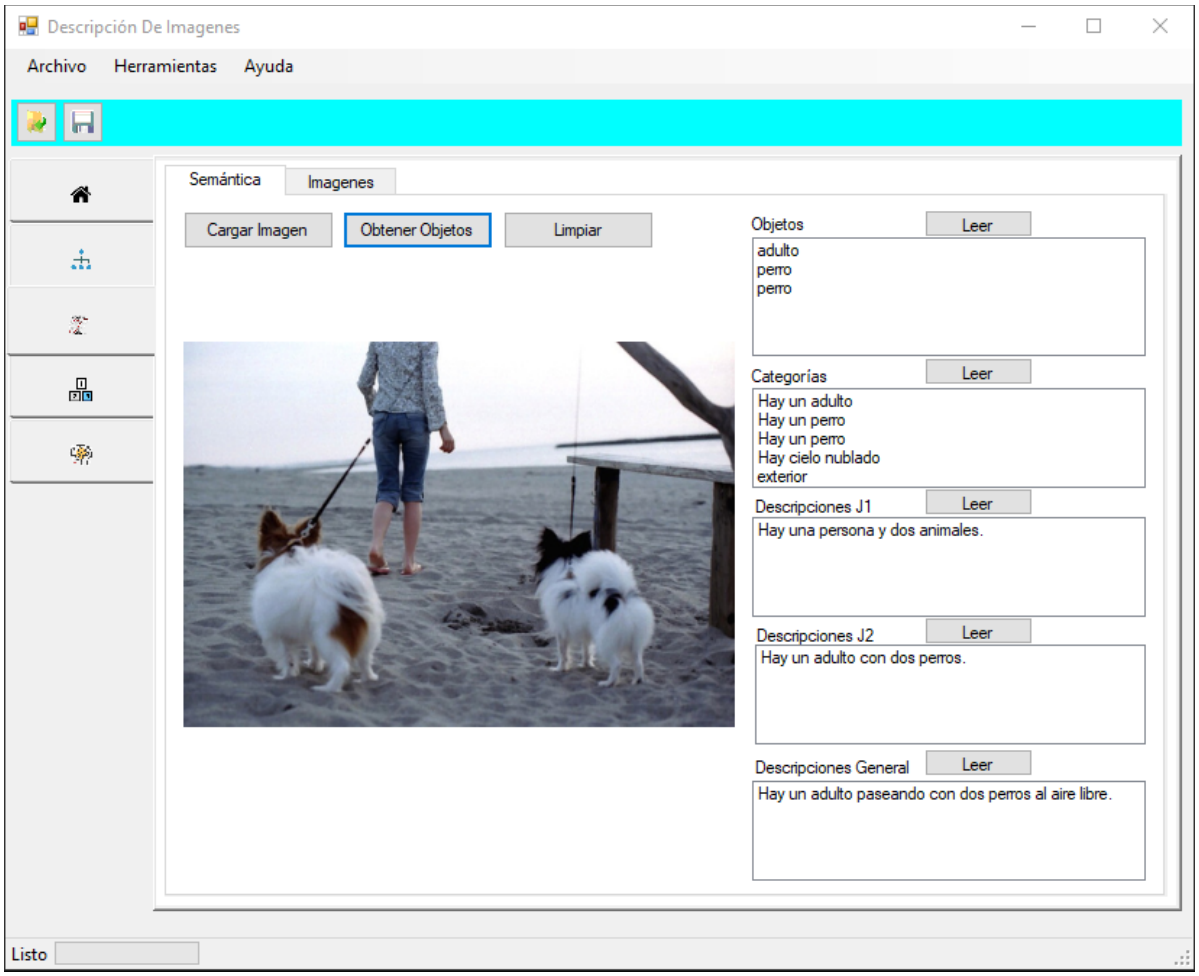


Figura 5.27 Descripción semántica de Figura 5.22

5.2 Resultados del modelo y descripciones de personas en RIDeCS con jerarquías

En esta etapa de experimentación, se presentan imágenes descritas por el modelo propuesto y las mismas imágenes, descritas por humanos en RIDeCS [19]. Para realizar la comparativa de las oraciones se utilizó la métrica *similitud de oraciones basado en fragmentos y contenido* [51], dicha métrica, no solo considera la similitud entre palabras sino el contenido de estas.

En la Tabla 5.2, se muestra la comparativa y similitud entre las descripciones generadas por el modelo para la imagen *a* y algunas respuestas de humanos tomadas de Ridecs.

para cada nivel de jerarquías, así como para la descripción general. Donde J1 es la descripción de un nivel de jerarquía y J2 es la descripción de dos niveles de jerarquía y DG descripción general. Como se puede apreciar para J1 la mayor similitud es de 87.5%, para J2 es de 95%.

Tabla 5.2 Comparativa entre descripciones para la imagen *a*

Imagen a	*Descripción del modelo*	*Descripción humanos*	*Edad*	*Porcentaje de similitud*
	J1: Hay un animal y 3 transportes.	Un auto transportando un animal	14	60.83%
		Unos vehículos y un animal	40	87.75%
		Transportando un elefante	25	82.88%
	J2: Hay un elefante, un camión y 2 carros.	Un carro con un elefante	54	85.50%
		Una camioneta con un elefante arriba	11	95.00%
		Un elefante con un auto	6	87.87%
	DG: Hay un elefante sobre un camión en un camino.	Una camioneta llevando a un elefante en una carretera	10	80.75%
		Un camión transportando un elefante sobre una carretera.	31	81.25%
		Un elefante en una camioneta	58	95.00%

Para la imagen *a,* las descripciones dadas por personas resultaron ser similares, en todos los casos se consideró la existencia de vehículo, aunque en diversos niveles de jerarquía de la categoría transporte, en algunas oraciones se describe coche, auto, camioneta o camión; al utilizar una métrica sensible a la similitud semántica el porcentaje de coincidencia aumenta.

En la Tabla 5.3, se muestra la comparativa y similitud entre las descripciones generadas por el modelo para la imagen *b,* y algunas respuestas de humanos. Como se puede apreciar en la Tabla 5.3, el índice de coincidencia fue mayor, en todos los casos superó el 85% de similitud con respecto a las descripciones dadas por humanos, esto, debido a que en la mayoría de los casos las oraciones se enfocan principalmente en dos elementos: el gato y la silla. En el caso de la imagen *b* la mayor similitud se tuvo con personas de 7, 37 y 6 años con una similitud de 95%.

Tabla 5.3 Comparativa entre descripciones para la imagen *b*

Imagen b	*Descripción del modelo*	*Descripción humanos*	*Edad*	*Porcentaje de similitud*
	J1: Hay un animal y 2 objetos.	Hay un animal	7	95.00%
		Hay una mascota en una silla	12	85.5%
	J2: Hay un gato, una silla y una maceta.	Gato en silla	37	95.00%
		Un gato en un asiento	58	85.5%
		Un gato con una silla	6	95.00%
	DG: Hay un gato sobre una silla y a la izquierda una maceta.	Un gatito en una silla	54	90.18%
		Un gato en un sillón	54	92.50%
		Un gato sobre una silla azul	27	83.25%

Para este caso se utilizaron menos descripciones en J1 Y J2 con respecto a la imagen anterior, debido a que incluso las personas de corta edad proporcionaron descripciones más detalladas, se asume que es debido a la familiaridad que se tiene con ambos elementos protagonistas (gato y silla). El modelo describe oraciones con todos los objetos recibidos, por ello la maceta fue considerada en la descripción generada por el.

En la Tabla 5.4, se muestra la comparativa y similitud entre las descripciones generadas por el modelo para la imagen *c* y algunas respuestas de humanos, de igual manera, en este caso se utilizó menor cantidad de descripciones en J1 Y J2 por la misma causa. El porcentaje de similitud más alto fue de 93.41%

Tabla 5.4 Comparativa entre descripciones para la imagen *c*

Imagen c	*Descripción del modelo*	*Descripción humanos*	*Edad*	*Porcentaje de similitud*
	J1: Hay 2 personas, 2 transportes y un objeto.	2 personas en bicicletas	17	86.45%
	J2: Hay 2 adultos, 2 bicicletas y una bolsa.	Adultos en bici	24	83.60%
	DG: Hay 2 adultos paseando en 2 bicicleta en el campo.	2 adultos en una bicicleta	11	87.87%
		2 hombres en bicicleta	28	82.32%
		2 personas manejando bicicleta	25	93.41%

En la Tabla 5.5, se muestra la comparativa y similitud entre las descripciones generadas por el modelo y algunas respuestas de humanos para la imagen *d.*

Tabla 5.5 Comparativa entre descripciones para la imagen *d*

Imagen d	*Descripción del modelo*	*Descripción humanos*	*Edad*	*Porcentaje de similitud*
	J1: Hay 5 animales	Manada de elefantes	28	85.00%
		Varios animales	40	94.57%
		Manada	32	86.77%
	J2: Hay 5 elefantes	5 elefantes	11	99.30%
		Un grupo de elefantes	30	78.00%
		Elefantes	58	93.60%
	DG: Hay elefantes cerca de un cuerpo de agua	Unos elefantes pasando un río	10	65.00%
		Unos elefantes tomando agua	13	80.43%
		Muchos elefantes en el agua	38	84.87%

En la imagen *d* presentada en la Tabla 5.5, los humanos que describieron en la imagen en la plataforma RIDeCS a excepción de un caso, mencionaron plural para los elefantes con diferentes palabras *manada, "varios, grupo, unos, muchos".* En J1 la mayor similitud es de 94.57%, en J2 99.30% y en DG de 84.87% la diferencia entre estas oraciones es "*cerca"* y "*en".* Pero el plural, elefantes y agua se conserva.

5.3 Resultados del modelo y descripciones generales de personas en RIDeCS

En esta sección, los resultados reportados se enfocan únicamente en las descripciones generales generadas por el modelo y su comparación con descripciones humanas.

En la Tabla 5.6, se describe la imagen *e*, donde se documentan dos tipos de descripciones, la primera corresponde a la descripción proporcionada por el modelo y la segunda a la descripción de seis humanos incluyendo su edad. Adicionalmente se muestra el porcentaje de similitud entre ambas columnas de descripciones.

Tabla 5.6 Comparativa entre descripciones para la imagen *e*

Imagen e	*Descripción del modelo*	*Descripción humanos*	*Porcentaje de similitud*
	DG: Hay personas trabajando en una oficina	Grupo de personas laborando en oficina	89.00%
		Hay un grupo de personas en una oficina de trabajo	95.41%
		Personas trabajando	95.00%
		Personas trabajando en una oficina	99.30%
		Algunas personas trabajando	85.75%
		Trabajo, oficina	81.10%

Para la imagen *e,* la descripción general del modelo es: "*Hay personas trabajando en una oficina".* La descripción con mayor similitud de las proporcionadas por personas que participaron en RIDeCS describiendo fue de 99.30%. Analizando las respuestas de esta imagen los participantes entendieron la escena como *trabajo,* independientemente de las palabras para describir la situación de la escena.

Si se tratara de una escuela el escenario sería muy similar, personas, bancas y sillas, pero en este caso no hay ni un profesor al frente dando la clase. Otro escenario posible es una imagen de biblioteca, pero todos los participantes describieron un área de trabajo, se asume que es por la ausencia de libreros, misma razón por la cual la descripción del modelo descartó que fuera una biblioteca.

En la Tabla 5.7, se muestra la imagen *f* con las descripciones del modelo, diez descripciones proporcionadas por humanos y la edad, en este caso se consideró agregar la edad de los participantes a la tabla debido a las variaciones en las descripciones.

Tabla 5.7 Comparativa entre descripciones para la imagen *f*

Imagen f	*Descripción del modelo*	*Descripción humanos*	*Edad*	*Porcentaje de similitud*
	DG: Hay un adulto en un yate	Adulto paseando en un bote	19	93.15%
		Bote en el agua	15	84.33%
		Un barco en el agua con un señor	7	89.93%
		Un barco en el agua y un joven	55	89.95%
		Un barco	42	88.30%
		Un barco en el agua	38	85.31%
		Un yate	58	83.10%
		Un señor dando un paseo en lancha	25	88.66%
		Un barco y un adulto	19	92.52%
		Un bote con un señor	14	94.00%

En la tabla anterior, se muestra una imagen donde una persona está sobre un barco pequeño, bote o yate, en las descripciones proporcionadas por los participantes se puede notar que en varios casos se enfocaron únicamente en él, seis participantes sí observaron a la persona, en este caso usaron diferentes palabras para describirlo *"señor, adulto, persona, joven"* para referirse a la persona que se encuentra en la escena, esto, debido a su perspectiva de la escena con base a lo que conocen o como le conocen, por ejemplo los de menor edad lo ven como señor o adulto, mientras que, los de mayor edad como persona o joven.

En la Tabla 5.8, se muestra la imagen *g,* la descripción proporcionada por el modelo y diez descripciones de humanos las descripciones de humanos mencionan una persona montando, sobre caballo, montando caballo. El mayor porcentaje de similitud fue de 98.85% con la oración "*Un señor montando un caballo".*

Tabla 5.8 Comparativa entre descripciones para la imagen *g*

Imagen g	*Descripción del modelo*	*Descripción humanos*	*Porcentaje de similitud*
	DG: Hay una persona montando un caballo	Persona montando	75.24%
		Persona montando a caballo en la calle	79.50%
		Un jinete que van en el caballo	85.15%
		Un señor en un caballo	75.24%
		Un señor montando un caballo	98.85%
		Una persona sobre caballo	77.22%
		Un muchacho cruzando la calle con su caballo	50.56%
		Un caballo con su jinete	72.96%
		Montar a caballo	76.00%
		Un señor montando un caballo pinto en un pequeño pueblo con terracería, y con casas un poco antiguas	77.54%

5.4 Resultados del modelo con otro trabajo

En esta sección se realiza una comparativa. En la Tabla 5.9, se muestra la descripción para la imagen *h* con el trabajo de Karpathy & Fei [77] el cual consiste en descripciones semánticas de imágenes mediante anotaciones y algoritmos computacionales, descripciones de RIDeCS, las cuales son proporcionadas por personas [19] y el modelo propuesto en este trabajo. Cabe mencionar que este trabajo es de los pocos que muestran las descripciones generadas, la mayoría únicamente se enfocan en medir el conjunto de información del banco de datos, lo que dificulta una comparativa cualitativa. Ambos trabajos se enfocan en descripciones semánticas, sin embargo se diferencian en que Karpathy & Fei adicionalmente realizan clasificación.

Tabla 5.9 Comparativa entre descripciones del modelo propuesto y Karpathy & Fei [77]

Comparativa Imagen h		*Porcentaje de similitud*
Este trabajo	**1** Hay 2 niños montando una jirafa en un parque	---
Karpathy & Fei [77]	**2** Una jirafa de pie junto a un árbol en un campo	---
Humanos RIDeCS	**D1** Una niña jugando con una jirafa	---
	D2 Unos niños montando en una jirafa	---
	D3 niñas en el zoológico montadas en una jirafa	---
	D4 Unos niños montando una jirafa	---
	D5 Varios niños jugando con una jirafa	---
	D6 4 niños en el pasto con un animal	---
	D7 Niños montando una jirafa	---
	D8 Niñas arriba de una jirafa	---
	D9 2 niñas y 2 niños en una jirafa	---
	D10 Jirafa	---
Este trabajo vs Humanos RIDeCS	1 vs D1	78.65%
	1 vs D2	93.73%
	1 vs D3	83.36%
	1 vs D4	87.75%
	1 vs D5	82.33%
	1 vs D6	80.75%
	1 vs D7	91.83%
	1 vs D8	90.18%
	1 vs D9	80.71%
	1 vs D10	50.00%
Karpathy & Fei [77] vs Humanos RIDeCS	2 vs D1	47.12%
	2 vs D2	48.75%
	2 vs D3	68.25%
	2 vs D4	66.00%
	2 vs D5	58.50%
	2 vs D6	68.25%
	2 vs D7	64.38%
	2 vs D8	76.00%
	2 vs D9	71.50%
	2 vs D10	60.00%
Este trabajo vs Karpathy & Fei [77]	1 vs 2	72.42%

En la Tabla 5.10, se muestra la descripción para la imagen *i* con el trabajo de Karpathy & Fei [77] descripciones de RIDeCS [19] y el modelo propuesto en este trabajo.

Tabla 5.10 Comparativa entre descripciones del modelo propuesto y Karpathy & Fei [77]

Comparativa Imagen i		*Porcentaje de similitud*
Este trabajo	**1** Hay una niña comiendo pastel	---
Karpathy & Fei [77]	**2** La niña está comiendo un pedazo de pastel	---
Humanos RIDeCS	**D1** Una niña pequeña comiendo pastel con las mano	---
	D2 Una niña comiendo un trozo de pastel	---
	D3 Niña comiendo pastel	---
	D4 Bebé comiendo pastel	---
	D5 Una niña comiendo	---
	D6 Una niña con la mano en la boca	---
	D7 Un bebé chupándose la mano con comida	---
	D8 Niña festejando con pastel	---
	D9 Fiesta de la niña con pastel	---
	D10 Niña y pastel	---
Este trabajo vs Humanos RIDeCS	1 vs D1	87.50%
	1 vs D2	97.50%
	1 vs D3	97.50%
	1 vs D4	91.95%
	1 vs D5	78.00%
	1 vs D6	71.50%
	1 vs D7	84.25%
	1 vs D8	76.00%
	1 vs D9	87.75%
	1 vs D10	86.12%
Karpathy & Fei [77] vs Humanos RIDeCS	2 vs D1	87.50%
	2 vs D2	99.20%
	2 vs D3	91.25%
	2 vs D4	86.53%
	2 vs D5	78.00%
	2 vs D6	71.50%
	2 vs D7	84.25%
	2 vs D8	71.50%
	2 vs D9	87.75%
	2 vs D10	86.12%
Este trabajo vs Karpathy & Fei [77]	1 vs 2	94.95%

En la Tabla 5.9, se muestra la información de descripciones para la imagen *h,* la cual corresponde a la escena donde aparecen varios niños y una jirafa como protagonistas. La descripción marcada con el número 1 *"Hay 2 niños montando una jirafa en un parque"*, corresponde a la proporcionada por el trabajo propuesto. La oración marcada con el número 2 "*una jirafa de pie junto a un árbol en un campo"* corresponde a la descripción de los autores Karpathy & Fei y de D1 a D10 corresponde a las descripciones de los participantes de la plataforma RIDeCS. En el caso de esta investigación y las descripciones en Ridecs para la imagen *h* tuvo una similitud promedio de 81.93% y para el caso del trabajo de Karpathy & Fei comparado con las respuestas en Ridecs para la imagen *h* tuvo una similitud promedio de 62.88%

Como se puede apreciar ambas descripciones son correctas, la diferencia es que cada uno de los resultados se enfocó en diversos objetos al momento de describir la escena. Para la imagen *h,* ninguno de los participantes considero al árbol en su descripción, pero ello no significa que éste no exista. En todos los casos se compara la descripción proporcionada por ambos trabajos con las respuestas obtenidas en RIDeCS debido a que son descripciones realizadas por humanos reales de diferentes rangos de edad y las respuestas de estos están relacionadas con el conocimiento adquirido a su edad. Se logró apreciar la diferencia entre descripciones para una misma imagen dada por personas de diferentes edades.

En la Tabla 5.10, Se muestra la información de las descripciones para la imagen *i,* la cual corresponde a la escena donde aparece una niña de entre 2 a 3 años comiendo pastel, ambos trabajos describen a una niña comiendo pastel. Las 10 descripciones de los participantes concuerdan con dichas descripciones a excepción de dos, que describen un festejo. En el caso de esta investigación y las descripciones en RIDeCS para la imagen *i* tuvo una similitud promedio de 85.81% y para el caso del trabajo de Karpathy & Fei comparado con las respuestas en RIDeCS para la imagen *h* tuvo una similitud promedio de 84.36%

En la Figura 5.28, se muestra una gráfica comparativa de las descripciones de la imagen *h* correspondiente a la Tabla 5.9, La línea roja grafica el porcentaje de similitud entre las respuestas del trabajo de Karpathy & Fei con respecto a las descripciones humanas de RIDeCS y la línea azul corresponde a este trabajo.

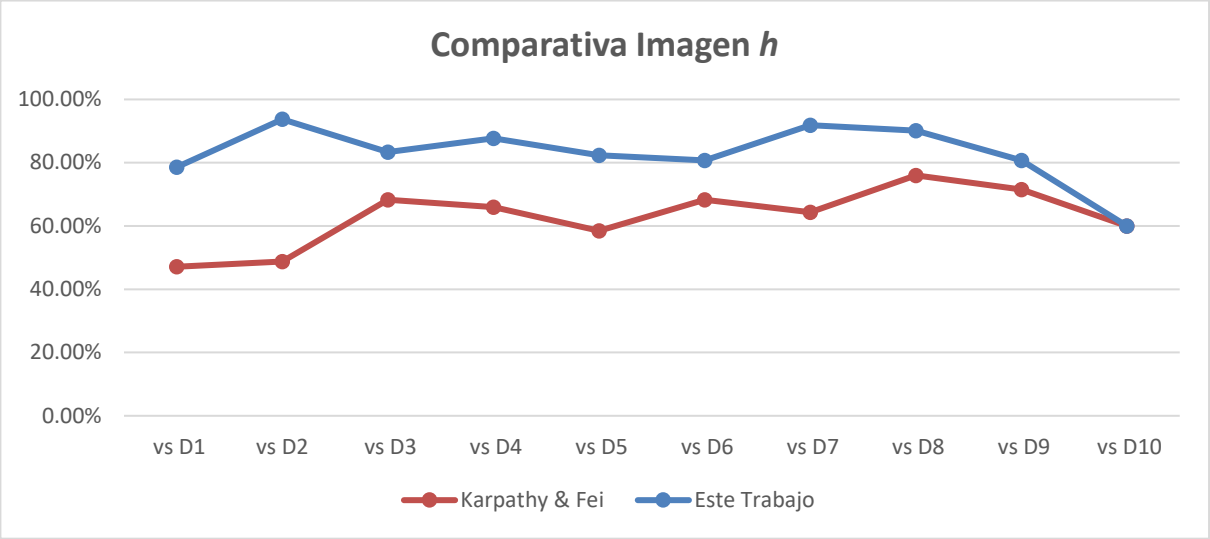


Figura 5.28 Gráfica comparativa de las descripciones de la Tabla 5.9

En la Figura 5.29, se muestra una gráfica comparativa de las descripciones de la imagen *h* correspondiente a la Tabla 5.10, La línea roja grafica el porcentaje de similitud entre las respuestas del trabajo de Karpathy & Fei con respecto a las descripciones humanas de RIDeCS y la línea azul corresponde a este trabajo.

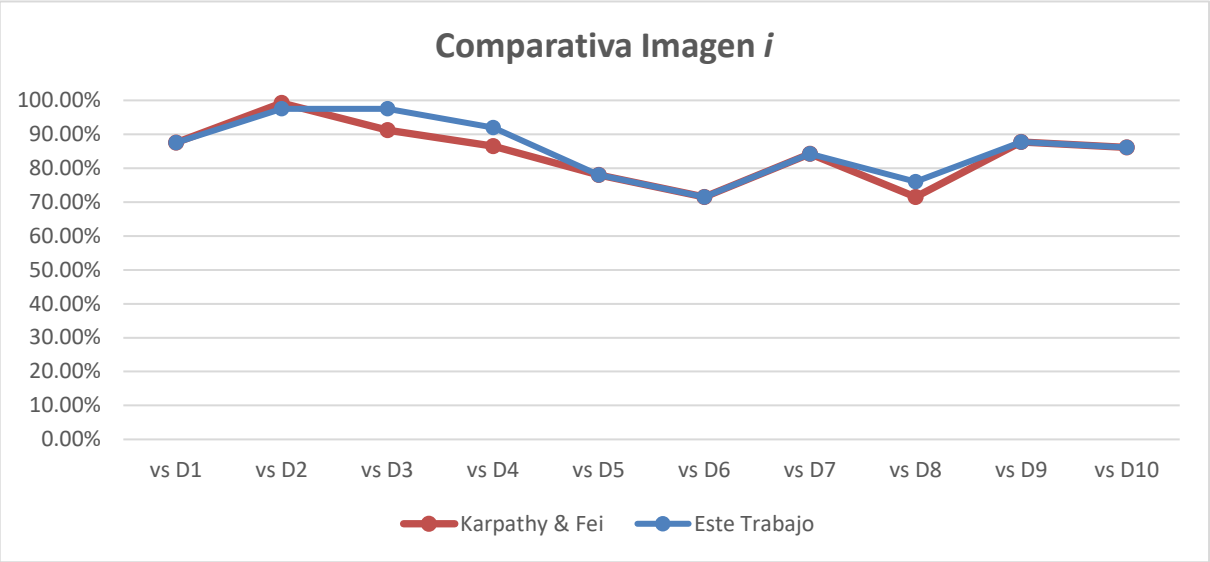


Figura 5.29 Gráfica comparativa de las descripciones de la Tabla 5.10

5.5 Análisis de resultados

Se detectó que existen trabajos con similitud en los elementos, por ejemplo, las distribuciones, en otros casos consideran las jerarquías, pero desde un enfoque diferente mediante la asignación de la jerarquía con respecto a las oclusiones de los objetos, sin embargo, no existe un modelo de representación del conocimiento que dicte el proceso a seguir para la descripción del contenido semántico en imágenes. En este trabajo, se propuso un modelo para ello, el cual ha mostrado que es útil para describir el contenido semántico de las imágenes de manera detallada mediante la medición y comparación con las respuestas de humanos y otros trabajos. Promediando la edad de los participantes se concluyó que el modelo describe similar a un niño de alrededor de 7 a 14 años cuando se consideran las jerarquías de nivel 1 y 2 de clase y para las descripciones generales que conllevan un nivel de especificación mayor fue similar al de personas con una edad de 38 a 42 años. Tras revisar las descripciones de los participantes detenidamente se pudo notar las variaciones en las oraciones sobre una misma escena, es muy interesante observar que para una misma imagen puede haber decenas de descripciones diferentes y esto dependerá de los conocimientos, intereses, gustos, regiones y edades de las personas, entre otros factores; dicho conocimiento es el que se modela con este trabajo para la descripción del contenido semántico en imágenes. Como modelo tiene la fortaleza de ser flexible para ser utilizado con cualquier conjunto de clases, si esta investigación fuera adaptada para teléfonos móviles podría ser una gran herramienta de apoyo para invidentes, incluso para conductores.

5.6 Conclusiones de la tesis

En las siguientes secciones se abordarán los últimos detalles de esta investigación, un breve repaso por los objetivos planteados y logrados, el trabajo resultante extra a lo propuesto inicialmente, así como el trabajo futuro y posibles áreas de oportunidad para este trabajo tanto la industria como en la vida diaria. En la sección 5.7, se presentan los objetivos planteados y si fueron logrados, posteriormente se muestran las conclusiones generales de esta tesis en la sección 5.8 y finalmente el trabajo futuro en

la sección 5.9, el cual no es parte de este trabajo, pero son áreas de oportunidad para aplicaciones y mejoras.

5.7 Objetivos cumplidos

En la siguiente sección se muestra el objetivo general propuesto al inicio de esta investigación y los objetivos específicos, se marcan con el símbolo ✓ los que ha sido completados al 100%.

5.7.1 Objetivo general

Proponer e implementar un modelo de descripción semántica del contenido en imágenes. ✓

5.7.2 Objetivos específicos

- Proponer e implementar un elemento para la categorización de objetos. ✓
- Proponer e implementar un elemento basado en distribuciones espaciales de los elementos que conforman una imagen. ✓
- Proponer e implementar un elemento basado en jerarquías para los elementos que conforman una imagen y que sea capaz de describir información con diferentes niveles de detalle. ✓
- Proponer e implementar un elemento de relaciones para que los objetos de una imagen puedan ser relacionados entre sí. ✓
- Proponer e implementar un elemento de inferencias para que las descripciones puedan ser realizadas de manera natural usando verbos en lugar de posiciones mediante el análisis de categorías, distribuciones, jerarquías y relaciones. ✓

5.8 Conclusiones de los logros

En este trabajo, se realizó un modelo de descripción semántica de imágenes debido a la ausencia de uno que dictara el proceso para realizar dicha tarea. Se construyó considerando sinónimos, esto debido a la variabilidad del lenguaje para referirse a un objeto o una situación. Dentro de la literatura de área se encontraron similitudes con este trabajo sin embargo carecían de un enfoque meramente del área semántica, todos los trabajos en su mayoría involucraban clasificación, segmentación etc. Algunos que eran únicamente de semántica eran muy específicos a una temática, lo que reducía el área de aplicación. Se espera que con la creación de este modelo se pueda crear la pauta que de inicio a trabajos de descripciones semánticas exclusivamente y siguiendo las bases del modelo propuesto, las cuales consideran jerarquías de clase, niveles de prioridad con base en los elementos en la escena e inferencias sobre las relaciones entre los objetos de una misma imagen. Al realizar esta investigación, se observó que a pesar de ser el español uno de los idiomas más hablados a nivel mundial hay pocos trabajos del área en dicho idioma, por ello fue necesario crear RIDeCS, el cual es un repositorio de representación del conocimiento y contiene cientos de descripciones de personas de diversos rangos de edad, quedará disponible en el sitio web www.ridecs.com para futuros trabajos e investigaciones.

5.9 Trabajo futuro

El trabajo futuro es posible dividirlo en dos vertientes, la primera en mejoras y extensiones al modelo y el segundo, aplicaciones del modelo.

Para el primer caso, este trabajo podría convertirse en multiplataforma y ser usado en teléfonos inteligentes. Otro aspecto que podría cubrirse en el futuro es un elemento de consultas para navegar entre imágenes mediante palabras o frases en lugar de realizar búsquedas manuales. También sería posible generar retroalimentaciones automáticas.

Para el segundo caso, las aplicaciones pueden ser infinitas desde el ámbito automotriz para describir en mapa la escena y sea más sencillo conducir sin copiloto, lectura de mensajes entrantes considerando descripciones de las fotos o imágenes recibidas,

también en los asistentes virtuales mejorando la lectura de documentos incluyendo las descripciones de las imágenes que vayan apareciendo y no únicamente de lectura de texto, también en video vigilancia describir automáticamente las escenas verbalmente haciendo uso de un sintetizador de voz, y finalmente un enfoque muy solicitado, como herramienta de apoyo a personas invidentes.

REFERENCIAS

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

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

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

[4]. J. M. Alvarez, M. Salzmann & N. Barnes, “Large-scale semantic co-labeling of image sets”, IEEE Winter Conference on Applications of Computer Vision, 1, pp. 501-508, doi: 10.1109/WACV.2014.6836060, 2014.

[5] H. Zhao, J. Shi, X. Qi, X. Wang & J. Jia, “Pyramid Scene Parsing Network”. IEEE Conference on Computer Vision and Pattern Recognition, 6230-6239, 2017, doi:10.1109/CVPR.2017.660.

[6] S. Pinker, F. Meler, Cómo funciona la mente. Ediciones Destino, Barcelona, ISBN 8423341178, 2000.

[7] R. Penrose, La nueva mente del emperador. Random House Mondadori, Barcelona, ISBN 10: 8439717865, 1991.

[8] K. L. V. Bertalanffy, J. Bertoglio, Teoría general de sistemas. Fondo de Cultura Económica, Madrid, 1987.

[9] M. R. Rosenzweig & A. I. Leiman. Psicología Fisiológica. McGRAW-HILL, Madrid, ISBN 10: 8476159277, 1992.

[10] N. M. Cruz, V. M. Pérez, C. T. Cantero, “Influencia de la motivación intrínseca y extrínseca sobre la transmisión de conocimiento. El caso de una organización sin fines de lucro”, CIRIEC-España, Revista de Economía Pública, Social y Cooperativa, núm. 66, pp. 187-211, 2009.

[11] J. G. Cillán, M. M. García, Atributos extrínsecos del producto: las señales de calidad ISBN 978-84-7595-244-4, (1998).

[12] J. A. Martínez, “La recuperación automatizada de imágenes: retos y soluciones”, 2013, DOI: 10.5209/rev_RGID. 2013.v23.n2.43137.

[13] R. Echeverria, Ontología del Lenguaje, N° de Inscripción: 67559 I.S.B.N.: 956-7802-33-5, 1994.

[14] R. Arillo, J. M. González, “La recuperación de la imagen fija. Perspectiva funcional de los sistemas automatizados de recuperación de imágenes”, en El análisis documental de la fotografía de prensa en entornos automatizados, Universidad Carlos III, Madrid, pp. 265-310, 2002.

[15] J. F. Serrano, “Recuperación de imágenes mediante rasgos descriptores globales y locales”, tesis para obtener el grado de Doctor en Ciencias de la Computación, 2011.

[16] A. Gómez, “Modelo de diseño orientado a marcos para sistemas basados en el conocimiento”, Universidad Politécnica de Madrid, tesis doctoral en informática, (1993).

[17] P. Young, A. Lai, M. Hodosh, J. Hockenmaier, “From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions”, TACL. 2. 67-78., 2014, doi: 10.1162/tacl_a_00166

[18] O. Ducrot, “La elección de las descripciones en semántica argumentativa léxica”, Revista iberoamericana de discurso y sociedad, ISSN 1575-0663, Vol. 2, Nº. 4, pp. 23-44, 2000.

[19] C. A. Vazquez, E, Pinto, “Repositorio de Imágenes para Descripciones de Contenido Semántico”, 2020. [En línea]. Disponible: https://www.ridecs.com/.

[20] C. Pérez, “Recuperación Automatizada de Imágenes mediante la Implementación de Descriptores del Estándar MPEG-7”, 2014.

[21] P. Troncoso, “Indexado y Recuperación de Imágenes por Contenido”, Maestría en Ciencias en Ciencias Computacionales, Sep 2007.

[22] F. Jing, M. Li, L. Zhang, H.-J. Zhang, B. Zhang, “Learning in region- based image retrieval”, Proceedings of the International Conference on Image and Video Retrieval (CIVR2003), pp. 206–215, 2003.

[23] A. Semenov, Las tecnologías de la información y la comunicación en la enseñanza, Instituto de Educación Abierta de Moscú (Federación Rusa), Montevideo, Uruguay. ISBN 9974-32-414-9, 2005.

[24] R. Quillian, “Semantic memory” in Semantic Information Processing, M. Minsky (Ed), MIT Press, Cambridge, Mass, 1968.

[25] B. Raphael, “A Computer Program for Semantic Information Retrieval” in Semantic Information Processing, M. Minsky (Ed), MIT Press, Cambridge, Mass, 1968.

[26] S. C. Shapiro, & G. H. Woddmansee, "A Net Structured Based Relational Question-Answerer", Proceedings International Joint Conference on AI. Washington DC, 325-346, 1971.

[27] M. Minsky, "A Framework for Representing Knowledge", en P. Winston (ed.) The Psychology of Computer Vision. New York: McGraw-Hill. 211-277, 1975.

[28] J. Sowa, Conceptual Structures: Information Processing in Mind and Machine. Reading, Mass: Addison-Wesley. 1984.

[29] D. M. Ramík, K. Madani, C. Sabourin, “From visual patterns to semantic description: A cognitive approach using artificial curiosity as the foundation”, Pattern Recognition Letters, vol. 34, no. 14, pp. 1577–1588, 2013. doi: 10.1016/j.patrec.2013.05.014.

[30] F. Arvidsson, A. Flycht-Eriksson, “Ontologies I”. [En línea]. Disponible: https://gndec.ac.in/~librarian/Articles/ontology/ontologies1.pdf [Accedido: Sep 2017]

[31] S. Nasiri, G. Zahedi, S. Kuntz, M. Fathi, “Knowledge representation and management based on an ontological CBR system for dementia caregiving”, Neurocomputing, vol 350, 20, pp 181-194, July 2019.

[32] D. Gürdür, A. V. Feljan, J. El-khoury, S. K. Mohalik, E. Fersman, “Knowledge Representation of Cyber-physical Systems for Monitoring Purpose”, Procedia CIRP, vol. 72, pp. 468-473, 2018.

[33] Z. Civelek, M. Lüy, E. Çam, H. Mamur, “A new fuzzy logic proportional controller approach applied to individual pitch angle for wind turbine load mitigation”, Renewable Energy, vol. 111, pp. 708-717, October 2017.

[34] L. Fei-Fei, R. Fergus, P. Perona. Banco de imágenes Caltech 101. [En línea]. Disponible: http://www.vision.caltech.edu/Image_Datasets/Caltech101/ [Accedido: septiembre 2017].

[35] L. J. Li, L. Fei-Fei, “Event Data Set”, [En línea]. Disponible: http://vision.stanford.edu/lijiali/event_dataset/ [Accedido: diciembre 2017]

[36] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba. “Places the Scene Recognition Data Base”. [En línea]. Disponible: http://places.csail.mit.edu [Accedido: febrero 2018]

[37] M. Everingham, L. Van~Gool, C. Williams, J. Winn, A. Zisserman, “Visual Object Classes Challenge 2012 (VOC2012)”. [En línea]. Disponible: http://host.robots.ox.ac.uk/pascal/VOC/ [Accedido: octubre 2017]

[38] D. Dua, C. Graff, “UCI Machine Learning Repository”. [En línea]. Disponible: http://archive.ics.uci.edu/ml/index.php [Accedido: marzo 2018]

[39] Eastman Kodak Company, “Kodak Lossless True Color Image Suite”. [En línea]. Disponible: http://r0k.us/graphics/kodak/ [Accedido: marzo 2018]

[40] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso and A. Torralba. “ADE20K Dataset”. [En línea]. Disponible: http://sceneparsing.csail.mit.edu [Accedido: Abril 2018]

[41] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, “The Cityscapes Dataset”. [En línea]. Disponible: https://www.cityscapes-dataset.com [Accedido: mayo 2018]

[42] I. Krasin, T. Duerig, N. Alldrin, V. Ferrari, S. Abu-El-Haija, A. Kuznetsova, H. Rom, J. Uijlings, S. Popov, S. Kamali, M. Malloci, J. Pont-Tuset, A. Veit, S. Belongie, V. Gomes, A. Gupta, C. Sun, G. Chechik, D. Cai, Z. Feng, D. Narayanan, K. Murphy. “OpenImages: A public dataset for large-scale multi-label and multi-class image classification”. [En línea]. Disponible: https://opensource.google/projects/open-images-dataset [Accedido: 2019]

[43] T. Zesch, O. Levy, I. Gurevych, I. Dagan. “UKP-BIU: Similarity and entailment metrics for student response analysis”, Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013), in Conjunction with the Second Joint Conference on Lexical and Computational Semantics (*SEM 2013), pp 285–289. Association for Computational Linguistics, 2013.

[44] S. D. Chowdhury, U. Bhattacharya, S. K. Parui. “Levenshtein distance metric based holistic handwritten word recognition”, Proceedings of the 4th International Workshop on Multilingual OCR, p 17. ACM, 2013.

[45] B. Balsmeier, G. Fierro, L. Fleming, K. Johnson, A. Kaulagi, G. Li, W. Yeh. “Weekly disambiguations of US patent grants and applications”. 2014.

[46] A. B. Kaufman, E. N. Colbert-White, C. Burgess. “Higher-order semantic structures in an african grey parrot’s vocalizations: evidence from the hyperspace analog to language (hal) model”. Animal cognition, 16(5):789–801, 2013.

[47] M. J. Babcock, V. P. Ta, W. Ickes. “Latent semantic similarity and language style matching in initial dyadic interactions”. Journal of Language and Social Psychology, 33(1): pp 78–88, 2014.

[48] P. A. Chew. “Automated account reconciliation method”, US Patent 8,639,596. January 28 2014.

[49] P. Resnik. “Using information content to evaluate semantic similarity in a taxonomy”. arXiv preprint cmp-lg/9511007, 1995.

[50] D. Ştefănescu, R. Banjade, V. Rus, A Sentence Similarity Method Based on Chunking and Information Content. Gelbukh A. (eds) Computational Linguistics and Intelligent Text Processing. CICLing 2014. Lecture Notes in Computer Science, vol 8403. Springer, Berlín, Heidelberg. 2014. doi.org/10.1007/978-3-642-54906-9_36

[51] D. Lin. “An information-theoretic definition of similarity”, ICML, vol 98, pp. 296–304, 1998.

[52] A. Bordes, X. Glorot, J. Weston, Y. Bengio. “A semantic matching energy function for learning with multi-relational data”, Machine Learning, 94(2): pp. 233–259, 2014.

[53] J. J. Jiang, D. W. Conrath. “Semantic similarity based on corpus statistics and lexical taxonomy”. arXiv preprint cmp-lg/9709008, 1997

[54] X. Zheng, A. P. Cai. “The method of web image annotation classification automatic”, Advanced Materials Research, 889: pp. 1323–1326, 2014.

[55] C. Leacock, M. Chodorow. “Combining local context and wordnet similarity for word sense identification”, WordNet: An electronic lexical database, 49(2): pp. 265–283, 1998.

[56] K. Abdalgader. “Word sense identification improves the measurement of short-text similarity”, The International Conference on Computing Technology and Information Management (ICCTIM2014), pp. 233–243. The Society of Digital Information and Wireless Communication, 2014.

[57] Z. Wu, M. Palmer. “Verbs semantics and lexical selection”, Proceedings of the 32nd annual meeting on Association for Computational Linguistics, pp. 133–138. Association for Computational Linguistics, 1994.

[58] C. Surianarayanan, G. Ganapathy, M. S. Ra-masamy. “A practical approach to enhancement of accuracy of similarity model using wordnet towards semantic service discovery”, Demand-Driven Web Services: Theory, Technologies, and Applications, pp. 245–266, 2014. doi:10.4018/978-1-4666-5884-4.ch011

[59] A. Budanitsky, G. Hirst. “Evaluating wordnet-based measures of lexical semantic relatedness”, Computational Linguistics, 32(1):13–47, 2006.

[60] W. Song, J. Z. Liang, S. C. Park. “Fuzzy control ga with a novel hybrid semantic similarity strategy for text clustering”, Information Sciences, pp. 156–170, 2014.

[61] J. Pérez, M. Merino “Definición de semántica”, [En línea]. Disponible: https://definicion.de/semantica/ [Accedido: marzo 2018]

[62] S.K. Chang, S.H. Liu, “Picture indexing and abstraction techniques for pictorial databases”, IEEE Trans. Pattern Anal. Mach. Intell. 6 (4), pp. 475–483, 1984.

[63] Y. Song, W. Wang, A. Zhang, “Automatic annotation and retrieval of images”, World Wide Web 6 (2), 209–231, 2003.

[64] A. Mojsilovic, B. Rogowitz, “ISee: perceptual features for image library navigation”, Proceedings of the SPIE, Human Vision and Electronic Imaging, vol. 4662, pp. 266–277, 2002.
[65] S.K. Chang, Q.Y. Shi, C.W. Yan, “Iconic indexing by 2D string”, IEEE Trans. Pattern Anal. Mach. Intell. 9 (3), pp. 413–428, 1987.

[66] Y. Liua, D. Zhanga, G. Lua, W. Maba, “A survey of content-based image retrieval with high-level semantics”, Gippsland School of Computing and Information Technology, Monash University, Vic 3842, Australia bMicrosoft Research Asia, No. 49 ZhiChun Road, Beijing, China Received 19 November 2005; received in revised form 20 March 2006; accepted 28 April 2006.

[67] R. Chiou, G. F. Humphreys, J. Jung, M. A. Lambon Ralph, “Controlled semantic cognition relies upon dynamic and flexible interactions between the executive semantic control and hub-and-spoke semantic representation systems”, Cortex, vol. 103, June 2018, pp. 100-116, 2018.

[68] R. Vedantam, C. Lawrence Zitnick, Devi Parikh, “CIDEr: Consensus-based Image Description Evaluation”, Submitted IEEE Xplore, 2014.

[69] A. Marie-Aude & H. Hicham, “Semantic Structuration of Image Annotations: A Data Mining Approach”, pp. 38-47, 2002.

[70] V. Lavrenko, R. Manmatha, J. Jeon, “A Model for Learning the Semantics of Pictures”, Advances in Neural Information Processing Systems 16, 2003.

[71] S. Pandey, P. Khanna, H. Yokota, “An Effective Use of Adaptive Combination of Visual Features to Retrieve Image Semantics from a Hierarchical Image Database”, Journal of Visual Communication and Image Representation. 30. 10.1016/j.jvcir.2015.03.010., 2015.

[72] M. H. Quinn, E. Conser, J. M. Witte, M. Mitchell, “Semantic Image Retrieval via Active Grounding of Visual Situations”, Computer Vision and Pattern Recognition, 2017.

[73] L. Xu, X. Wang, “Semantic Description of Cultural Digital Images: Using a Hierarchical Model and Controlled Vocabulary”, D-Lib Magazine vol 21, n. 5/6, 2015.

[74] D. Lin, S. Fidler, C. Kong, R. Urtasun, “Visual Semantic Search: Retrieving Videos via Complex Textual Queries”, Proceedings of the IEEE Conference on

Computer Vision and Pattern Recognition (CVPR ’14). IEEE Computer Society, USA, pp. 2657–2664, 2014. doi: https://doi.org/10.1109/CVPR.2014.340

[75] O. Vidal Pino, E. R. Nascimento, M. Campos, “Prototypicality effects in global semántica description of objects”, 2019 IEEE Winter Conference on Applications of Computer Vision (WACV)}, Hawái, EE. UU, 2019.

[76] G. Nascimento, C. Laranjeira, V. Braz, A. Lacerda, E. R. Nascimento, “A Robust Indoor Scene Recognition Method based on Sparse Representation”, Springer International Publishing, 22nd Iberoamerican Congress on Pattern Recognition. CIARP, 2017.

[77] A. Karpathy, L. Fei-Fei, “Deep Visual-Semantic Alignments for Generating Image Descriptions”, CVPR 2015 Paper, Department of Computer Science, Stanford University, 2015.

[78] P. Rosado-Rodrígo, E. Figueras-Ferrer, M. Planas-Rosselló, F. Reverter-Comes. “La visión artificial, un nuevo aliado para el análisis de imágenes artísticas”. Arte, Individuo y Sociedad Universidad de Barcelona. 2016.

[79] P. Sinha, R. Russell. “A perceptually based comparison of image similarity metrics”. Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology Massachusetts Avenue, Cambridge, MA 02139, USA. Department of Psychology, Gettysburg College, Gettysburg, PA 17325, USA. Perception, 2011, vol. 40, pp 1269-1281. 2011.

[80] D. D. Thang, S. C. Hidayati, Y. Chen, W. Cheng, S. Sun and K. Hua, "A Spatial-Pyramid Scene Categorization Algorithm based on Locality-aware Sparse Coding," 2016 IEEE Second International Conference on Multimedia Big Data (BigMM), pp. 342-345, 2016. doi: 10.1109/BigMM.2016.93.

[81] M. Marszalek, C. Schmid. “Semantic Hierarchies for Visual Object Recognition”. CVPR-IEEE Conference on Computer Vision & Pattern Recognition, Minneapolis, United States. pp.1-7, Jun 2007. ff10.1109/CVPR.2007.383272ff. ffinria-00548680.

[82] Y. Chen, J. Z. Wang, R. Krovetz. “Content-based image retrieval by clustering”. Proceedings of the 5th ACM SIGMM International Workshop on Multimedia Information Retrieval, MIR 2003. Association for Computing Machinery, Inc, pp. 193-200, 2003.

[83] A, Mojsilović, J. Gomes, B. Rogowitz. “Semantic-Friendly Indexing and Quering of Images Based on the Extraction of the Objective Semantic Cues”. International Journal of Computer Vision vol. 56, pp.79–107 2004. https://doi.org/10.1023/B:VISI.0000004833.39906.33

[84] Convex Optimization for Scene Understanding, Mohamed Souiai, Claudia Nieuwenhuis, Evgeny Strekalovskiy and Daniel Cremers, IEEE International Conference on Computer Vision Workshops, Technical University of Munich, 2013.

[85] A. Delong, L. Gorelick, O. Veksler, Y. Boykov. “Minimizing energies with hierarchical costs”. International Journal of Computer Vision, 2, 2012. doi=10.1.1.229.11.

[86] C. Nieuwenhuis, E. Töppe, D. Cremers. “A Survey and Comparison of Discrete and Continuous Multi-Label Optimization Approaches for the Potts Model”. Int J Comput Vis. 2013. doi:10.1007/s11263-013-0619-y

[87] J. Yao, S. Fidler, R. Urtasun. “Describing the Scene as a Whole: Joint Object Detection, Scene Classification and Semantic Segmentation”. Computer Vision and Pattern Recognition (CVPR), IEEE Conference, 2012.

[88] E.B. Sudderth, A. Torralba, W.T. Freeman, A.S. Willsky. “Learning hierarchical models of scenes, objects, and parts”. IEEE International Conference on Computer Vision. IEEE International Conference on Computer Vision, November 2005.

[89] M. Alberti, J. Folkesson, P. Jensfelt. “Relational Approaches for Joint Object Classification and Scene Similarity Measurement in Indoor Environments”. Qualitative Representations for Robots: Papers from the AAAI Spring Symposium, Stockholm, Sweden, 2014.

[90] J. Redmon, A. Farhadi, “YOLOv3: An Incremental Improvement”, [En línea]. Disponible: https://pjreddie.com/darknet/yolo/ [Accedido: 2018]

[91] J. Nelson, Mastering Redis, Packt Publishing Ltd, ISBN: 978178398819, 2016.

[92] JSON (JavaScript Object Notation) (1999), [En línea]. Disponible: https://www.json.org/json-es.html

[93] FLEX Analizador Léxico (April 2004) [En línea]. Disponible: http://gnuwin32.sourceforge.net/packages/flex.htm

[94] Código Internacional de Nomenclatura Zoológica, Comisión Internacional de Nomenclatura Zoológica, ISBN:84-607-0588-9, 4ª Edición, Madrid, España, 2009.

Anexo A

Tabla A.1 Animales faltantes de la Figura 5.4

Aves	Mamíferos	Mamíferos	artrópodos	Reptiles
urraca	alpaca	gato	Aaaña	Cocodrilo
arrendajo azul	antílope	hámster	artrópodo	Lagartija
avestruz	ardilla	hipopótamo	abeja	Serpiente
pájaro carpintero	armadillo	jaguar	cangrejo	Tortuga
pingüino	ballena	jirafa	cien pies	tortuga de mar
cuervo	caballo	koala	escarabajo	
pollo	cabra	león	escorpión	
águila	camello	leopardo	garrapata	
búho	canguro	lince	Hormiga	
pato	cebra	lobo marino	Insecto	
canario	cerdo	mapache	mariquita	
ganso	chita	mono	oruga	
cisne	ciervo	mula	polilla	
Alcón	conejo	murciélago		
loro	delfín	nutria		
gorrión	elefante	oso		
	erizo	oso café		
	puerco Espín	oso panda		
	ratón	oso polar		
	rinoceronte	oso rojo		
	tigre	oveja		
	toros	perro		
	vaca	zorrillo		

Anexo B

Tabla B.1.1 Objetos faltantes de la Figura 5.5

Ropa	calzado	accesorios	Herramienta
guantes	botas	cinturón	Desatornillador
short	sandalia	lentes de sol	Taladro
minifalda	tacón	cartera	Motosierra
vestido		tiara	llave inglesa
traje de baño		collar	Tijera
pantalón		aretes	escalera de tijera
camisa		sombrero vaquero	Engrapadora
calcetín		sombrero	Trinquete
corbata		Bufanda	Daga
saco		sombrero fedora	Espátula
traje		sombrero sol	Plomería
Falda		bolso de mano	martillo
chaqueta		lentes	quita nieve
		reloj inteligente	Trípode
			Estetoscopio
			hacha
			Cincel

Tabla B.1.2 Objetos faltantes de la Figura 5.5

blancos	armas	Juguetes	obj deportivos
toalla	antorcha	muñeca	arco y flecha
almohada	misil	pelota	banco de entrenamiento
pañuelos	rifle	globo	barra horizontal de gimnasia
sabanas	escopeta	Dado	bate de beisbol
colchas	espada	patines	bici fija
edredón	tanque	disco volador	Binoculares
almohada	arma	patineta	Caminadora
funda de almohada	cohete	patín eléctrico	casco de bicicleta
	pistola	oso de peluche	casco de futbol
	bomba	cubilete	chaleco salvavidas
		monociclo	equipo de bolos
			equipo de senderismo
			Esquí
			googlees de nado
			guante de beisbol
			mesa de billar
			Parachute
			pelota de criquet
			pelota de golf
			pelota de rugby
			pelota de tenis
			Pesa
			raqueta
			raqueta de tenis
			raqueta de tenis de mesa
			saco de boxeo
			Squash
			tabla de nieve
			tabla de pádel
			tabla de surf
			Tablero
			Tienda
			uniformes deportivos
			viga de equilibrio

Tabla B.1.2 Objetos faltantes de la Figura 5.5

cuidado personal	decoración	Variado	Papelería
cepillo dental	cortina	Contenedor	Mochila
jeringa	escultura	carro súper	Maleta
cosméticos	escultura de bronce	placa de coche	Pizarrón
toallas de papel	espejo	Linterna	Sacapuntas
polvo facial	florero	escalera	Borrador
isopo	lampara	poster	cinta adhesiva
gorro de baño	maceta	cañón	argolla
espray de cabello	marco de foto	moneda	estuche lápiz
labial	persiana	caja	cortadora de papel
papel de baño	planta	puerta	contenedor de basura
apósito curativo	reloj	camilla	Sobre
perfume	reloj de pared	bandera	Regla
toallas de papel	vela	cabina de baño	Bolígrafo
pañal	ventana	muleta	Maletín
		caja de plástico	Libro
		canasta de picnic	
		sombrilla	

Tabla A.2.3 Objetos faltantes de la Figura 5.5

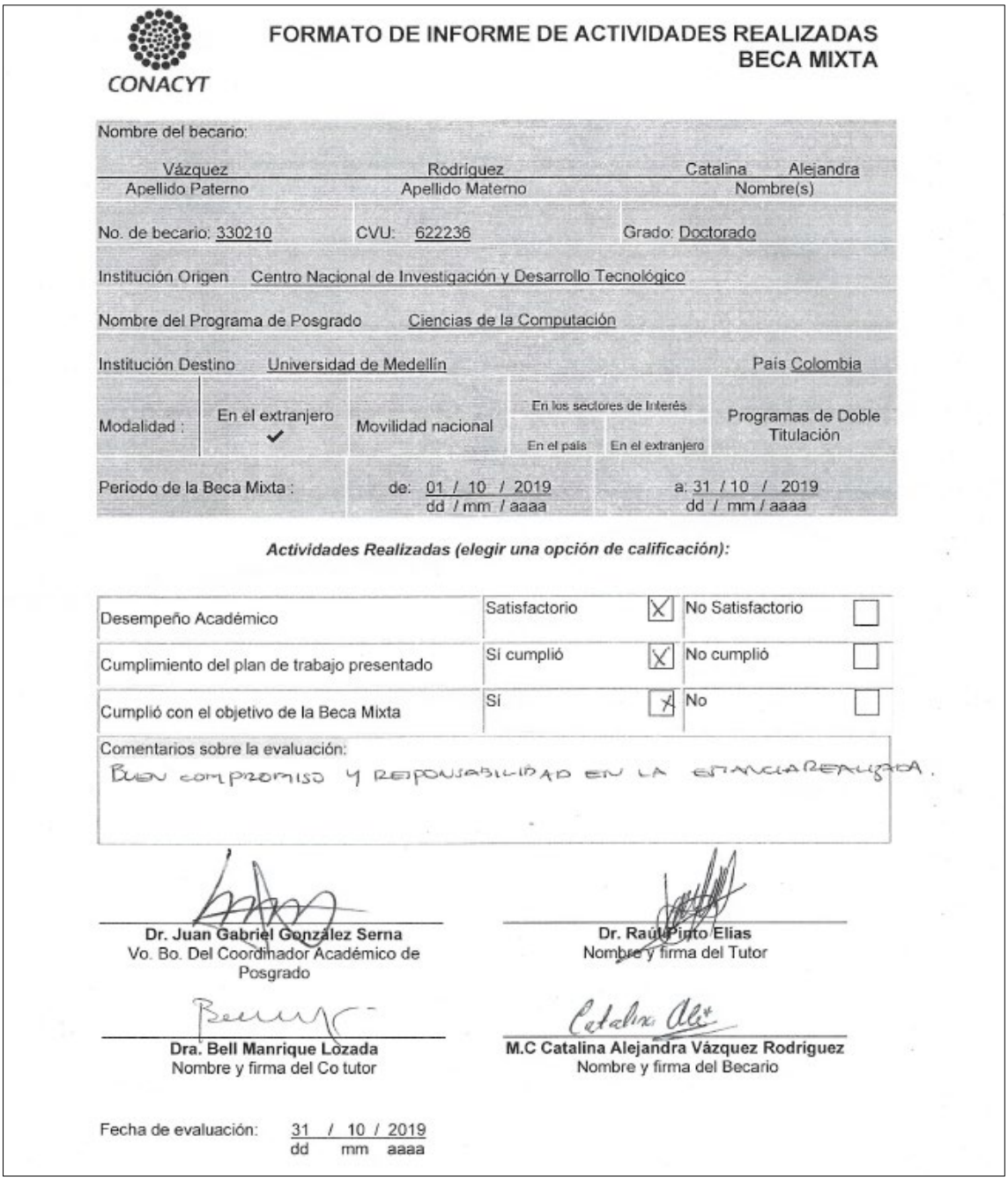
utensilios de cocina	instrumentos musicales	Otros
abrebotellas	acordeón	Auriculares
abrelatas	armónica	Báscula
botella	arpa	Bombilla
cafetera	banjo	Calentador
coctelera	flauta	Cámara
copa de vino	guitarra	Celular
cortador de pizza	oboe	control remoto
cuchara	órgano	dispensador de jabón
cucharón	piano	enchufe toma corriente
cuchillo	saxofón	Humidificador
cuchillo	tambor	interruptor de luz
especiero	teclado musical	iPod
jarra	trombones	Micrófono
lata	trompeta	reloj digital
olla de presión	violín	reproductor de casete
palillos chinos	violonchelo	secadora de manos
pimentero		teléfono de cable
plato		Ventilador
popote		ventilador de techo
salero		
Sartén		
espray cocina		
tabla de picar		
taza de cabe		
taza medidora		
tazón		
tazón para mezclar		
tenedor		
Tetera		
Vajilla		
Wok		

Tabla B.2.4 Objetos faltantes de la Figura 5.5

Hogar	oficina
lavabo	calculadora
accesorios de baño	fax
alacena	impresora
archivero	laptop
banca	monitor pc
banco	ratón de computadora
bañera	tablet
baño	teclado
bidé	
cajón	
cajonera	
cama	
cama de perro	
cama infantil	
closet	
ducha	
encimera	
escritorio	
estante para vinos	
estufa de leña	
gabinete	
grifo	
guardarropa	
jacuzzi	
librero	
mesa	
mesita de café	
mesita de noche	
mueble de gato	
porta pastel	
silla de ruedas	
sofá cama	
sofá de estudio	

Anexo C

En esta sección se muestran las actividades adicionales realizadas durante el periodo doctoral, tales como: participaciones de congreso, estancia, publicaciones, etc.



CONACYT

FORMATO DE INFORME DE ACTIVIDADES REALIZADAS
BECA MIXTA

Nombre del becario:

Vázquez	Rodríguez	Catalina Alejandra
Apellido Paterno	Apellido Materno	Nombre(s)

No. de becario: 330210 CVU: 622236 Grado: Doctorado

Institución Origen: Centro Nacional de Investigación y Desarrollo Tecnológico

Nombre del Programa de Posgrado: Ciencias de la Computación

Institución Destino: Universidad de Medellín País: Colombia

Modalidad :	En el extranjero ✓	Movilidad nacional	En los sectores de Interés		Programas de Doble Titulación
			En el país	En el extranjero	

Periodo de la Beca Mixta : de: 01 / 10 / 2019 (dd / mm / aaaa) a: 31 / 10 / 2019 (dd / mm / aaaa)

Actividades Realizadas (elegir una opción de calificación):

Desempeño Académico	Satisfactorio	X	No Satisfactorio	
Cumplimiento del plan de trabajo presentado	Sí cumplió	X	No cumplió	
Cumplió con el objetivo de la Beca Mixta	Sí	X	No	

Comentarios sobre la evaluación:
BUEN COMPROMISO Y RESPONSABILIDAD EN LA ESTANCIA REALIZADA.

Dr. Juan Gabriel González Serna
Vo. Bo. Del Coordinador Académico de Posgrado

Dr. Raúl Pinto Elías
Nombre y firma del Tutor

Dra. Bell Manrique Lozada
Nombre y firma del Co tutor

M.C Catalina Alejandra Vázquez Rodríguez
Nombre y firma del Becario

Fecha de evaluación: 31 / 10 / 2019 (dd mm aaaa)

Figura C.1 Comprobante de estancia internacional

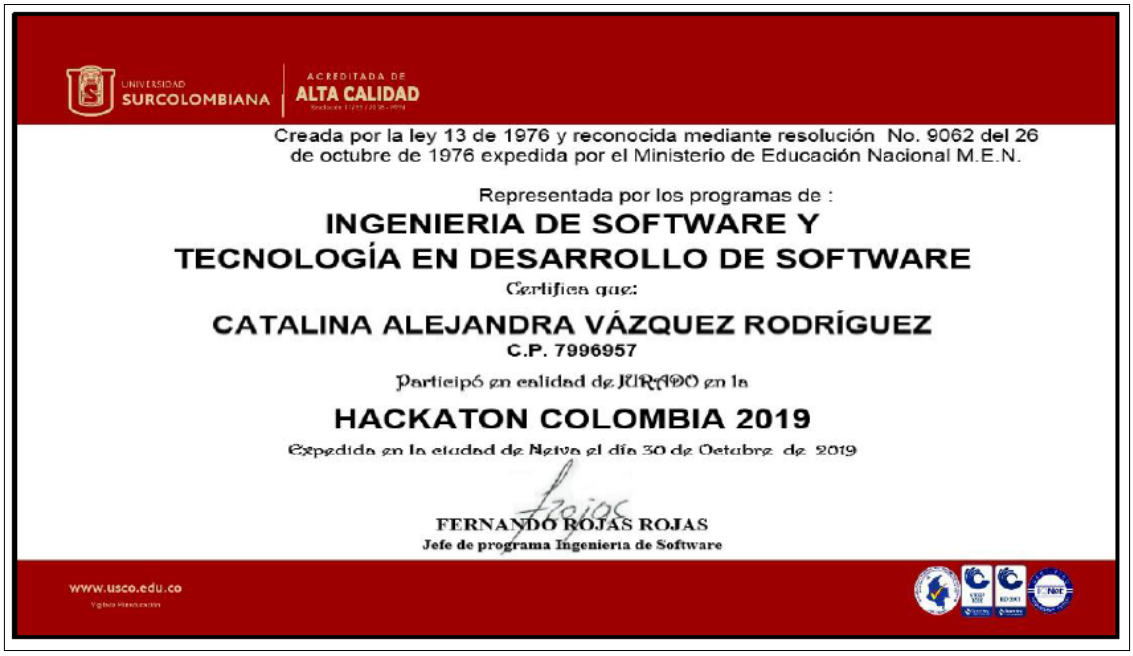


Figura C.2 Comprobante de participación como jurado en hackaton Colombia 2019

Figura C.3 Comprobante de participación como ponente en SICC Colombia 2019

Año: 8, No. 46 Septiembre - Octubre 2020

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.

Figura C.4 Comprobante de publicación articulo Conacyt

risti

CONSTANCIA DE ACEPTACIÓN

En mi calidad de Editor de la Revista RISTI - Revista Ibérica de Sistemas e Tecnologias de Informação Ingenierías Universidad de Medellín, categoría Q3 de Scopus/SCImago Journal Rank, con ISSN 1646-9895.

HAGO CONSTAR:

Que el siguiente artículo fue aceptado para publicación en nuestro número 34 de octubre de 2019, y se encuentra en estado de *pre-print* para publicación a la fecha:

Título:	**La Semántica de las Imágenes y el Análisis de su Contenido.**
Autor (es):	*Catalina-Alejandra Vázquez-Rodríguez, Raúl Pinto-Elías.*

Portugal, 5 de noviembre de 2019

ÁLVARO ROCHA
Director
Consejo Editorial
Departamento de Engenharia Informática
Pólo II - Pinhal de Marrocos, 3030- 290
Coimbra, Portugal.
amrocha@dei.uc.pt

Figura C.5 Comprobante de publicación articulo risti