risti
Revista Ibérica de Sistemas e Tecnologias de Informação
Revista Ibérica de Sistemas y Tecnologías de Información

Recebido/Submission: 05/08/2019
Aceitação/Acceptance: 17/09/2019

La Semántica de las Imágenes y el Análisis de su Contenido

Catalina-Alejandra Vázquez-Rodríguez, Raúl Pinto-Elías

cvazquez@cenidet.edu.mx, rpinto@cenidet.edu.mx

Tecnológico Nacional de México/CENIDET, Depto. Ciencias Computacionales, Interior Internado Palmira S/N, Col. Palmira, C.P. 62490, Cuernavaca, Morelos, México.

DOI: 10.17013/risti.34.20–28

Resumen: Las imágenes transmiten información de diferentes maneras, la descripción puede ser referente a la forma de los objetos, colores, texturas; sin embargo, también pueden transmitir conceptos, los cuales estarán formados por el conjunto de objetos presentes en una imagen. En este trabajo se presenta un enfoque de descripción de imágenes desde la perspectiva semántica y se evalúa mediante un análisis léxico-sintáctico para comprobar que las descripciones dadas sean parte del lenguaje definido, por ello se utiliza un módulo de verificación sintáctica el cual resultó ser de gran utilidad dado que no siempre los datos de entrada pertenecerán al conjunto de trabajo, si bien es un modelo adaptable y con posibilidades de crecimiento, es necesario definir los datos a trabajar; por ello la verificación sintáctica evita añadir "ruido" al modelo para garantizar que las clases que entren sea una clase válida y pueda realizarse un refinamiento de las descripciones.

Palabras-clave: Semántica del contenido de imágenes; estructuración de la información; análisis léxico del contenido de imágenes.

The Semantics of Images and the Analysis of their Content

Abstract: the images transmit information in different ways, the description can be referring to the shape of objects, colors, textures; however, they can also convey concepts, which will be formed by the set of objects present in an image. In this work, an image description approach is presented from the semantic perspective and is evaluated through a lexical-syntactic analysis to verify that the descriptions given are part of the defined language, so a syntactic verification module is used which turned out to be very useful since the input data will not always belong to the work set, although it is an adaptable model and with growth possibilities, it is necessary to define the data to work; Therefore, the syntactic verification avoids "noise" to the model to ensure that the classes that enter are a valid class and a refinement of the descriptions can be made.

Keywords: Coupling; flexibility; interface inheritance; implementation inheritance.

1. Introducción

La descripción de las imágenes es una tarea que ha intentado resolverse computacionalmente por muchos años, dicha tarea muchas veces es confundida con el reconocimiento de objetos (Mottaghi et al., 2014), el cual consiste en identificar los elementos y objetos de las imágenes (Álvarez, Salzmann, & Barnes, 2014), mientras que la descripción de las imágenes va enfocada a describir el contenido de ellas como un todo, es decir, interpretar la escena presente (Xu, Schwing, & Urtasun, 2014).

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

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

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

Al profundizar en el campo de las propiedades de las imágenes, más aumenta la complejidad de la traducción o interpretación de la relación entre los atributos intrínsecos y objetos de una imagen (Cruz, Pérez, & Cantero, 2009). 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.

A continuación, se describe el contenido de este trabajo. En la sección 2 se encuentra la información de los datos con los que se trabajó, en la sección 3 el estado del arte, en la sección 4 la estructura del sistema, en la sección 5 la experimentación y resultados, en la sección 6 las conclusiones, en la sección 7 trabajo futuro, y finalmente las referencias.

2. Datos y clases

En el presente trabajo se realizan descripciones semánticas del contenido de las imágenes. Se utilizó el banco Pascal VOC 2012, el cual es un banco de imágenes muy

popular, para construir y evaluar algoritmos para la clasificación de imágenes, detección de objetos y segmentación. Pascal VOC agrupa sus imágenes de tres formas: el primer grupo corresponde a las imágenes originales, el segundo grupo corresponde a los Ground Truth, y el tercer grupo corresponde a la clasificación de las clases existentes en las imágenes.

Se trabajó con 5 categorías y 26 subcategorías las cuales se muestran en la Tabla 1.

Categorías	Subcategorías
Personas	Adulto, Niño, Bebé
Animales	Gato, Vaca, Perro, Ave, Caballo, Oveja
Transporte	Avión, Bicicleta, Bote, Autobús, Coche, Motocicleta, Tren
Objetos	Botella, Silla, Mesa, Maceta, Sofá, Pantalla
Entorno	Construcciones, Cielo, Pavimento, Vegetación

Tabla 1 – Categorías y subcategorías trabajadas para la descripción semántica y sintáctica del contenido de imágenes.

3. Estado del arte

Para describir imágenes es necesaria una explicación de los elementos de las imágenes. 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.

La 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 suelen denominarse descripciones de bajo nivel (Civelek, Lüy, & Mamur, 2017). Las propiedades extrínsecas de la imagen corresponden a todo lo contenido en la imagen no propiamente visual, estas propiedades están divididas en nivel medio y nivel alto. En el nivel medio se trabaja la detección automática de límites, contornos, objetos como rostros, sillas, coches etc. y de conceptos extraídos de la imagen como la identificación de día o de noche, si es verano o invierno.

En el nivel alto se trabajan los elementos que se incluyen en los metadatos como autor, título, localización geográfica, 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 (Iancu, 2018).

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, pues estos atributos son inherentes a la imagen. El problema se complica cuando se aborda la representación automática de las propiedades extrínsecas, describir el concepto de frío,

rostro, mascota 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, al profundizar 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 (Xu et al., 2014).

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 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 mediante técnicas de retroalimentación entre otras para mejorar la respuesta del sistema (García & Cillán, 1998).

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 consideraremos como el punto de partida de los llamados modelos de razonamiento. En cualquier dominio, la propagación del conocimiento (Xu et al., 2014), por medio de programas de Inteligencia artificial 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, si se habla de cualquier otro tipo de sistemas de IA, y contribuyen de manera decisiva a organizar correctamente la búsqueda de soluciones.

Pueden plantearse diversas maneras de representar el conocimiento para posteriormente lograr descripciones semánticas, mediante, ontologías, lógica de primer orden, lógica difusa etc. (Ducrot, 2000). La representación del conocimiento y el razonamiento juegan un papel central en la Inteligencia Artificial. La investigación en IA comenzó tratando de identificar los mecanismos generales responsables del comportamiento inteligente, rápidamente se hizo evidente que los métodos generales y poderosos no son suficientes para obtener el resultado deseado, es decir, el comportamiento inteligente (Martínez, 2013).

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 (Álvarez et al., 2014).

El tema de las descripciones semánticas se ha abordado mediante diversos enfoques a lo largo de los años, sin embargo, sigue siendo un problema sin resolver por

completo, es sencillo describir los elementos que conforman una imagen, pero interpretar la escena dados los elementos es una tarea más compleja (Pérez & Merino 2018).

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

Figura 1 – Diagrama del proceso de descripción y verificación semántica.

4. Estructura del sistema

Cuando entran los datos pertenecientes a los Ground Truth de Pascal VOC se realizan 4 pasos. En el primer paso se extraen los nombres de las clases presentes en la imagen, en el segundo se realiza una verificación sintáctica de las clases para evitar añadir ruido al sistema, en el tercero se enumeran la cantidad de objetos detectados de cada clase, en el cuarto se calculan las coordenadas x,y de cada objeto y finalmente se obtienen descripciones semánticas de la escena de la imagen y algunas inferencias. En la Figura 2 se muestra el diagrama del proceso de descripción semántica de imágenes.

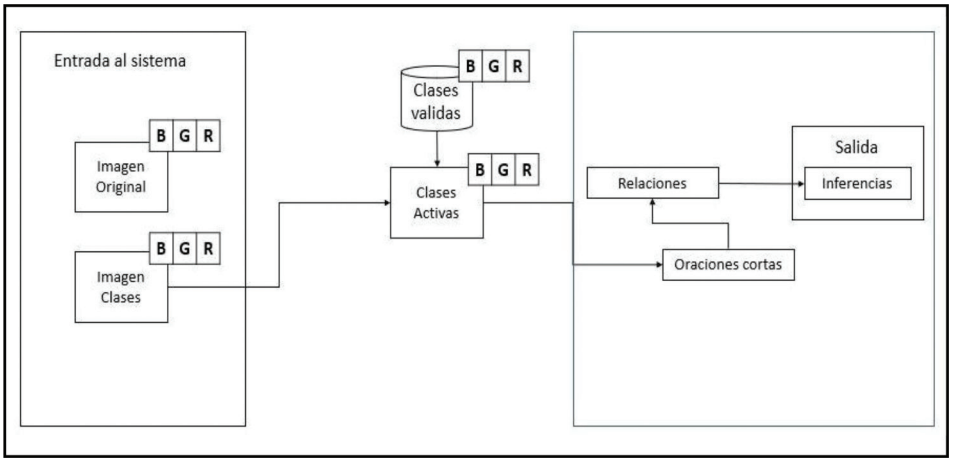
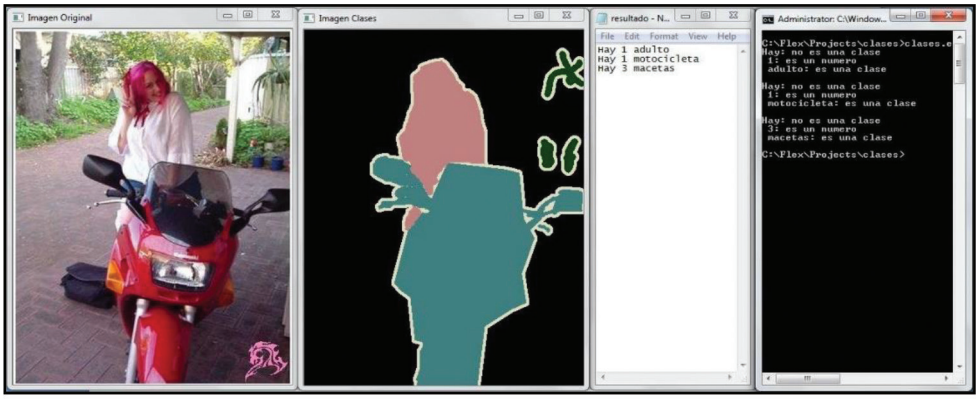


Figura 2 – Diagrama del proceso de descripción y verificación semántica.

5. Experimentación y resultados

La experimentación se llevó a cabo con imágenes del banco Pascal VOC 2012, se utilizaron los originales y los Ground Truth, los cuales muestran a detalle donde se ubican los elementos descritos de las imágenes, por lo cual se puede obtener la ubicación de cada objeto con precisión y de esta manera saber si un objeto se encuentra, sobre, al lado, debajo etc. lo cual apoya a la descripción y a las inferencias que pudieran realizarse sobre la escena de la imagen. En este trabajo las descripciones fueron variantes en el aspecto de algunas más detalladas que otras, pues la descripción depende directamente de la cantidad de objetos contenidos en las imágenes. En la Figura 3a se muestra la imagen original, en la Figura 3b el Ground Truth de la original, en la Figura 3c la descripción de su contenido y finalmente, en la Figura 3d el análisis sintáctico del contenido mediante Lex-Bison para verificar que las clases son las definidas; así como su correcta escritura para evitar errores ortográficos, de singular-plural, etc.



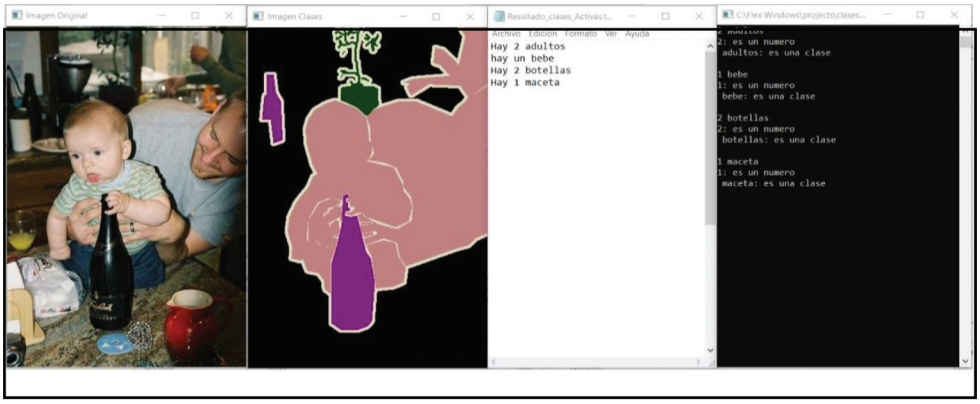
a) b) c) d)

Figura 3 – Imagen de adulto en moto, su *Ground Truth*, la descripción semántica y verificación sintáctica.

Como se puede apreciar en la Figura 3a se tiene una persona adulta, una moto y vegetación. En la Figura 3b que corresponde al Ground Truth es posible ver fácilmente de color verde los objetos "macetas" contenidos en la imagen original, posteriormente en la Figura 3c, el algoritmo de descripción de contenido semántico de imágenes realizó su trabajo enlistando los elementos y la cantidad de objetos que están dentro del conjunto de clases con las que se está trabajando. Finalmente, en la Figura 3d se realiza el análisis sintáctico de las sentencias descriptivas dadas por el sistema de descripción semántica de imágenes, donde verifica si cada palabra es parte de la gramática creada correspondiente a las clases con las que se trabaja. La Figura 4 muestra una escena de casa con dos adultos y un bebé.

En la Figura 4a, se tiene la imagen original tomada del banco Pascal VOC 2012 al igual que su correspondiente Ground Truth de la Figura 4b. En la Figura 4c, se muestra la descripción de los elementos y se puede observar como el sistema es capaz de diferenciar

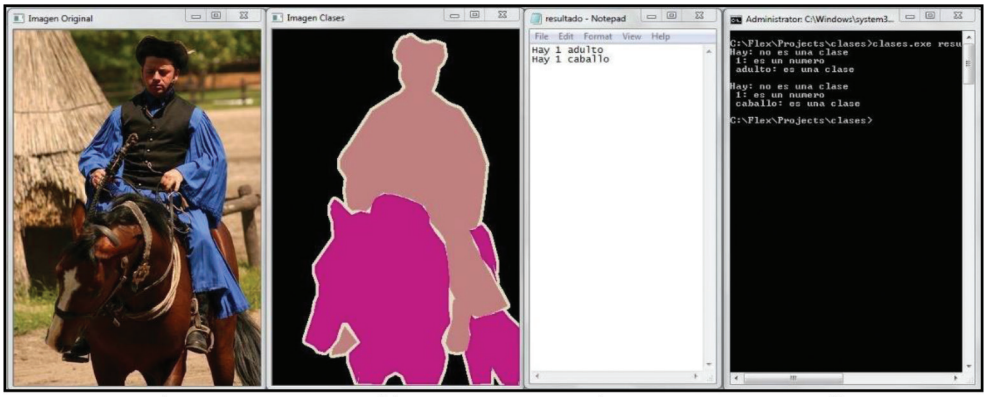
entre una persona adulta y una persona pequeña (niño, bebé); en este caso, se trata de un bebé y en la Figura 4d su verificación sintáctica.



a) b) c) d)

Figura 4 – Imagen de familia en casa, su Ground Truth, la descripción semántica y verificación sintáctica.

En la Figura 5 se muestra una escena de una persona montando un caballo.



a) b) c) d)

Figura 5 – Imagen de persona montando un caballo. Su Ground Truth, la descripción de los elementos de la imagen acorde a las clases definidas y su verificación sintáctica.

La Figura 5a corresponde a la original, la Figura 5b al Ground Truth, la Figura 5c la descripción de su contenido y finalmente en la Figura 5d, al análisis sintáctico de las descripciones. En los Ground Truth, se puede apreciar que el color de las clases se mantiene, para la clase “persona” se utiliza el color rosa claro.

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

6. Conclusiones

El detalle de la descripción de las imágenes cambia entre una y otra, esto debido a que, en algunos casos las imágenes contarán con una cantidad mayor de objetos con respecto de otras. La riqueza de las descripciones será mayor cuando haya más objetos en las imágenes, cuando tengan visibles detalles del entorno tales como: vegetación, cielo, edificios, etc. Se logró que el sistema haga una diferenciación de las etapas de la clase "persona", es decir, si se trata de un adulto o de un bebé se considera la diferencia ya que son derivaciones de dicha clase y al tener la información estructurada en árboles jerárquicos se logra obtener ese detalle de las subcategorías. Finalmente, se ha mostrado la validación sintáctica la cual realiza una verificación de las clases para solo procesar las válidas y de esta manera evitar añadir ruido al sistema con palabras que no son clases reconocidas tales como: conectores, numerales, etc.

7. Trabajo futuro

El sistema de descripciones y análisis semántico de las imágenes se seguirá complementando para que sea capaz de dar descripciones más naturales como en el caso de la Figura 3, en lugar de enunciar los elementos que aparecen, sea posible realizar interpretaciones a manera que pueda decirse que se está montado un caballo.

El entorno de la imagen será tomado en cuenta para describir si se trata de una imagen de interior, exterior, rural o urbano, así como inferir si esta frio, si es noche, día, etc. Finalmente se considera realizar una estructura de almacenamiento de la información que permita acceder y a diferentes niveles de la información, para de esta manera tener la posibilidad de crear descripciones desde diferentes perspectivas.

Agradecimientos

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

Referencias

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

Álvarez J.M., Salzmann M., & Barnes N. (2014). Large-scale semantic co-labeling of image sets. In *IEEE Winter Conference on Applications of Computer Vision*, Steamboat Springs, CO, USA, pp. 501-508. IEEE. doi: 10.1109/WACV.2014.6836060.

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

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

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

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

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

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

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

Martínez, C. J. (2013). La recuperación automatizada de imágenes: retos y soluciones. *Revista General De Información Y Documentación*, 23(2), 423–436. doi.orG/10.5209/rev_RGID.2013.v23.N2.43137.

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

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

Xu, J., Schwing, A.G., & Urtasun, R. (2014). Tell Me What You See, and I Will Show You Where It Is. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3190-3197.

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

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