Representación y Reconocimiento de Categorías Visuales mediante Gramáticas Visuales y Modelos Gráficos Probabilistas

por

Elías Ruiz Hernández

Tesis sometida como requisito parcial para obtener el grado de **Doctor en Ciencias en el Área de Ciencias Computacionales** en el Instituto Nacional de Astrofísica, Óptica y Electrónica

Supervisada por:

Dr. Luis Enrique Sucar Succar, INAOE

Representación y Reconocimiento de Categorías Visuales mediante Gramáticas Visuales y Modelos Gráficos Probabilistas

Tesis Doctoral

Por:

Elías Ruiz Hernández

Tesis sometida como requisito parcial para obtener el grado de

DOCTOR EN CIENCIAS EN EL ÁREA DE CIENCIAS COMPUTACIONALES

en el

Instituto Nacional de Astrofísica, Óptica y Electrónica

Febrero 2016

Tonantzintla, Puebla Supervisada Por:

Dr. Luis Enrique Sucar Succar

INAOE 2015

Resumen

En este trabajo de tesis se plantea la construcción de un modelo jerárquico para tareas de representación y reconocimiento de objetos mediante el uso de gramáticas visuales en combinación con modelos gráficos probabilistas. El reconocimiento de categorías de objetos es un problema abierto de investigación en donde los enfoques más comunes se han centrado en tratar de hacer reconocimiento a partir de la descripción de características locales. Si bien estos trabajos han logrado buenos resultados, debe decirse que se han centrado más bien en reconocimiento de objetos específicos, antes que en categorías de objetos. Lo anterior ha dado lugar a enfoques más orientados en atacar esta problemática. En esta área se han encontrado algunas limitaciones, tales que los modelos construidos suelen ignorar la idea de una estructura común del objeto a aprender y que los modelos son usualmente cajas negras donde no hay alguna interpretación de lo que el modelo está aprendiendo a partir de las imágenes.

El enfoque seguido en esta propuesta de tesis parte de incorporar el entendimiento de un modelo estructurado basado en gramáticas visuales a fin de obtener un modelo expresivo en el cual se puedan modelar tanto los elementos simples que componen a un objeto como las relaciones entre éstos. Para poder hacer inferencia se consideran los modelos gráficos probabilistas, que permiten incorporar incertidumbre. Se busca obtener un modelo que aproveche las capacidades de los modelos jerárquicos y los modelos gráficos para ser robustos ante ruido y oclusión en las imágenes. Las principales contribuciones del modelo propuesto en esta tesis son: i) el desarrollo de un nuevo modelo genérico de visión para reconocimiento de categorías de objetos que involucra el uso de gramáticas visuales y modelos gráficos probabilistas, ii) un algoritmo de aprendizaje de gramáticas simbólico-relacionales a partir de ejemplos, iii) un algoritmo de transformación de la gramática visual a modelos gráficos probabilistas y iv) extender las gramáticas simbólico-relacionales a gramáticas relacionales-temporales añadiendo descripción explícita de relaciones temporales de modo que se puedan representar no solamente objetos en imágenes, sino en secuencias de imágenes. Los experimentos realizados abarcan por separado los objetivos que se plantean lograr: i) verificar que una gramática puede ser capaz de distinguir al objeto descrito en un conjunto de

imágenes, ii) verificar que se puede aprender una gramática automáticamente para hacer reconocimiento a partir de ejemplos, iii) mostrar una comparación de como representan la incertidumbre tres modelos gráficos probabilistas y iv) mostrar la viabilidad de las gramática relacional-temporal para aprender y detectar secuencias visuales. El modelo propuesto se compara con otros algoritmos del estado del arte dentro del área de modelos composicionales. Los resultados muestran que el modelo sí aprende a partir de la gramática a reconocer el objeto en las imágenes con pocos ejemplos de entrenamiento. También se observó que el modelo es robusto a oclusión al poder reconocer sólo partes del objeto. Aunque otros modelos mostraron mejores resultados en exactitud del reconocimiento, se observó que el modelo propuesto es más transparente al utilizar una gramática que representa la información visual de manera más sencilla y estructurada que en otros casos. La comparación de los modelos gráficos probabilistas arrojó información de cómo puede compactarse la representación de la gramática en estos modelos que involucran incertidumbre. Finalmente, se probó la gramática relacional temporal en secuencias de imágenes que describen errores en el proceso de vestimenta de personas, mostrando que dicha representación es capaz tanto de compactar la información como de reconocer errores en secuencias de vestimenta incorrectas usando menos información que el método con el cual se comparó.

Palabras Clave: gramáticas visuales, gramáticas simbólico-relacionales, redes bayesianas, modelos relacionales probabilistas, reconocimiento de categorías de objetos, modelos composicionales.

Abstract

In this thesis, the construction of a hierarchical model for representation and object recognition tasks is presented. The proposed model involves the use of visual grammars in combination with probabilistic graphical models. Recognition of object categories is an open research problem where common approaches have focused on trying to recognize objects from the description of local characteristics. While these works have achieved good results, they have focused on recognition of specific objects, rather than object categories. In object category recognition tasks, we have found some limitations, i.e. the built models often ignore the idea of a common structure of the object to be learned and those models are usually black boxes where there is no interpretation of what the model is learning from the images.

The approach taken in this thesis is to incorporate the understanding of a structured model based on visual grammars to obtain an expressive model which can model the simple elements and the relationships among them. Inference is performed with probabilistic graphics models that incorporate uncertainty. The objective is to obtain a model that leverages the capabilities of hierarchical models and graphical models that are robust to noise and occlusion in images. The main contributions in this thesis are: i) the development of a novel model to recognize visual object categories that involves the use of visual grammars and probabilistic graphical models, ii) an algorithm to learn symbolic-relational grammars from visual examples, iii) a transformation algorithm to convert a visual grammar into probabilistic graphical models iv) an extension of a symbol-relational grammar into a temporal-relational grammar to represent not only objects in static images, but by using temporal relations in image sequences . Experiments are performed to evaluate several objectives: i) verify that a grammar can distinguish the object described in a set of images, ii) verify that a grammar can learn a grammar from examples iii) show a comparison between three models based on probabilistic graphical models, and iv) show the viability of temporal relational grammars to learn and detect visual sequences. The proposed model is compared with other algorithms of the state of the art in the area of compositional models. The results show that the model does learn from the grammar and recognizes the object in images with few training examples. It was

also observed that the model is robust to occlusion because it can recognize only parts of the object. Although other approaches showed better results in recognition accuracy, it was observed that the proposed structured model based on visual grammars is more transparent because the grammar representation provided visual information in a simpler structured way than in other cases. Comparison of the representation and inference in three probabilistic graphical models provided information about a more compact representation of the grammar when graphical models are used. Finally, the temporal-relational grammar was tested in image sequences that describe errors due to dressing activity, showing that such representation is capable of recognizing errors in dressing activity sequences using less information than other method.

Índice general

Índice de figuras

Índice de tablas

Índice de algoritmos

Tabla de Abreviaturas

BC: Base de conocimientos.

BD: Base de datos.

FHD: Full High Definition. Define la resolución de una imagen de 1920x1080 píxeles.

GDA: Grafo dirigido acíclico.

G-SR: Gramática Simbólico Relacional.

G-TR: Gramática Temporal Relacional.

HD: High Definition. Define la resolución de una imagen de 1280x720 píxeles.

HoG: Histogram of Gradient. Método para describir en un histograma las orientaciones de una imagen.

JPEG: Joint Photographic Experts Group. Formato comprimido de una imagen.

LBP: Local Binary Patterns. Método para describir patrones binarios locales en una imagen.

Lex: Lexicón.

MGP: Modelo Gráfico Probabilista.

RB: Red Bayesiana.

RBR: Red Bayesiana Relacional.

RFID: Radio Frequency IDentification. Identificación por radiofrecuencia.

RGB - Espacio de colores primarios: rojo, verde y azul.

RLM: Red Lógica de Markov.

SNM: Supresión de No Máximos.

SVM: Support Vector Machines. Máquina de vectores de soporte.

Capítulo 1

Introducción

Actualmente un tema de interés dentro de la visión computacional es el reconocimiento de objetos. Este problema se ha abordado desde dos enfoques principalmente: los sistemas de visión orientados a reconocer objetos específicos en una imagen, tal como el rostro de una persona específica; y sistemas de visión que reconocen categorías de objetos, por ejemplo, diversos tipos de sillas, autos en general, etc.

Los mecanismos que logran esto son muy diversos. Por una parte se extraen características basadas en apariencia para toda la imagen (características globales), o por regiones de la misma (algoritmos de segmentación), o basándose en la forma de la imagen (detección de bordes y curvas) o en descripción de puntos característicos de la imagen (características locales). Sin embargo, hay relativamente poco trabajo en el entendimiento de una imagen por sus componentes mediante un modelo jerárquico que defina los elementos dentro de una imagen, descomponiendo elementos complejos en otros más sencillos, con la intención de representar de manera abstracta al objeto visual y posteriormente realizar el reconocimiento de objetos en la imagen a partir de esta representación.

Una de las ventajas de considerar estos enfoques jerárquicos consiste en que se logra representar de manera estructurada un objeto visual, lo cual permite construir un modelo más general de reconocimiento (más apto para lograr reconocimiento en diversas categorías de objetos) ya que

basta con aprender la estructura del objeto a partir de los elementos más simples que se logren identificar, previo entrenamiento. Asimismo, al aprender de elementos más sencillos, es posible ganar cierta robustez, es decir, reconocer el objeto a pesar de ciertas condiciones de ruido u oclusión, puesto que si faltan unos pocos elementos sencillos, aún es posible reconocer parte de la estructura del objeto. El reto que presenta la construcción de estos modelos consiste en cómo construir una estructura abstracta que defina al objeto de manera flexible (una estructura aplicable no solamente para un objeto específico, sino para una categoría), y preferentemente, entendible para una persona, a fin de lograr avanzar en la representación y entendimiento de los sistemas de visión. También es deseable que incorpore algún mecanismo para hacer inferencia.

El reconocimiento de categorías de objetos es un problema abierto de investigación en donde los enfoques recientes se han centrado en tratar de reconocer características específicas, aunque poniendo poco énfasis en construir un modelo expresivo, que describa las características visuales encontradas y determine cómo se relacionan entre ellas a fin de tener una representación que explique la estructura del objeto. Enfoques como aprendizaje profundo *(deep learning)* [103], obtienen una representación numérica poco expresiva que pretende aprender todos los posibles ejemplos del dominio desde un sistema de caja negra. El modelo aprende de los datos aunque teniendo aún poca claridad sobre que tipo de estructura aprendió. En este trabajo se busca modelar un sistema que, partiendo de aspectos básicos de los sistemas de visión, logre una representación de éste y además sirva para tareas de reconocimiento de categorías de objetos, partiendo de una gramática visual y de modelos gráficos probabilistas, que permiten incorporar incertidumbre.

1.1. Motivación

Dentro del trabajo previo sobre reconocimiento de objetos se ha encontrado que la mayoría consideran pobremente la estructura del objeto, esto es, tratan de reconocer el objeto a partir de las características “más fuertes” que lo discriminan (color, textura, forma, esquinas, etc.), mientras que se deja de lado la idea de la representación y composición de un objeto (tal como una taza se

compone de un vaso y una asa, o que un vehículo se compone de puerta, vidrios, llantas y lámina de cierto color). En este sentido muchos modelos que logran buenas tasas de reconocimiento son poco entendibles para un humano. Los modelos más recientes basados en estrategias de aprendizaje profundo [103, 52] son un resurgimiento de las redes neuronales que han logrado posicionarse como los modelos con mayores tasas de reconocimiento, pero con capacidad poco clara para describir "qué" es lo que están aprendiendo. De este modo, en esta tesis se propone una alternativa a los modelos actuales que si bien también busca realizar tareas de reconocimiento de objetos, también trata de describir lo aprendido a través de una gramática. Nuestro enfoque es similar a lo que se hace para el lenguaje hablado o escrito, se sigue una gramática que estructura el conocimiento encontrado. Este enfoque permite hablar de "partes del objeto" que pueden estar presentes o no, según si hay oclusión del objeto visual a reconocer. Una representación así, resulta más robusta puesto que no exige el reconocimiento de todas las partes aprendidas por la gramática. Por otra parte, aprender de manera compacta en una gramática un objeto visual es una tarea que puede realizarse a partir de pocos ejemplos. Se aprenden distintas poses o variantes del objeto en una gramática al incluir reglas de producción disyuntivas (de tipo *Or*). La mayoría de los enfoques actuales pretenden construir una representación a partir de *miles* de ejemplos. En esta tesis, partiendo de elementos simples que describen regiones o partes de imágenes, los modelos pueden representarse a partir de un ejemplo. Las herramientas para construir este modelo son básicamente dos. Por un lado las gramáticas visuales, que dan transparencia al proceso de estructuración de la información aprendida y por otro lado los modelos gráficos de tipo probabilista (con un mayor interés en redes bayesianas), que manejan la incertidumbre que caracteriza a los sistemas automáticos que buscan reconocer objetos, sobre todo cuando hay problemas de oclusión, o problemas en los que no todas las partes que componen al objeto se pueden detectar debido a cambios de iluminación, cambio de apariencia o pose, entre otros.

1.2. Problemática

La problemática se centra en ¿Cómo construir un modelo composicional que represente categorías de objetos visuales basado en gramáticas visuales y que permita incorporar incertidumbre, para mostrar que también es capaz de hacer inferencia para realizar tareas de reconocimiento de objetos? Este problema ha sido atacado desde diversos enfoques, aunque sus dificultades se encuentran en cómo conjuntar aspectos de reciente interés, tales como el enfoque de representación del conocimiento a utilizar, debido al tradicional enfoque de *caja negra* que está centrado en obtener altas tasas de reconocimiento *antes* de representar qué es lo que se está aprendiendo. Otro aspecto poco entendido es cómo establecer conexiones entre la representación estructurada y el modelo que añada el manejo de incertidumbre. en ocasiones este modelo de incertidumbre se reduce a una estrategia de clasificación, haciendo estos modelos poco robustos a casos de oclusión.

Por otra parte existe literatura [88, 89, 34, 36] que combina partes de modelos estructurales con modelos basados en características, ya que incluyen tanto algún enfoque de tipo jerárquico junto con clasificación simple empleando elementos de bajo y alto nivel. Es también usual que no haya ningún consenso en el uso de un diccionario visual (que ayude a describir los elementos sencillos que se detecten). Por lo regular, el término de diccionario visual se deja de lado y en ocasiones se usa implícita o parcialmente, según resulte más conveniente. En otros casos [47, 75] los modelos tienen predefinida una gramática o quizás parte del modelo, de tal suerte que se terminan adaptando a ciertas tareas específicas antes que construir un modelo más genérico. Por genérico se entiende que sea flexible a utilizarse en dominios distintos con pocos cambios. Otra cosa observada es que aunque algunos modelos incorporan una gramática que termina dejándose de lado, teniendo una participación sólo representativa pero no activa en la construcción del modelo (por ejemplo, definen la gramática pero se utiliza un grafo de tipo árbol en su lugar). Finalmente, si el modelo permite la inclusión de modelos relacionales probabilistas[1] para representar la información visual de la gramática (algo que se considera factible, aunque poco estudiado hasta ahora) se abren las puertas a la incorporación de diversos enfoques relacionales en áreas de visión donde hasta la fecha han

[1] También son conocidos como modelos estadísticos relacionales [45].

tenido importancia limitada.

1.3. Preguntas de Investigación

En el desarrollo de esta investigación, se plantearon las siguientes preguntas:

¿El modelo a desarrollar en el presente trabajo de tesis puede representar la estructura de un objeto a partir de ejemplos de una categoría visual y representar dicha abstracción a través de una gramática visual?

¿La representación visual obtenida a través de las gramáticas visuales permite realizar reconocimiento de la misma en otras imágenes de prueba, cuantificable mediante una medida de recuerdo?

¿Cómo podrían combinarse las capacidades de representación del conocimiento de una gramática y las tareas de aprendizaje e inferencia de un modelo gráfico probabilista, para poder utilizarse en tareas de reconocimiento de categorías de objetos en imágenes sobre ciertos dominios?

¿Una gramática visual puede describir también información temporal de modo que abstraiga información de una secuencia visual?

Al finalizar esta tesis se han contestado estas preguntas de investigación a partir de los resultados encontrados.

1.4. Hipótesis

Una gramática visual puede abstraer la estructura de un objeto a partir de ejemplos de una categoría visual y tener una inferencia a través de una transformación hacia una red bayesiana, verificando la abstracción lograda a través de experimentos que involucren recuerdo.

1.5. Objetivo General

Desarrollar un modelo jerárquico que sea capaz de representar y reconocer categorías de objetos en imágenes a partir de gramáticas visuales y modelos gráficos probabilistas.

1.6. Objetivos Específicos

- Definir un alfabeto visual a utilizar en una gramática visual.
- Definir la gramática visual y sus restricciones.
- Desarrollar un método para aprender gramáticas visuales a partir de ejemplos.
- Desarrollar una transformación de la gramática visual a un modelo gráfico probabilista (red bayesiana, red bayesiana relacional y red lógica de Markov).
- Desarrollar un método para aprender los parámetros del modelo generado a partir de ejemplos.
- Desarrollar un método de inferencia para reconocer el objeto en nuevos ejemplos.
- Extender el modelo para reconocer secuencias visuales incorporando relaciones temporales.
- Validar el modelo en dominios diferentes (imágenes naturales de objetos, repositorios de categorías de objetos como Caltech [29], secuencias visuales, entre otros).

1.7. Resumen del trabajo realizado

En esta tesis se plantea la incorporación de una gramática visual que permita elaborar un modelo jerárquico de modo que a partir de elementos básicos (obtenidos mediante algún algoritmo de segmentación de imágenes, obtención de recuadros, etc.) se construyan formas más complejas mediante ciertas reglas de composición definidas en la gramática, a fin de alcanzar el reconocimiento de una categoría de objeto visual, en un contexto determinado. La importancia de atacar este problema desde un enfoque jerárquico con modelos que incorporen incertidumbre consiste en que se puede construir un modelo más robusto a problemas de oclusión en un objeto, puesto que el modelado con incertidumbre permite trabajar con evidencia incompleta. Entre las ideas que se

incorporan se consideran algunas reglas de agrupamiento perceptual de la visión [68], así como modelos gráficos probabilistas [91], a fin de tener algunas pistas inspiradas en la visión para construir una gramática y sus elementos terminales, así como también tener un enfoque estructurado como el que proveen las redes bayesianas, con la ventaja de incorporar conocimiento previo mediante éstas.

Teniendo en cuenta lo anterior, se busca construir un modelo que, definiendo una gramática visual a utilizar, los elementos visuales más básicos para hacer su detección y el modelo gráfico para realizar inferencia, esté constituido de la siguiente manera:

1. Aprenda, a partir de ejemplos, una gramática que represente a una categoría de un objeto visual.

2. Transforme la gramática a un modelo gráfico que permita considerar incertidumbre a fin de realizar inferencia con esta representación.

3. Realizar un aprendizaje de los parámetros del modelo a partir de ejemplos de entrenamiento.

4. Desarrolle una estrategia eficiente de paso de evidencia sobre nuevos ejemplos a fin de reconocer esta categoría en imágenes de prueba.

5. Se permita probar el modelo en algunos dominios, para evaluar el grado de aprendizaje de esta representación así como realizar comparativas con los trabajos relacionados.

6. Comparar la transformación a incertidumbre obtenida en el caso de redes bayesianas con modelos relacionales probabilistas.

7. Extender esta idea, en secuencias de imágenes, a fin de representar no solamente un objeto visual sino una *secuencia visual,* de manera estructurada.

De esta manera se tendrían las siguientes contribuciones con el desarrollo del modelo propuesto. Por un lado se contaría con un nuevo modelo genérico para la representación de categorías de objetos y secuencias visuales, basado en gramáticas que incorporen relaciones de tipo espacial y temporal (para el caso de secuencias visuales) mediante lógica de predicados, y el uso de modelos gráficos probabilistas que soporten estas estructuras relacionales. Por otro lado, se buscaría

automatizar la generación de un lexicón visual, algo poco considerado dentro de este tipo de trabajos. Este lexicón visual que alimenta a la gramática, sería aprendido automáticamente mediante ejemplos vía entrenamiento previo. También se contribuiría con un algoritmo que transforme la gramática a fin de que sea portable su descripción al modelo gráfico probabilista considerado. Finalmente se propondría una gramática explícita que describa relaciones temporales entre objetos para no solamente representar y reconocer objetos visuales, sino también secuencias de imágenes. Una aplicación directa de ello es poder describir vídeos, debido a que un vídeo puede ser entendido como una secuencia de imágenes.

En los experimentos realizados se probó primero la viabilidad de la representación propuesta en un caso muy simple con pocas relaciones espaciales y un sencillo diccionario visual. Posteriormente se probó la capacidad de aprender la gramática a partir de ejemplos. Después se evaluó las capacidades de precisión y recuerdo del modelo para reconocer objetos. Como la transformación a un modelo de incertidumbre es algo poco explorado en la literatura, se compararon 3 formas de hacer esta transformación. Finalmente se evaluó la capacidad de encontrar errores en secuencias visuales usando una gramática que incluye relaciones temporales.

Los resultados mostraron que la representación obtenida con gramáticas permite abstraer la estructura de la categoría visual (o la secuencia visual) y además, tiene capacidad para reconocer objetos a partir de pocos ejemplos, a pesar de contar con elementos limitados en la gramática (como el lexicón que puede ser mejorable). También se muestra que se obtienen gramáticas automáticamente a partir de ejemplos, que resultan útiles para realizar posteriores transformaciones a modelos gráficos probabilistas.

1.8. Contribuciones

- Un nuevo modelo genérico para representar categorías de objetos visuales que incorpore gramáticas visuales y modelos gráficos probabilistas

- Una variante de gramáticas visuales que incorpore relaciones temporales para representar

secuencias visuales

- Un algoritmo que aprenda gramáticas visuales automáticamente a partir de ejemplos.
- Un algoritmo de transformación de la gramática visual a un modelo gráfico probabilista.
- Un mecanismo de inferencia considerando incertidumbre que permita reconocer al objeto aprendido por la gramática.
- Un mecanismo de inferencia basado en reglas que reconozca secuencias visuales aceptadas por la gramática.

1.9. Estructura del Documento

La estructura de la tesis es como sigue. En el capítulo 2 se da una revisión de los temas necesarios para comprender los aspectos jerárquicos del modelo propuesto. En el capítulo 3 se da una revisión de los modelos gráficos probabilistas que se utilizan en esta tesis. Posteriormente se revisan trabajos previos de los modelos composicionales en el capítulo 4. En los capítulos 5, 6 y 7 se describe el método para construir el modelo de visión propuesto con sus distintas variantes. En todas las variantes se utilizan gramáticas visuales de tipo relacional. La inferencia, en una de estas variantes se realiza con redes bayesianas (cap. 5), otra con modelos relacionales probabilistas (cap. 6), y en una más se utiliza inferencia basada en reglas y se añaden relaciones temporales a la definición de la gramática (cap. 7). En el capítulo 8 se muestran los resultados alcanzados en comparación con otros trabajos similares y finalmente en el capítulo 9 se exponen las conclusiones y el trabajo futuro que se desprenden de esta tesis.

Capítulo 2

Representación y Reconocimiento Visual de Objetos

2.1. Introducción

En esta sección se describen fundamentos teóricos necesarios para construir la representación del conocimiento que se usará en nuestro modelo. Se abordarán dos aspectos importantes vinculados a este trabajo: reconocimiento visual de objetos y representación estructurada de objetos mediante gramáticas visuales.

2.2. Conceptos sobre Reconocimiento de Objetos

Representación de un Objeto. Este concepto se refiere a la manera en que un elemento visual se expresa en términos numéricos para su procesamiento en un algoritmo computacional. Normalmente un objeto es representado mediante características de diversa índole como pueden ser: color, textura, forma, etc. Cuando se habla de reconocimiento de objetos en un modelo jerárquico significa que un elemento visual es descompuesto en otros objetos, que a su vez también pueden descomponerse. Al final los elementos más simples son descritos igualmente mediante características diversas. La hipótesis que subyace para construir un modelo jerárquico es que esta forma de

modelar es más expresiva y comprensible para un humano, a diferencia de aquellos casos donde un objeto se representa computacionalmente mediante miles o quizás decenas de miles de características.

Categorías de Objetos. Los objetos pueden agruparse por tener características similares. Por ejemplo, manzanos, robles y pinos pertenecen a la categoría de árboles. Mazda, Pontiac y Mercedes Benz pertenecen a la categoría de automóviles. Los trabajos en reconocimiento de objetos se distinguen por perseguir dos tareas:

1. Reconocer objetos específicos.

2. Reconocer categorías de objetos.

Si bien existe una gran cantidad de trabajos en ambas tareas, la segunda tiene un mayor interés. Los objetos específicos se han logrado detectar con ayuda de algoritmos basados en descripción de puntos de interés locales [69]. En cambio, para reconocer categorías aún no hay consenso sobre cuál podría ser una buena estrategia a seguir. Sin embargo, está habiendo un interés creciente en tratar de construir modelos estructurales que lleven a cabo esta tarea. En este trabajo se plantea un modelo que sirva para reconocer categorías de objetos sobre ciertos dominios como escenas de interiores, por poner un ejemplo.

Cuando se habla de reconocimiento de objetos, filosóficamente existen dos maneras de construir un modelo para esta tarea: los modelos discriminativos y los modelos generativos. Los modelos discriminativos tratan de aprender los límites que existen entre la distribución de las clases de objetos. En términos bayesianos, los modelos discriminativos aprenden la probabilidad posterior de encontrar el objeto dada una imagen, teniendo la ventaja de aprender a discriminar clases. Por el contrario, los modelos generativos aprenden la distribución de las posibles clases de objetos, lo cual permite generar ejemplos una vez entrenado el modelo. En términos bayesianos, los modelos generativos aprenden la probabilidad conjunta del objeto y la imagen. Debido a que en esta tesis se plantea la representación y el reconocimiento de categorías visuales de objetos (distintas entre sí) antes que la generación de ejemplos sintéticos de categorías visuales. En este trabajo se propone

un modelo discriminativo que aprende categorías de objetos mediante gramáticas visuales y redes bayesianas.

Inferencia. La inferencia consiste en el proceso necesario para decidir si existe un objeto en una imagen. En este trabajo se realiza inferencia de tipo bayesiana sobre un modelo gráfico. La inferencia puede realizarse de dos formas:

- **Arriba-Abajo**: se realiza la inferencia a partir de los parámetros globales del modelo. Este tipo de inferencia trata de predecir los efectos a partir de las causas.
- **Abajo-Arriba**: se realiza inferencia a partir de los elementos encontrados para construir hipótesis más generales en cada paso. Se trata de predecir la causa dados los efectos de algún fenómeno.

En esta tesis se considera la inferencia de abajo hacia arriba puesto que se plantea encontrar los elementos terminales que constituyen a un objeto y posteriormente inferir si pertenece a cierta categoría.

2.3. Enfoques utilizados en reconocimiento de objetos

Aunque es difícil discriminar entre enfoques seguidos para construir modelos de reconocimiento de objetos, se seguirá la discriminación de enfoques dada por [119], en el cual se distinguen tres formas distintas de construir los modelos de reconocimiento:

1. **Modelos basados en características invariantes.**

   Básicamente consiste en obtener por alguna vía características globales o locales de una imagen, globales como el histograma de color de la imagen o locales como el uso de descriptores de regiones. El objetivo es quedarse con aquellas características que mejor definan al objeto, llamadas características invariantes. Una vez obtenidas las características más relevantes, el problema pasa a ser considerado un problema de clasificación. Estos modelos son los más ampliamente utilizados y tienen la ventaja de lograr para algunos objetos tasas muy altas

de reconocimiento [8]. En particular, el uso de descriptores locales (tales como SIFT [69] y SURF [4]) le han dado un mayor auge. No obstante aún no hay consenso claro de qué características invariantes utilizar para lograr el reconocimiento de objetos en general. De esta manera, los modelos basados en características invariantes suelen esforzarse en aprender la mayor cantidad de características y seleccionar las mejores (de manera manual o automática) antes de entender de una manera más estructurada los componentes de un objeto. Los modelos más recientes basados en aprendizaje profundo [52, 103], también se ubican aquí.

2. **Modelos basados en la descripción de la estructura y sus partes.**

   Este enfoque surge ante el entendimiento de cómo los objetos pueden ser descompuestos en elementos cada vez más sencillos que podemos llamar *componentes* del objeto. Esta idea tiene inspiración en una forma más "natural" o más "humana" de reconocer un objeto. Como ejemplo, nosotros reconocemos una casa porque se compone de paredes, puertas y ventanas. Un rostro se compone de ojos, nariz y boca. Lo que plantea este enfoque es tratar de aprender la organización de los elementos más simples para primero reconocer estos elementos simples y luego encontrar la organización de los mismos, para tratar de inferir el reconocimiento del objeto completo. Una de las ventajas de estos enfoques es que es posible reconocer inclusive en situaciones de falta de alguna de las partes. De una manera análoga que en el enfoque basado en características, no hay un consenso sobre los elementos simples que deben utilizarse. Es común en la literatura denominar como *diccionario visual* al conjunto de estos elementos simples. Este diccionario visual normalmente se compone de bordes, esquinas, formas geométricas, regiones y descriptores locales. Este enfoque es el que nos interesa para la revisión bibliográfica. No obstante, aún este enfoque es bastante extenso. En la siguiente sección se hará una división un poco más estricta para poder quedarnos con los trabajos que hagan un reconocimiento basado en este enfoque jerárquico además de otras restricciones.

3. **Modelos basados en alineamiento.**

   Este enfoque se centra en tener un modelo computacional del objeto a reconocer y compensar las variaciones que pueden tener los objetos en el mundo real en términos de orientación,

escala o apariencia mediante un conjunto de transformaciones permitidas. La idea es obtener una correspondencia entre el objeto en una imagen y el modelo después de aplicarle a éste una serie de transformaciones. Estos modelos son muy utilizados para tareas de reconocimiento óptico de caracteres [13], donde se busca reconocer letras sin importar su posición, orientación o escala. Su aplicación a objetos reales en 3D es aún difícil debido a que no resulta trivial determinar la apariencia del objeto real una vez que ha sido movido, girado, distorsionado u ocluido parcialmente.

Considerando que se busca aprender categorías visuales con cierta estructura que pueda representarse mediante una gramática visual, el trabajo propuesto queda englobado dentro del segundo punto. En el siguiente capítulo se tratará de dividir este enfoque de estructura y partes, a fin de hacer una subdivisión más precisa de las diferencias que presentan este tipo de modelos.

2.4. Relaciones Espaciales y Temporales

Las relaciones espaciales permiten describir el mundo referenciando objetos en términos de la posición de otros objetos. De acuerdo a [27, 106] existen relaciones topológicas, de orden, métricas y difusas. Las relaciones topológicas se preservan ante una variedad de transformaciones espaciales como rotación y escalamiento. Ejemplos de relaciones topológicas son $Adyacente$, $DentroDe$, $Traslapado$. Las relaciones de orden son aquellas que no son invariantes a cambios de rotación. Ejemplos: $ArribaDe$, $IzquierdaDe$. Las relaciones métricas describen cuantitativamente la relación sostenida entre dos objetos. $Distancia(A,B) = 20$ píxeles, sugiriendo que hay una distancia de 20 píxeles entre dos regiones de una imagen es un ejemplo de una relación métrica. Finalmente las relaciones difusas son aquellas que presentan ambigüedad en el lenguaje tal como $Lejos$ y $Cerca$. Aunque estas relaciones espaciales pueden definirse en sentido estricto, normalmente son difusas porque suelen definirse bajo un margen que puede ser incluso traslapado (ambas relaciones coexisten), según en el dominio donde se trabaje. La manera más simple de describir relaciones espaciales entre objetos es con predicados o funciones, por ejemplo: $ArribaDe(A,B)$ donde los argumentos A

y *B* son objetos que sostienen una relación espacial: el objeto *A* se encuentra *arriba de* el objeto *B*. Si la definición lo incluye, también puede indicar que las regiones son adyacentes. En esta tesis se utilizarán las relaciones de topológicas y de orden. Las relaciones métricas no son consideradas debido a que añaden una restricción de escala que reduce las posibilidades de reconocer objetos de diversos tamaños. Las relaciones difusas no son utilizadas debido a que, al coexistir dos relaciones que en inicio son opuestas, puede haber una pérdida de precisión del objeto, es decir, una gramática podría decir que un árbol se compone de un follaje y un tronco que están *cerca y lejos* uno del otro.

Por otro lado las relaciones temporales se han usado para representar una secuencia entre objetos a través del tiempo. Normalmente los objetos son eventos del mundo que ocurren en cierto intervalo de tiempo. A manera de ejemplo $Después(Levantarse, Bañarse)$ sugiere que la actividad *Levantarse* ocurre primero y acto seguido la actividad *Bañarse* vendrá *después*. Allen [2] propuso que tanto las relaciones espaciales como las temporales pueden combinarse dando lugar a un espacio de configuraciones entre objetos o eventos. En la Fig. 2.1 se muestra esta combinación de relaciones espaciales y temporales. En muchos ejemplos prácticos las relaciones que aparecen en recuadros de color son las más usadas. En esta tesis se trabaja principalmente con relaciones de tipo espacial para describir una imagen y en un caso especial: investigar la inclusión de relaciones temporales para descubrir secuencias correctas de imágenes, donde cada imagen también debe tener una configuración espacial determinada. De manera más genérica, las relaciones espaciales y temporales son capaces de describir vídeos. En el capítulo 7 se muestra el método para describir estas secuencias en un dominio particular.

2.5. Conceptos sobre Gramáticas Visuales

Los lenguajes visuales permiten, a partir de ciertos elementos y relaciones, construir un concepto o comunicar una idea. Señales, diagramas y fotografías, son expresión de un lenguaje visual. Los propósitos en la investigación de lenguajes visuales son variados, uno de ellos es entender cómo pueden clasificarse éstos de manera natural y cómo pueden ser especificados de manera formal.

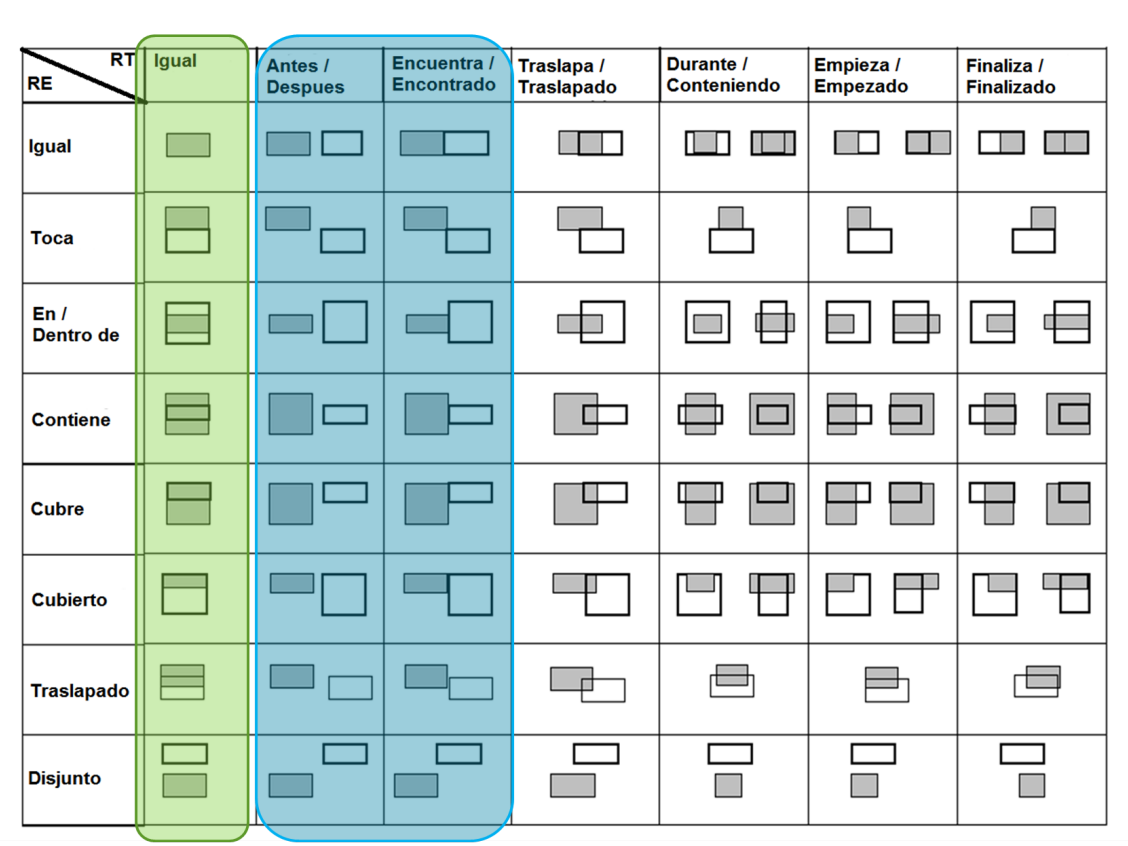


Figura 2.1: Tipos de relaciones espaciales (RE) y sus combinaciones con relaciones temporales (RT). Las relaciones en el recuadro de la izquierda (columna igual) no consideran temporalidad. Las relaciones en el recuadro derecho (columnas antes / después y encuentra / encontrado) de mayor tamaño son las relaciones espaciales más básicas añadiendo temporalidad. Imagen tomada de [19].

Algo que motiva la investigación en esta área es facilitar la interacción humano-computadora. Las gramáticas visuales surgen como una manera de representar los lenguajes visuales de una manera formal, resultan de una extensión de las gramáticas usadas en teoría sintáctica para describir lenguajes formales. Existen variados formalismos que permiten definir gramáticas visuales, los cuales se constituyen por una serie de reglas que restringen las configuraciones que puede admitir el lenguaje visual.

2.5.1. Tipos de gramáticas

En esta sección se definen algunos tipos de gramáticas previas a la gramática que se utilizarán en esta tesis. Más adelante se propone en esta tesis las gramáticas temporales relacionales como una extensión a las gramáticas simbólico–relacionales.

Gramática Simbólica (Tradicional). De acuerdo a [73], las gramáticas simbólicas se componen de cuatro elementos: símbolo inicial S, conjunto de elementos no terminales V_N, conjunto de elementos terminales V_T y reglas de producción P. suele abreviarse como: $G = (S, V_N, V_T, P)$. Si además la gramática es libre de contexto, las producciones P son de la forma: $v \rightarrow (V_N \cup V_T)*, v \in V_N$, donde el asterisco indica una secuencia de 0 a n elementos del conjunto $(V_N \cup V_T)$. Las demás gramáticas definidas aquí, son una extensión de esta definición.

Gramática Transformacional. Descritas por [17], estas gramáticas describen un elemento complejo en partes más sencillas que los componen. En modelos visuales la descomposición tiene como límite los píxeles de los que una imagen se compone. Sin embargo, rara vez se llega a este nivel. Por lo regular los elementos más sencillos suelen ser líneas, curvas, conectores, figuras geométricas, puntos locales, entre otros. Las reglas de producción tienen relaciones espaciales implícitas. Un ejemplo es: $A \rightarrow ab$ sugiere que los elementos a y b sostienen una relación de adyacencia no definida explícitamente.

Gramática Estocástica. Comentadas en [11, 12], las gramáticas estocásticas extienden las gramáticas transformacionales para definir un marco para representar un objeto y su descomposición en elementos más simples mediante un modelo probabilista. En particular, las gramáticas

estocásticas incorporan un valor de probabilidad para cada regla de producción P de la gramática. Un ejemplo sencillo es:

$p_i\, v \rightarrow (V_N \cup V_T)*, v \in V_N,\, p_i \in [0 \ldots 1],$

donde p_i es un valor de probabilidad (entre 0 y 1) y v debe instanciarse más de una vez. La suma de los valores de probabilidad entre las diversas instancias de v debe ser uno. La interpretación de esto es que ante dos posibles caminos que pueden seguir las reglas de producción de la gramática, hay cierta probabilidad de escoger las reglas de producción, haciendo que ciertas configuraciones sean más probables que otras.

Gramática Relacional. Descritas en [121], las gramáticas relaciones tienen la propiedad de que cada regla de producción está compuesta por un conjunto de símbolos $V_N \cup V_T$ que representan objetos visuales, los cuales están relacionados por un conjunto de elementos relacionales binarios (de aridad = 2). Es usual que la descripción de los elementos relacionales sea a través de un predicado: $relación(A, B)$, donde $A, B \in V_N \cup V_T$ y $relación \in V_R$ es un elemento relacional que describe una interacción entre A y B.

Después de esta revisión de tipos de gramáticas, Se observa que un modelo que estructure información visual (información en dos dimensiones) requiere una gramática de tipo relacional, en donde regiones o descripciones de la imagen se comporten como símbolos y se conecten mediante relaciones de tipo espacial. Una alternativa para evitar una gramática relacional, es codificar la imagen de manera unidimensional, perdiendo algo de claridad en cuanto a la estructura del objeto en la imagen que se quiere aprender.

2.5.2. Alfabeto Visual y Lexicón

Las gramáticas suelen describir en un apartado especial (el lexicón) el conjunto de elementos terminales que admite dicha gramática. Si dichos terminales son palabras, tales palabras se componen de una sucesión de letras que pertenecen a un alfabeto. Existen definiciones análogas para lenguajes visuales descritas a continuación:

Alfabeto Visual. Es la manera en como se extraen los elementos más primitivos con los que una

gramática trabaja. Así como el alfabeto se compone de letras, un alfabeto visual puede componerse desde lo más simple (píxeles) hasta elementos más complejos: recuadros, puntos de interés local, regiones de una imagen, bordes, formas curvas, etc.

Lexicón Visual. Es un conjunto de descripciones de los elementos terminales de una gramática. Cada elemento del lexicón está constituido a partir de los elementos del alfabeto visual. Así como en textos una palabra se compone de varias letras, una palabra visual puede estar constituida a partir de uno o varios píxeles, de uno o más puntos de interés local, de un recuadro, de un borde o la fusión de varios bordes, etc. En textos, si una gramática pide un verbo como elemento terminal, el lexicón describe cuáles verbos (y sus variantes) son aplicables para dicho terminal (por ejemplo: $Verbo := \{comer, comiendo, come, comes\}$). En gramáticas visuales se suele prescindir del lexicón debido a que un terminal describe sólo una región específica; sin embargo, en este trabajo se propone incorporar un *Lexicón Visual,* que admita variantes del objeto descrito por un elemento terminal.

En este sentido, un lexicón visual consiste en describir qué elementos de una imagen son candidatos a ser denominados como elementos terminales que aparecerán en una gramática. El lexicón varía si se cambia el alfabeto visual. En otras palabras, el lexicón escoge los elementos más idóneos para ser elementos terminales y el alfabeto visual es el que provee todo el universo posible de elementos terminales; es decir $L \subseteq \mathscr{A}$, donde L es un lexicón y $\mathscr{A}$ es un alfabeto visual. En ocasiones, por claridad, los elementos de L, se renombran. El lexicón visual sirve para representar en abstracto regiones o partes de una imagen. Un lexicón permite restringir los terminales a utilizar, de modo que pueda descartar elementos de la imagen que no se correspondan con algún elemento de L, ayudando a reducir ruido.

Después de ver esta revisión de las gramáticas, se definirá la gramática a usar en el modelo propuesto.

2.5.3. Gramática Simbólico-Relacional

En este trabajo se usará este tipo de gramática para construir el modelo propuesto debido, por un lado, a sus propiedades relacionales. Por otro lado, estas gramáticas se utilizan para codificar

información visual de diagramas de flujo, de modo que hay un potencial de uso para representar objetos visuales.

Definición. Descritas en [33], las gramáticas simbólico-relacionales son una extensión a las gramáticas relacionales, las cuales permiten proporcionar un alto nivel de descripción de los lenguajes visuales. De manera formal una *gramática Simbólico-Relacional* (abreviadas como *gramáticas SR*) es una 6-tupla: $G = (V_N, V_T, V_R, S, P, R)$ donde:

- V_N es un conjunto finito no vacío de símbolos *no terminales*.
- V_T es un conjunto finito no vacío de símbolos *terminales*. Siempre se cumple que $V_N \cap V_T = \varnothing$
- V_R es un conjunto finito de símbolos relacionales.
- $S \in V_N$ es el símbolo inicial.
- P es un conjunto finito de reglas de reescritura, llamadas *producciones-s*, de la forma

$$l : Y^0 \rightarrow \langle \mathbf{M}, \mathbf{R} \rangle$$

  donde

  - l es un entero que etiqueta las *producciones-s* y es un identificador único
  - $\langle \mathbf{M}, \mathbf{R} \rangle$ es una sentencia sobre V_R y $V_N \cup V_T$,
    - donde $\mathbf{M}$ es un conjunto de *s-ítems* (v, i) con $v \in V_N \cup V_T$ y el símbolo i es un número natural usado para distinguir entre diferentes ocurrencias de el mismo símbolo. Cada *s-ítem* se puede abreviar de la forma v^i.
    - $\mathbf{R}$ es un conjunto de *r-ítems* de la forma $r(X^i, Y^j)$, con $X^i, Y^j \in \mathbf{M}$ y $r \in V_R$
  - También se cumple siempre que $Y^0 \in V_N$ y además $Y^0 \notin \mathbf{M}$

- R es un conjunto finito de reglas de reescritura, llamado *producciones-r*, de la forma

$$r(Y^0, X^1) \rightarrow [l]\,\mathbf{Q}$$

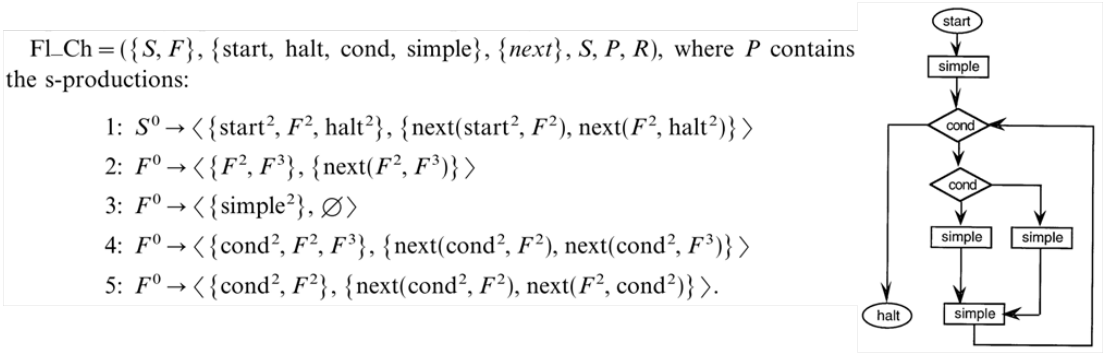


Figura 2.2: Representación mediante una gramática SR de un diagrama de flujo de un algoritmo. Las flechas son descritas mediante la relación "next" en la gramática. Debe tenerse en cuenta que existe una analogía entre una imagen y un diagrama de flujo. En el ejemplo, dado que no se presentan reglas de reescritura, se define $R = \varnothing$. Imagen tomada de [33].

o

$$r(X^1, Y^0) \rightarrow [l]\mathbf{Q}$$

donde:

- $r \in V_R$,
- l es la etiqueta de una *producción-s* $Y^0 \rightarrow \langle \mathbf{M}, \mathbf{R} \rangle$,
- $\mathbf{Q} \neq \varnothing$ is un conjunto finito de *r-ítems* de la forma $r(\mathbf{Z}, X^1)\, o\, r(X^1, \mathbf{Z})$, con $\mathbf{Z} \in \mathbf{M}$

Los exponentes cero sobre las variables del lado izquierdo en reglas de producción, y los exponentes uno sobre las variables del lado derecho, son un formalismo para indicar el lado en el que se encuentra la variable en las reglas de reescritura. Un ejemplo de una gramática SR que define un diagrama de flujo de un algoritmo se ilustra en la Fig. 2.2. Un *s-ítem* es un elemento simbólico que se escribe de manera compacta de la forma v^i y representa a un elemento visual (una región, un píxel, recuadro, etc.). Un *r-ítem* es un elemento relacional que se escribe de forma compacta como $r(X^i, Y^j)$ y representa una relación (descrita como un predicado) que sostienen los dos elementos simbólicos X^i y Y^j.

2.6. Resumen

En esta sección se estudiaron los conceptos básicos relacionados con el reconocimiento de objetos, las principales tareas asociadas al reconocimiento de categorías y las gramáticas visuales, como un instrumento para estructurar el conocimiento visual que se puede obtener a través de las imágenes. En el siguiente capítulo se estudiarán los modelos gráficos probabilistas que permitirán incorporar incertidumbre a la representación del conocimiento provista por las gramáticas visuales. También se verá una extensión de estos modelos que permiten tratar con datos en donde la estructura y relaciones entre las variables son modeladas de manera conjunta. El estudio de estos modelos relacionales probabilistas permitirá proponer un modelo que combine las capacidades de las gramáticas visuales con los modelos relacionales probabilistas para hacer un análisis comparativo con el modelo que se propone en esta tesis.

Capítulo 3

Modelos Gráficos Probabilistas

En este capítulo se abordarán los modelos gráficos probabilistas y una extensión de los mismos conocida como modelos relacionales probabilistas. Los primeros permiten modelar dependencias entre variables, y los segundos son una generalización de los primeros que buscan modelar estas dependencias sabiendo que las variables guardan una estructura subyacente entre ellas. El modelo de visión propuesto en esta tesis maneja incertidumbre utilizando este tipo de modelos. Los conceptos más importantes asociados a estos modelos se presentan en este capítulo.

3.1. Introducción

Los Modelos Gráficos Probabilistas (MGP) [91, 62] son tipos especiales de modelos estadísticos que caracterizan variables aleatorias y las relaciones de dependencia que hay entre ellas. Una variable aleatoria puede admitir valores de un conjunto dado. Si el conjunto es numerable la variable aleatoria es de tipo discreto, mientras que si no es numerable, la variable aleatoria es de tipo continuo. En esta tesis se trabajó con variables discretas solamente. Las variables aleatorias pueden definir la probabilidad de que en un lugar determinado del planeta esté nublado, llueva o haya mucho viento. Las variables aleatorias discretas asociadas admitirían valores como si-llueve, no-nublado, mucho-viento, etc. Las relaciones de dependencia permiten establecer conexiones entre pares de variables, tal como definir la probabilidad de ciertos hábitos alimenticios de las personas

(primera variable) dado que viven en cierto lugar del planeta (segunda variable). Es de observarse que todo MGP siempre es un modelo estadístico. En esta tesis se utilizan los MGPs para poder modelar las relaciones de dependencia que existen entre elementos que se descubren en una imagen para poder describir el contenido de la misma. Por ello, en esta sección se definen algunos tipos de MGPs.

Por otra parte, los modelos estadístico-relacionales (o simplemente modelos relaciones probabilistas) [45] resultan de una extensión de los MGP o bien son una extensión de modelos lógicos. A diferencia de los MGP, los modelos relacionales tratan con información estructurada (tal como una base de datos o una base de conocimiento) en donde aprenden relaciones causales entre variables de forma parecida que un MGP, pero tomando ventaja de la estructura subyacente en los datos. El resultado es que los modelos relacionales pueden instanciar MGPs de acuerdo a consultas que ellos reciban.

3.2. Conceptos sobre Modelos Gráficos Probabilistas.

En esta sección se presenta brevemente la teoría sobre redes bayesianas y las redes bayesianas relacionales, que serán utilizadas en nuestro modelo. En términos generales se definen a los modelos estadísticos como pares del tipo $(\mathscr{S}, \mathscr{P})$ donde $\mathscr{S}$ son un conjunto de variables aleatorias y $\mathscr{P}$ son distribuciones de probabilidad asociadas a ellas. Variantes de distintos modelos gráficos probabilistas añaden restricciones y extensiones a esta definición. En las siguientes secciones se definen los modelos probabilistas que son utilizados en esta tesis.

3.2.1. Redes Bayesianas

Una red bayesiana [91] es un grafo dirigido acíclico (GDA) (es decir no tiene ciclos dirigidos), que consiste de:

1. Un conjunto de variables aleatorias que constituyen los nodos de la red.

2. Un conjunto de arcos dirigidos que conectan pares de nodos (que se pueden interpretar como relaciones de dependencia).

3. Para cada nodo se especifica su función de distribución condicional que cuantifica los efectos que los padres tienen sobre el nodo.

Se utilizará la definición discreta, de modo que la función de distribución condicional es una tabla de las probabilidades (denominada TPC, tabla de probabilidad condicional) que un nodo hijo puede tomar para cada combinación de valores de sus nodos padres. De manera más formal se dice que una red bayesiana es un par $(\mathscr{G}, \mathscr{P})$, donde $\mathscr{G}$ es un GDA sobre las variables aleatorias $X, X_2, ...X_n$ y $\mathscr{P}$ es la distribución de probabilidad conjunta sobre las variables. De la definición anterior sigue que la distribución de probabilidad conjunta está dada por:

$$P(X_1, X_2, ...X_n) = \prod_{i=1}^{n} P(X_i | Pa_G(X_i)), \tag{3.1}$$

donde Pa_G representa a los nodos padres de la variable aleatoria X_i

Una red bayesiana permite representar de manera gráfica las dependencias entre variables aleatorias, de modo que permiten simplificar la representación del conocimiento y razonamiento. Un ejemplo de una red bayesiana se ilustra en la Fig. 3.1.

El aprendizaje en las redes bayesianas consiste en deducir un modelo, en particular una estructura para la red y los parámetros asociados a cada nodo. Así hay dos tipos de aprendizaje para construir una red bayesiana:

Aprendizaje paramétrico: si se considera que la estructura de la red ya está dada, la manera más común de obtener los parámetros del modelo es estimar las probabilidades a partir de las frecuencias de los datos. Este método se conoce como estimador de máxima verosimilitud. Hay dos casos posibles [83]:

- Nodos sin padres: sólo precisan la probabilidad marginal $P(X_i = x_i)$, que es equivalente al número de ocurrencias del estado x_i de la variable X_i sobre el total de datos obtenidos.

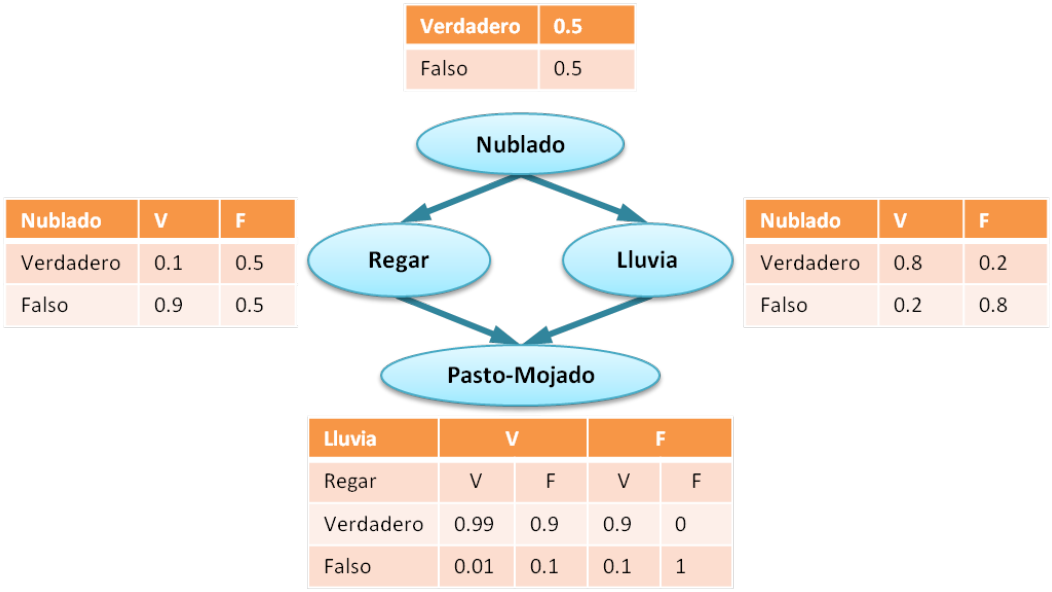


Figura 3.1: Ejemplo de una red bayesiana. Las variables aleatorias que se modelan se pueden representar mediante nodos o círculos. Las relaciones de dependencia se ilustran gráficamente con flechas entre las variables. Las distribuciones de probabilidad asociadas a cada variable se muestran mediante tablas de probabilidad condicional que en el ejemplo de la imagen, se muestran con el caso discreto.

- Nodos con padres: se requiere estimar la probabilidad condicional dados sus padres $P(X_i = x_i | Pa_G(X_i))$, que equivale al número de ocurrencias del estado x_i de la variable X_i con cada combinación de estados de los padres de X_i, sobre el total de casos en que se cumple $Pa_G(X_i)$ para esa misma combinación de estados.

En casos donde no sea posible obtener información estadística de todas las variables, se puede utilizar otro tipo métodos como Esperanza-Maximización (EM) [23, 9]. Este método permite encontrar estimadores de máxima verosimilitud en casos donde hay variables no observables. El algoritmo consiste en unos pasos iterativos que realizan a grandes rasgos los siguiente:

- Paso 1: estimar los valores faltantes de las variables no observables de alguna forma (inclusive aleatoria).
- Paso 2: estimar los parámetros del modelo a partir de estos valores introducidos.
- Paso 3: volver a estimar los valores faltantes a partir de los parámetros obtenidos (ya no es aleatorio, es una iteración).

- Paso 4: repetir desde el paso dos.

Los pasos anteriores se repiten hasta llegar a un criterio de convergencia. De manera más formal, son dos pasos, uno de esperanza y otro de maximización:

Paso E (esperanza): se calcula un función Q para estimar el valor de los datos Y (observados y ocultos) suponiendo unos parámetros conocidos h y teniendo los datos observados Z:

$$Q(h'|h) \leftarrow E[\ln p(Y|h')|h,Z]. \tag{3.2}$$

Paso M (maximización): con los datos obtenidos se sustituye la hipótesis h por la hipótesis h' que maximiza esta función Q:

$$h \leftarrow \arg\max_{h'} Q(h'|h). \tag{3.3}$$

dado que a veces se tienen variables no observables, este método será el preferido para el aprendizaje de parámetros en el método desarrollado en esta tesis.

Aprendizaje estructural: este aprendizaje consiste en encontrar las relaciones de dependencia e independencia entre las variables involucradas. Hay en general, dos maneras de aprendizaje de la estructura de la red:

- Métodos que detectan las independencias entre variables.
- Métodos que usan criterios de evaluación y búsqueda de estructuras.

Los primeros están basados en el concepto de relaciones de independencia entre variables (también llamado separación-d [91]) y los segundos están basados en la creación de una función heurística (normalmente es un algoritmo voraz) que evalúa las estructuras tratando de podar el espacio de búsqueda. La manera de evaluarlas es diversa, aunque suele usarse un criterio de entropía mínima: las variables aceptan aquella configuración que minimice la entropía entre los datos, de modo que se favorezcan las conexiones entre variables con alto grado de dependencia. Del primer método, la técnica más conocida es *PC* [111] y de los segundos, los dos trabajos más conocidos son los

de Chow-Liu [18] y los de Rebane-Pearl [98]. En algunas ocasiones un experto en el dominio puede determinar *a priori* estas relaciones, por lo que la estructura de la red puede ser construida manualmente.

En el modelo propuesto en esta tesis, se considera un aprendizaje estructural a partir de las reglas de producción provistas por la gramática SR. Para el caso del aprendizaje paramétrico se considera tanto un enfoque de estimación subjetiva de las probabilidades para las TPC, como un aprendizaje estadístico a partir de ejemplos. Como algunas variables que se crean son ocultas, este aprendizaje estadístico está basado en el algoritmo EM antes mencionado.

3.2.2. Redes de Markov

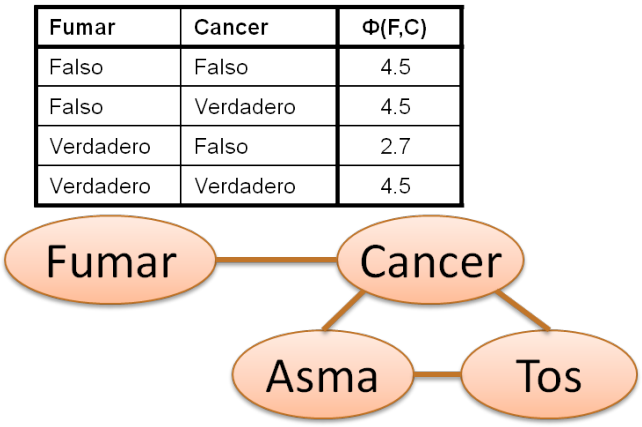
Una red de Markov (que también es conocida como campo aleatorio de Markov o CAM) es un modelo que representa la distribución conjunta de una serie de variables $X = (X_1, X_2, ..., X_n) \in \mathscr{X}$ [62]. La red de markov se compone de un grafo no dirigido G y un conjunto de funciones potenciales llamadas ϕ_k. El grafo asociado tiene un nodo por cada variable y además una función potencial ϕ_k por cada cliqué que se encuentre en el grafo. De manera sencilla, una función potencial es solamente una función que mapea a números reales positivos ($\Re^+$) los estados del cliqué correspondiente. Para simplificar la tarea, se trabajará con estados discretos de las variables únicamente. Con ello, la distribución conjunta representada por la red de Markov está dada por:

$$P(X = x) = \frac{1}{Z}\prod_k \phi_k(x_{\{k\}}), \tag{3.4}$$

donde $x_{\{k\}}$ es el estado del cliqué k-ésimo y Z es una función de partición, dada por:

$$Z = \sum_{x \in \mathscr{X}} \prod_k \phi_k(x_{\{k\}}), \tag{3.5}$$

Un ejemplo de una red de Markov se ilustra en la Fig. 3.2. Las funciones de partición se representan por tablas con pesos asociados a cada combinación de estados de las variables involucradas en cada



Fumar	Cancer	Φ(F,C)
Falso	Falso	4.5
Falso	Verdadero	4.5
Verdadero	Falso	2.7
Verdadero	Verdadero	4.5

Asma	Cancer	Tos	Φ(A,C,T)
F	F	F	5.1
F	F	V	3.2
F	V	F	1.2
F	V	V	1.9
V	F	F	1.3
V	F	V	3.6
V	V	F	2.9
V	V	V	4.7

Figura 3.2: Ejemplo de una red de Markov y sus tablas de pesos asociados a cada cliqué. Para calcular la probabilidad conjunta, es necesario computar la función de partición Z, que es resultado de la suma del producto de todas las configuraciones de los cliqués en la red.

cliqué. Es importante decir que las redes de Markov también pueden representarse mediante modelos log-lineales, de modo que sustituyen el producto de las funciones potenciales por la función exponencial de una suma pesada:

$$P(X = x) = \frac{1}{z} \exp \sum_j w_j f_j(x), \tag{3.6}$$

donde w_j es un peso (un valor en $\Re^+$) y f_j es, para los objetivos de esta tesis, una fórmula binaria $f_j(x) \in \{0,1\}$.

A diferencia de las redes bayesianas, las redes de Markov representan relaciones entre variables de manera no dirigida de modo que la influencia entre dichas variables es en ambos sentidos. Una aplicación muy común de los CAM es en tareas de procesamiento de imágenes ya que representan las relaciones espaciales entre los píxeles o segmentos de la escena. Se entiende que existe una influencia de una región con otra que se encuentre cercana a ella. [62].

Dado que el cálculo de la función de partición Z depende en tiempo de la cantidad de estados discretos y el número de cliqués, se sigue que el cálculo anterior es NP-completo [62]. Algunos trabajos tratan de aproximar el cálculo para hacerlo más eficiente [46]. Sin embargo, para modelos pequeños, el cálculo aún en inferencia exacta puede ser eficiente.

3.3. Modelos Relacionales Probabilistas

Los primeros modelos relacionales probabilistas surgieron en un tiempo relativamente reciente [54, 37, 44, 99, 45] (finales de los noventa e inicios de este siglo) para tratar de generalizar el tratamiento a datos que tienen una estructura intrínseca conocida. El objetivo de estos modelos consiste en aprovechar esa estructura conocida de los datos e incorporarla al modelo que aprende a predecir dichos datos. A diferencia del uso de clasificadores, los datos nos son tratados como un vector de características, sino que se interpretan como datos que tienen cierta relación y estructura entre ellos. Un ejemplo de datos estructurados son las bases de datos relacionales.

En esta sección se abordarán algunos conceptos correspondientes a la fusión de modelos de inferencia basados en lógica con aquellos modelos que realizan inferencia basados en el manejo de incertidumbre. Los modelos basados en lógica siempre utilizan una base de conocimiento y generan reglas lógicas que permiten inferir conocimiento cuando se presenta una nueva instancia de ejemplo. Los modelos probabilistas por su parte construyen una distribución de probabilidad de las variables a partir de un conjunto de datos de entrenamiento. La fusión de los modelos lógicos con los probabilistas permite la construcción de modelos más robustos y expresivos. En la presente tesis se explora el uso de los modelos relacionales probabilistas para compararlos con el modelo propuesto, debido a que dicho modelo también trata de trasladar una representación estructurada (una gramática visual de una imagen) en un modelo gráfico probabilista que maneje incertidumbre.

En esta sección se aborda un bosquejo general acerca de los modelos relacionales probabilistas y los trabajos más cercanos que se han presentado al tratar de representar conocimiento estructurado, aunque no necesariamente de tipo visual.

3.3.1. Lógica de primer orden

Obsérvese los siguientes enunciados:

“La probabilidad de que una región (elegida aleatoriamente) de cualquier imagen de un conjunto sea un árbol, es igual a 0.7”.

"La probabilidad de que la región r de la imagen $\mathscr{I}$ sea un árbol, es igual a 0.7".

Notar que el primer enunciado se refiere a una distribución de probabilidad que se aplica a todo un conjunto de imágenes determinado, mientras que el segundo enunciado solamente hace referencia a una probabilidad sobre una imagen en particular (la imagen $\mathscr{I}$). Además, el segundo enunciado puede asumir un valor de verdad dependiendo del contexto: puede ser verdad para cierta imagen $\mathscr{I}$ y puede ser falso para otras imágenes.

Aquí la lógica de primer orden es útil para representar este tipo de enunciados. Si se considera que $P(r,\mathscr{I})$ representa la probabilidad de la región r en una imagen $\mathscr{I}$ dada, el primer enunciado puede ser rescrito como $\forall r \forall \mathscr{I} : P(r,\mathscr{I}) = 0.7$. Esta representación sólo es a manera de ejemplo y no está vinculada a algún formalismo. En las siguientes secciones se mencionan dos tipos de formalismos que tratan de combinar la lógica y la incertidumbre, uno de ellos más cercano al manejo de la incertidumbre (redes bayesianas relacionales) y otro más orientado a la lógica (redes lógicas de Markov). Es importante notar que todos los modelos relacionales, necesitan una "base de conocimiento", abreviada como *BC*. Esta base de conocimiento les permite construir el modelo. De manera análoga, un modelo de aprendizaje automático requiere un conjunto de datos de entrenamiento.

De manera más amplia, la base de conocimiento es un conjunto de enunciados o fórmulas escritas usualmente en lógica de primer orden [43]. Las fórmulas se construyen a partir de cuatro tipos de símbolos: *constantes*, *variables*, *funciones* y *predicados*. Las *constantes* representan siempre objetos o elementos del dominio de interés. A manera de ejemplo, si uno se interesa en un conjunto de imágenes segmentadas, las *constantes* serán las regiones o segmentos de dichas imágenes. Las *variables* son símbolos que actúan sobre los objetos del dominio. Las *funciones* proyectan tuplas de objetos a otros objetos (reciben un objeto de argumento y devuelven otro objeto). Finalmente los *predicados* representan relaciones entre dos o más objetos tal como $Amigos(Juan, Pedro)$ o bien, establecen un atributo a un objeto: $EsRojo(sombrero)$. Los *predicados* también admiten un valor de verdad: $Amigos(Juan, Pedro)$ puede devolver *verdadero* si realmente son amigos, o *falso*, en caso de que no lo sean. Cuando un *predicado* carece de variables (o que sólo contienen constan-

tes) se dice que está "aterrizado", de modo que también se les suele llamar predicados o "átomos aterrizados"[1].

Como dos bases de conocimientos distintas describen escenarios o situaciones distintas, se dice que cada base de conocimiento describe un *mundo*. De este modo, cuando se dice "todos los *mundos* posibles", se hace referencia a todas las posibles bases de conocimiento que se pueden construir con las constantes, variables, funciones y predicados antes mencionados.

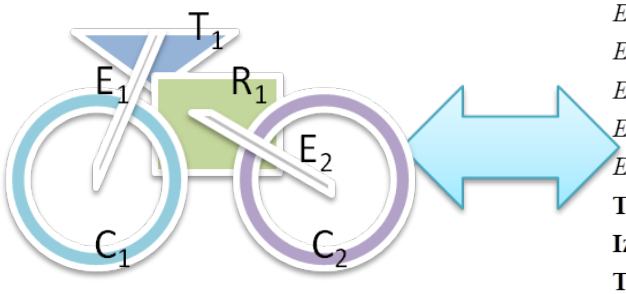
Adicionalmente, una $\mathscr{L}$-interpretación especifica cuáles símbolos representan las funciones, variables y predicados del dominio. Las variables usualmente tienen un *tipo*, es decir, sólo pueden admitir un subconjunto de los objetos del universo. A manera de ejemplo la variable *s* podría representar segmentos verdes o rojos en una serie de imágenes. A diferencia, una constante sólo hace referencia a una región específica.

Para los propósitos de la presente tesis, se busca representar relaciones espaciales de ciertas regiones, recuadros, píxeles o puntos de interés local a lo largo de un conjunto de imágenes. La idea es representar esta información de modo que se pueda construir una *BC*. Como no es trivial esta representación, el objetivo es representar una estructura aprendida a partir de una imagen (o a un objeto visual). Si se logra abstraer esta estructura de un objeto en términos de una *BC*, entonces es posible construir un modelo que aprenda a reconocer esta estructura en otras imágenes. Un ejemplo de esto se puede ilustrar en la Fig. 3.3.

3.3.2. Taxonomía de modelos relacionales probabilistas

Para situarse en contexto, hay una gran diversidad de modelos relacionales probabilistas. Algunos de los modelos pueden ser más útiles para representar información probabilista con extensiones de reglas lógicas o a la inversa. Otros modelos permiten establecer relaciones de dependencia condicional apoyándose en modelos gráficos dirigidos, mientras otros relajan esta restricción con modelos no dirigidos. Para simplificar un poco las cosas, se presenta una propuesta de taxonomía de algunos modelos relacionales y de ellos se han seleccionado dos para incorporarlos en los expe-

[1]El término proviene del inglés "ground atoms". Para darse una idea, un predicado aterrizado no está "flotando" entre un conjunto de variables que puede admitir.



$EsCuadro(R_1) \wedge EsVerde(R_1)$;
$EsTriangulo(T_2) \wedge EsAzul(T_2)$;
$EsBorde(E_1)$;
$EsBorde(E_2)$;
$EsCírculo(C_1)$;
$EsCírculo(C_2)$;
$\mathbf{Traslape}(R_1, E_2) \wedge \mathbf{Traslape}(R_1, C_1) \wedge \mathbf{Traslape}(R_1, C_2)$;
$\mathbf{Izquierda}(C_1, C_2)$;
$\mathbf{Traslape}(T_1, E_1) \wedge \mathbf{Traslape}(T_1, R_1)$;
$\mathbf{ArribaDe}(T_1, R_1)$;

Figura 3.3: Izquierda: representación visual de un objeto visual (bicicleta). Derecha: representación de la bicicleta usando fórmulas atómicas. *Traslape*, *Izquierda* y *ArribaDe* son relaciones espaciales (predicados) que devuelven un valor de verdad (*verdadero* / *falso*). Si un predicado aparece en la *BC*, se considera con valor de verdad *verdadero*. Si un modelo recibe diversos modelos de bicicletas visuales y los representa en una *BC* y se modela la incertidumbre asociada, se podría generar un modelo que aprende a "ver" bicicletas.

rimentos de la presente tesis. Esta clasificación está inspirada parcialmente en [45].

- Modelos relacionales vistos como una extensión de la lógica:
  - Evolucionan a modelos gráficos no dirigidos.
    - Redes lógicas de Markov. [99]
  - Evolucionan a modelos gráficos dirigidos:
    - Programas lógico bayesianos [61]
    - Redes lógico bayesianas [56]
- Modelos relacionales basados en una extensión de los modelos probabilistas:
  - Basados en modelos gráficos dirigidos
    - Redes bayesianas relacionales [54]
    - Modelos relacionales probabilistas [37]
  - Modelos Basados en modelos gráficos no dirigidos:
    - Redes relacionales de Markov [114]
    - Redes de dependencia relacional [84]

    - Campos aleatorios condicionales [44]

- Modelos relacionales probabilistas que extienden algún lenguaje de programación:

    - Programas lógico estocásticos [82]
    - Programación lógica inductiva probabilista [97]
    - Lógica bayesiana (BLOG) [78]
    - Lenguaje de modelado probabilista (IBAL) [92]

De todos ellos, se seleccionaron dos modelos para los experimentos: uno de ellos es una extensión o abstracción de las redes bayesianas y el otro es una extensión de los modelos basados en lógica. Aunque esta selección no fue rigurosa, se tomaron algunos aspectos como: i) un modelo (en apariencia) simple de describir, posiblemente basado en lógica, ii) un modelo que instancie redes bayesianas para compararlas con el método propuesto en la tesis iii) modelos que mediante predicados o reglas, describan una gramática de manera análoga. Para el último punto, los modelos que extienden lenguajes de programación parecen tener fines más ambiciosos, de modo que no fueron considerados. Para clarificar un poco esta idea, se presenta la Tabla 3.1 que considera ciertos aspectos de uso de estos modelos relacionales con respecto a las aplicaciones que se han desarrollado con ellos.

A continuación se muestran las definiciones para dos de estos tipos de modelos y se ven algunos trabajos que utilizan modelos relacionales en ciertas aplicaciones prácticas. No es muy usual aún encontrar esto en la literatura, pero el potencial a futuro es promisorio.

3.3.3. Redes Bayesianas Relacionales

Las redes bayesianas relacionales [54] (RBR) surgen como una manera de extender las redes bayesianas de manera que las variables involucradas puedan representar atributos y relaciones con otros objetos. Aunque las RBR tienen similitud con el trabajo de Friedman [37], las RBR permiten un uso más general desde que usan una *BC* y evitan acercarse a la nomenclatura común de las bases

Tabla 3.1: Aspectos generales de los modelos relacionales probabilistas.

MRP	Aspectos	Utilidad
Redes Lógicas de Markov	descripción simple con predicados, aplicaciónes en literatura	Alchemy
Programas lógico bayesianos	abarcan el área logica y de RBs	Balios
Redes lógico bayesianas	requieren definiciones típicas de un lenguaje de programación	ProbCog - ROS
Redes Bayesianas Relacionales	la definición de la RBR no es trivial	Primula
Modelos Relacionales Probabilistas	la notación es cercana a BDs	
Redes Relacionales de Markov	descubren información en documentos - Generalizan CRFs	
Redes de Dependencia relacional	potencial para extracción de información	
Campos aleatorios condicionales	potencial en descubrir información en documentos.	
Programas Lógico estocásticos	son cercanos a las gramáticas estocásticas	
Programación lógica Inductiva probabilista	requieren adaptarse a la ILP	
Logica Bayesianas (BLOG)	es una extensión a prolog	BLOG
lenguaje de modelado probabilista	es más un lenguaje de programación	IBAL

de datos (lo cual es más común en el trabajo de Friedman). Para el propósito de esta tesis, imágenes que contienen objetos visuales con estructura subyacente se pueden modelar en una *BC* (y por ende en una base de datos relacional), en donde dicha información puede ser descubierta por un modelo relacional. A continuación se describe el formalismo usado en las RBR.

Dado:

1. Un conjunto de símbolos relacionales S, llamados relaciones predefinidas (son predicados y devuelven un valor de verdad),

2. Un conjunto de símbolos relacionales R, llamados relaciones probabilistas (son predicados y devuelven un valor de probabilidad),

3. Un conjunto finito D, llamado dominio (los elementos u objetos a describir).

Se define una estructura S^D que es una interpretación de las relaciones S sobre el dominio D. Dicho de otra manera, S^D es una función que proyecta cada relación sobre el dominio $s(d), s \in S, d \subseteq D$

a valores de verdad (verdadero o falso)[2]. También se define un modelo de la Estructura Aleatoria-Relacional (*RRSM*[3]) como una función parcial, la cual recibe la estructura S^D como entrada y al compilarse devuelve una distribución de probabilidad sobre todas las estructuras R^D como salida (una red bayesiana). La función es parcial debido a que no todas las estructuras S^D de entrada que pueden existir en el universo pueden generar siempre estructuras R^D a partir del *RRSM*. En particular, esto ocurre cuando las redes bayesianas que deberían crearse son cíclicas, lo cual es un error. En la práctica cuando esto ocurre, se considera como un error de la definición del *RRSM* para la estructura S^D dada.

Antes de describir un ejemplo conviene explicar cada uno de los componentes del modelo relacional. Las relaciones predefinidas S, son relaciones en lógica de predicados que devuelven un valor de verdad para ciertas variables v que pertenecen a un domino D. Estas relaciones pueden ser unarias, binarias o de aridad n. Un ejemplo de una relación unaria podría ser $EsFruta(v)$. v es una variable de un dominio D. Si se consudera el ejemplo de dominio D siguiente: $D = \{manzana, gato, vaca, limón, Juan, Jorge, Pedro\}$. Entonces ocurre que $EsFruta(manzana)$ devolvería verdadero, $Esfruta(vaca)$ devolvería falso. $Persona(gato)$ devolvería falso y $Persona(Pedro)$, verdadero. Una relación binaria podría ser $Vecinos(v, w)$, donde v y w son variables que pertenecen al dominio D. $Vecinos(Juan, Jorge)$ podría ser verdadero, mientras que $Vecinos(Juan, manzana)$ devolvería falso. Todas las relaciones predefinidas, deben ser descritas explícitamente en el modelo, es decir qué variables pertenecientes al dominio D dan falso o verdadero para cada relación predefinida en S. Lo anterior constituye S^D.

Las relaciones probabilistas R por su parte, devuelven un valor de probabilidad a partir del estado que puedan admitir. Por ejemplo la relación unaria $EstaPodrida(manzana) = .05$ define que existe un 5 % de probabilidad de que una manzana esté podrida. La relación $Tiene(Juan, manzana) = 0.1$ define que existe un 10 % de probabilidad que Juan tenga una manzana. Nosotros podríamos tener evidencia de que una manzana está podrida si la tenemos con nosotros, sin embar-

[2]Notar que en este formalismo la estructura S^D es equivalente a la base de conocimiento *BC* mencionada anteriormente.

[3]Definido como Random-Relational Structure Model, de acuerdo a [54]

go, si carecemos de ésta, el modelo tiene un valor de probabilidad que se asigna *a priori*, o bien que haya sido aprendido a partir de ejemplos. El modelo permite "encadenar" estas relaciones probabilistas con otras e incluso, con relaciones predefinidas. Como ejemplo, se podría pensar que la distribución de probabilidad de la relación $Amigos(v, w)$ depende de si v y w son vecinos o no. En este sentido, se dice: $Amigos(v,w) : (Vecinos(v,w) : 0.7, 0.5)$. En otras palabras, existe un 0.7 de probabilidad que v y w sean amigos, dado que son vecinos, y 0.5 de probabilidad que sean amigos dado que no son vecinos. Para comprender la flexibilidad del modelo, *Vecinos* podría ser en su lugar una relación probabilista, definida como $Vecinos([Persona]v, [Persona]w) = 0.4$; es decir, la probabilidad que v y w sean vecinos es de 0.4. la notación $[Persona]v$ exige que la variable v sea verdadero para la relación predefinida *Persona*. Esto permite restringir el modelo a las variables que se quieran del dominio, o que satisfagan las relaciones predefinidas deseadas. Encadenar relaciones predefinidas y probabilistas constituye la Estructura Aleatoria-Relacional (*RRSM*) definida anteriormente.

De esta manera, a partir de relaciones predefinidas y probabilistas con sus definiciones sobre un dominio de variables, se puede obtener una estructura R^D sobre la cual es posible aplicar inferencia de tipo bayesiano. Notar que una red bayesiana podría abstraer esta información en un caso específico, pero no en todos los casos. Es decir, se debe tener una red bayesiana con ciertos parámetros cuando Juan y Pedro son vecinos, pero se debe tener otra red bayesiana con otros parámetros cuando Juan y Pedro no son vecinos. Como las redes bayesianas relacionales tratan con esta variación, permiten compilar la red bayesiana más adecuada según lo descubierto en el dominio S^D. Ejemplos de las estructuras *RRSM* y S^D se ilustran en las Figs. 3.4 y 3.5. El ejemplo de las redes bayesianas instanciadas (la estructura R^D) a partir del *RRSM* y el S^D de ejemplo, se muestran en la Fig. 3.6.

De esta manera, partiendo únicamente de las relaciones predefinidas y probabilistas, y describiendo la estructura de éstas, es posible compilar el modelo y obtener una red bayesiana sobre la cual se pueda realizar inferencia utilizando los métodos tradicionales que se aplican en las RBs.

Visto lo anterior, es posible sugerir una conexión entre las gramáticas relacionales y este tipo

```
        Estructura Aleatorio-Relacional
robo([persona]v)    = 0.005;
alarma([persona]v) = (robo(v):0.95, 0.01);
llamada([persona]v,[persona]w) =
(sformula(vecinos(v,w)):
        (sformula(bromista(v)):
                (alarma(w):0.9, 0.05),
                (alarma(w):0.9, 0)), 0);
alarmado([persona]v)=
        n-or{llamada(w, v)|w:vecinos(w, v)};
```

Figura 3.4: Ejemplo de un *RRSM*. En él se describen en cascada, las relaciones probabilistas entre los elementos v y w del dominio D. la definición puede involucrar diversas fórmulas, tanto probabilistas como predefinidas (en el ejemplo se denotan con el predicado *sformula).* Este ejemplo ilustra un caso hipotético de la probabilidad asociada a que una persona v se sienta preocupado de recibir una llamada de su vecino (o sus vecinos) indicándole que sonó la alarma de su casa. Se considera el caso de que el vecino sea un bromista y le llame sólo para molestarlo. Se modela que los robos en ese vecindario hipotético tienen una probabilidad de 0.5 % de ocurrir. También se modela la probabilidad de que la alarma de la casa suene dado que ha habido un robo (95 %) y se sugiere que hay un 1 % de que suene sin ocurrir (falso positivo). Este ejemplo ha sido tomado de [15].

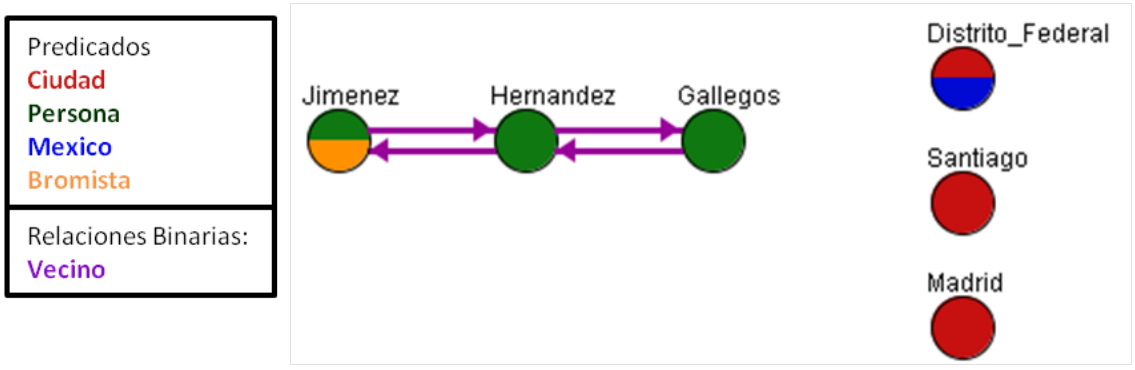


Figura 3.5: Descripción de S^D. Los círculos representan elementos del dominio D. Los atributos son relaciones unarias que pueden devolver verdadero o falso. $Bromista(Jimenez)$ y $Persona(Jimenez)$ devuelven verdadero, mientras $Persona(Madrid)$ devuelve falso. Las relaciones binarias se representan mediante flechas en el grafo. $Vecino(Jimenez, Hernandez)$ es verdadero, mientras $Vecino(Jimenez, Gallegos)$ es falso. Esta base de conocimiento es usada en la estructura aleatorio-relacional para instanciar redes bayesianas. Para obtener R^D, ambas estructuras son necesarias.

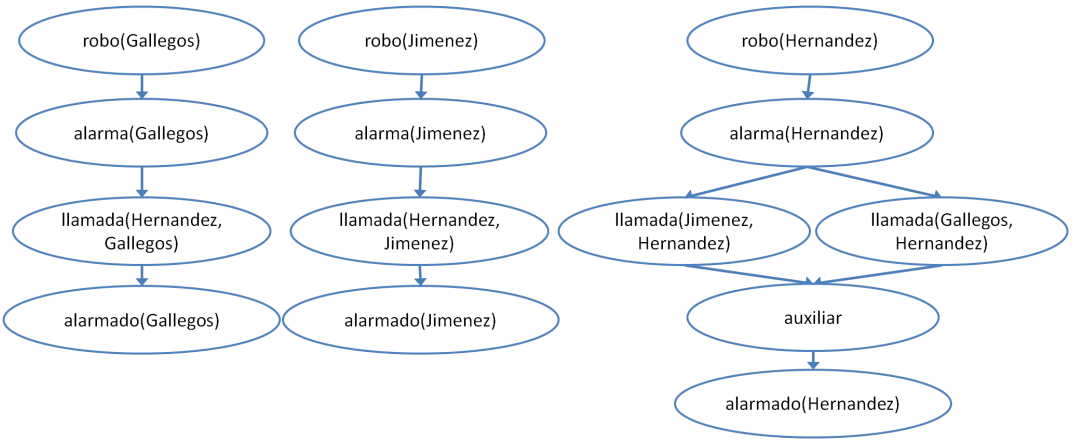


Figura 3.6: Ejemplo de la estructura R^D. Hay tres posibles redes bayesianas sobre las cuales se puede aplicar inferencia dado el dominio S^D. Como el dominio tiene tres personas, hay tres casos posibles de personas alarmadas. Esto da lugar a las tres redes bayesianas mostradas. Dado que Hernández tiene *dos* vecinos, su estructura es diferente que para Jimenes y para Gallegos: tiene dos maneras diferentes de informarse sobre el robo a partir de dos posibles llamadas telefónicas de sus dos vecinos. Por lo anterior, aparecen dos nodos de llamada. Las redes bayesianas relacionales añaden nodos auxiliares cuando se presentan situaciones opcionales. Como en los casos de Jimenez y Gallegos solo tienen un vecino, el nodo auxiliar no se utiliza. la relación *bromista*, dado que es predefinida y no probabilista, no aparece en las redes bayesianas.

de redes bayesianas relacionales podría darse mediante la lógica de predicados, lo cual resultaría en un modelo más expresivo y una vía más directa hacia la realización de inferencia. En esta tesis se plantea incorporar estos modelos relacionales mediante una transformación que construya una RBR a partir de una gramática relacional.

3.3.4. Redes Lógicas de Markov

En lógica, si una $\mathscr{L}$-interpretación[4] viola una sola fórmula dada en una *BC*, siempre tendrá probabilidad cero, de modo que ningún "mundo posible" debe contradecir la *BC*. En las redes lógicas de Markov (RLM) [99], esta restricción se relaja, de modo que cuando una interpretación contradice a la *BC*, sólo tiene una probabilidad menor (pero no cero) de ocurrir. Para lograr esto, en las RLM se añade un peso a cada fórmula. Pesos más altos indican que dicha fórmula castiga más una interpretación cuando no se satisface dicha fórmula. Una red lógica de Markov se define de la siguiente forma:

[4]Es decir, la definición de cuáles símbolos especifican las funciones, variables y predicados del dominio.

Definición: una RLM L es un conjunto de pares (F_i, w_i) donde F_i es una fórmula en lógica de primer orden y w_i es un número real (por ejemplo: $1.6 Vecino(v,w) \vee Amigos(v,w)$) . Dado un conjunto de constantes $C = \{c_1, c_2, ..., c_{|C|}\}$ que se pueden aplicar sobre cada una de las fórmulas F_i, se define una red de Markov (CAM) $M_{L,C}$ como:

- $M_{L,C}$ contiene un nodo binario por cada posible átomo aterrizado de cada fórmula que aparece en la RLM L. El valor del nodo es 1 si el átomo aterrizado es verdadero y 0 en caso contrario.

- $M_{L,C}$ contiene una característica por cada posible átomo aterrizado de cada fórmula F_i in L. El valor de esta característica es 1 si el átomo aterrizado es verdadero y 0 en caso contrario. El peso de la característica es el peso w_i asociado con la fórmula F_i en L.

Se forman cliqués entre los nodos cuando éstos aparecen en una fórmula F_i. Para mostrar como una RLM permite instanciar redes de Markov, se presenta el siguiente ejemplo:

sea una red lógica de Markov L formada por los siguientes dos pares de fórmulas pesadas:

$w_1 \forall x Fumar(x) \Rightarrow Cancer(x)$

$w_2 \forall x, y Amigos(x,y) \Rightarrow (Fumar(x) \Leftrightarrow Fumar(y))$

Para instanciar una red de Markov, a partir de la anterior RLM se necesita el conjunto $C = \{Ana, Bob\}$ (definido en una BC) que ayude a describir en forma de fórmulas aterrizadas los predicados de la RLM. El diagrama que instancia la anterior RLM se ilustra en la Fig. 3.7. Dado que $Fumar$ y *Cáncer* aparecen en la primera fórmula, se exige que estos nodos estén conectados cuando tienen el mismo argumento (sea Ana o Bob). $Fumar(A)$ y $Fumar(B)$ se conectan debido a que aparecen en la segunda fórmula del lado derecho. Cada par de $Fumar$ debe conectarse con el nodo $Amigos(x,y)$ pues así lo exige también la segunda fórmula. Una RLM no distingue cuando en las fórmulas los argumentos son equivalentes ($x = y$), de modo que ocurren nodos con argumentos iguales: $Amigos(A,A)$.

De maneja semejante a las RBR, las RLM son una generalización de las redes de Markov. Las RLM se pueden entender como una plantilla para construir o instanciar redes de Markov. En el proceso de inferencia, cuando se tiene un nuevo ejemplo a evaluar (representado a través de una

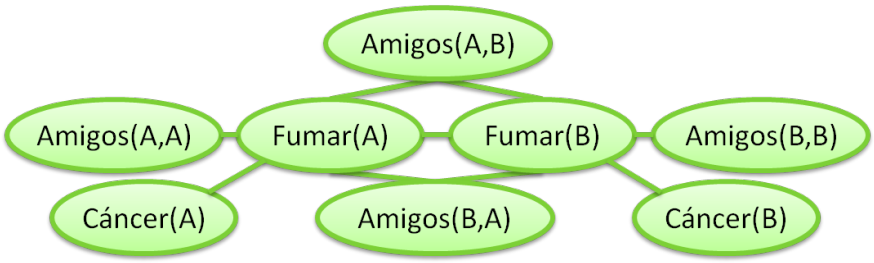


Figura 3.7: Instanciación de una red de Markov a partir de una red lógica de Markov. En este ejemplo los argumentos A y B pueden ser dos personas (por ejemplo, Ana y Bob) que se encontraron en el conjunto de constantes C. Las relaciones entre los nodos se construyen a partir de los predicados que aparecen unidos por un operador relacional en cada fórmula pesada de la RLM. Ejemplo tomado de [25] .

BC, de forma similar al S^D de las RBR), la RLM genera una instancia de red de Markov para ese ejemplo, y entonces la inferencia se puede hacer de manera similar que en una red de Markov. Haciendo una interpretación más general, las RLM tratan la inferencia como hallar la probabilidad de que una fórmula sea verdadera, dado que otros conjuntos de fórmulas lo son. Es decir, hallar:

$$P(F_1|F_2) = \frac{\sum_{x \in X_{F_1} \cap X_{F_2}} P(X = x)}{\sum_{x \in X_{F_2}} P(X = x)} \tag{3.7}$$

donde F_1 es la fórmula a inferir y F_2 es el conjunto de fórmulas que son verdaderas de acuerdo a la *BC*. Este procedimiento ocupa mucha memoria [99], puesto que se tienen que evaluar todos los posibles mundos donde las fórmulas son verdaderas. Explorar todos los posibles mundos (todas las posibles bases que satisfacen las fórmulas F_2) a partir del número de fórmulas dadas en una RLM, se sigue que dicha búsqueda es equivalente a un problema NP-completo [99]. Por ello, la inferencia en su lugar trata de aproximar la fórmula de la Ec. 3.7 mediante algoritmos de muestreo como Monte Carlo [46] o muestreo de Gibbs [41]. Para el caso de aprendizaje de la estructura de las fórmulas en una RLM, se puede usar cualquier enfoque de programación lógica inductiva [105, 81] para generar las fórmulas, aunque es deseable generar fórmulas de cualquier tipo, no solamente cláusulas de Horn. Como se ha visto, de manera análoga a las RBR, las RLM crean redes de Markov distintas según las constantes de *C* que aparezcan en la base de conocimientos

BC. A manera de comparación mientras en las RBR se expresan las relaciones o predicados a través de anidamientos, en RLM se expresan en enunciados pesados. Las RBR manejan siempre probabilidades y las RLM, pesos asociados a cada fórmula lógica en primer orden.

3.4. Aplicaciones con modelos relacionales probabilistas

Hay pocas aplicaciones prácticas usando modelos relacionales. En [57] se presenta un primer enfoque para utilizar las RBR en una aplicación que descubre comunidades en redes sociales. Se aprovecha el caso de que las redes sociales presentan diversos atributos y contienen múltiples relaciones entre cada variable. Las relaciones no son triviales, sino que los nodos tienen atributos haciendo que las relaciones entre ellos tengan sentido diferente: no es lo mismo la relación compañía-persona que una relación persona-persona. Las relaciones entre personas también tienen distinta dimensionalidad, pues pueden ser de amistad o de intereses comunes. Este conocimiento es altamente relacional, es decir, existe una diversidad de relaciones entre dos o más personas. Este método se probó en dos redes de comunidades, una simple y otra compleja, de tamaño pequeño. En las simples se tenía que encontrar la comunidad a la cual pertenecía un individuo de la red a partir de relaciones simples de amistad. En las redes complejas se incluyen relaciones más variadas, no necesariamente de amistad que pueden ser en sentido positivo o negativo. En [100] utilizan una RLM para un sistema de extracción de conocimiento para temas de biología en donde la RLM permite suavizar las reglas obtenidas del lenguaje para incrementar la cobertura de ejemplos. El objetivo es poder compactar la información en fórmulas lógicas pero evitando el enfoque duro de la programación lógica inductiva. Aunque estas aplicaciones no son propiamente del área de visión, ayudan a entender el potencial de los modelos relacionales para la extracción, representación e inferencia con información que tiene una naturaleza relacional.

En los trabajos de Snidaro [110, 109] se propone un modelo para utilizar redes lógicas de Markov en un modelo para un sistema de vigilancia de materiales peligrosos en buques de carga. Los autores buscan manejar datos de fuentes diversas, que además son posiblemente heterogéneos

y también tener en cuenta la posibilidad de que dichos datos sean incompletos. El procesar esta información implica manejo de incertidumbre. El manejo de los datos es a través de eventos simples y complejos (un evento complejo es la fusión de eventos simples o complejos) tales como llegada de buque, entrada del buque en determinada línea, declaración de material peligroso, etc. Si bien este trabajo no aborda modelos relacionales con imágenes, sí resulta un trabajo de aplicación de los modelos relacionales de manera más práctica, además de explotar los beneficios de combinar la expresividad de la lógica con el tratamiento de incertidumbre en el problema.

En Tran y Davis [118], los autores buscan reconocer eventos visuales en un entorno de vigilancia en video. Aquí primero se procesa una secuencia de video donde se elimina el fondo y se detectan objetos que se mueven, tales como personas y vehículos. Después de filtrar esta información, se crean predicados a partir de la localización de los mismos tales como "persona cerca de cajuela del vehículo" o "persona entra a vehículo" (escrito como $entrar(C_x, H_y)$) donde *entrar* es el predicado que indica a una persona H_x entrando al vehículo C_x. En sus pruebas, los autores muestran avances iniciales sobre como el modelo logra, conforme avanza el tiempo, mejorar sus predicciones según se va adquiriendo mayor información de evidencia. Aunque aquí no se utiliza la información cruda de las imágenes, los predicados generados sí provienen de una secuencia de imágenes. De manera implícita, algunos predicados sugieren relaciones espaciales entre los elementos involucrados. Estos últimos artículos [110, 109, 118] parecen tener de manera indirecta un mayor acercamiento a la representación de información visual, aunque la orientación sea principalmente a resolver una aplicación específica.

3.5. Resumen

En este capítulo se presentaron los aspectos teóricos relacionados con esta tesis. Se presentaron los modelos gráficos que se utilizarán: redes bayesianas, redes de Markov y sus extensiones en modelos relacionales probabilistas, que son las redes bayesianas relacionales y las redes lógicas de Markov. Se incluyó una taxonomía de estos modelos relacionales a manera de comparativa de

las características principales de estos modelos. Se comentaron también cuatro aplicaciones con los modelos relacionales a fin de esbozar el potencial a futuro de este tipo de modelos. En el siguiente capítulo se analizarán algunos trabajos que utilizan composición para realizar las tareas de reconocimiento de categorías de objetos.

Capítulo 4

Modelos Composicionales para Reconocimiento de Objetos

4.1. Introducción

En esta sección se analizarán los trabajos más representativos de modelos que utilizan una descomposición jerárquica para hacer reconocimiento de objetos. Se han dejado de lado los modelos orientados a reconocimiento de categorías de objetos donde la estrategia es clasificación sin alguna jerarquía clara. Hay un enfoque preferente en aquellos modelos que consideran la incorporación de gramáticas, a fin de evaluar sus ventajas y desventajas, así como de poder establecer algún criterio de comparación con dichos trabajos. Aunque es difícil establecer comparativas justas, se ha encontrado cierto consenso en el uso de repositorios de imágenes (como Caltech 256 [29]) y en particular, el uso de ciertas categorías de dicho repositorio (por ejemplo, motocicletas). Al final se establece una tabla de comparación que sintetiza las estrategias seguidas por cada enfoque a fin de discriminar las diferencias que tienen estos enfoques con el modelo presentado en esta tesis.

4.2. Enfoques que aprenden la estructura y partes de objetos.

Para efectos de acercarse más a los trabajos que estén vinculados en mayor medida con esta tesis, se ha decidido hacer una división más exhaustiva que la mostrada en la sección 2.3, de modo que se ahondara en el segundo enfoque, es decir, en los modelos basados en estructura y partes. Tomando inspiración de algunas divisiones sugeridas en [24], se crearon cuatro subdivisiones. Debe tenerse en cuenta que incluso en ocasiones es difícil determinar en qué subdivisión recae algún modelo y puede abarcar dos o más de las aquí presentadas.

1. **Modelos basados en clasificación de las partes.**

   Estos modelos son los más simples dentro de los modelos estructurados. Básicamente se centran en descomponer al objeto en elementos sencillos y en tratar de considerar la tarea como clasificación sobre estos elementos. Un inconveniente es que no es tan fácil modelar objetos que tengan presentaciones muy distintas, puesto que el modelo se reduce a determinar la presencia del objeto de acuerdo a la presencia o ausencia de las partes que lo componen, sin indicar explícitamente alguna relación o incluso variantes del objeto, como una mesa de cuatro patas o de una sola. Por lo regular este enfoque considera a los objetos de distintas presentaciones como categorías distintas.

2. **Modelos basados en la descripción de relaciones entre las partes.**

   Estos modelos se idearon para reconocer como categorías distintas a objetos que presentan partes similares o que son visualmente similares, pero lo que los hace distintos es el arreglo u organización presentado entre dichas partes. Un ejemplo muy simple es que tanto el agua como el cielo son regiones de un color azul muy similar pero la relación espacial que tienen con la tierra es distinta. Esta relación permite con facilidad distinguir que el cielo normalmente estará “arriba de” otros objetos (una montaña) y el agua por lo regular estará “debajo de” otros objetos (un barco). El reto común en estos modelos consiste en definir el conjunto de relaciones que mejor define al objeto.

3. **Modelos basados en métricas de la estructura de grafos.**

Estos modelos básicamente mapean una estructura de un objeto a un grafo compuesto de nodos como elementos simples y arcos como relaciones entre las partes. El objetivo consiste en encontrar una medida de similitud entre 2 grafos, para determinar si un objeto mapeado a un grafo coincide con el modelo del objeto de la categoría aprendida. Uno de los problemas que presentan estos modelos es que el isomorfismo de grafos es considerado un problema NP [40], por lo que si bien puede ser tratable en modelos pequeños, la variabilidad de estructuras que puede presentar un modelo podrían convertirlo en un problema difícil de computar.

4. **Modelos basados en gramáticas.**

   Estos enfoques parten de la idea de representar mediante el formalismo de una gramática la estructura de un objeto estableciendo la composición de elementos simples con otros para dar lugar a elementos más complejos hasta llegar a describir a un objeto. Alternativamente el modelo puede descomponer mediante reglas de producción propias de la gramática un elemento compuesto en otros elementos, más simples, llamados terminales. Las reglas de producción de la gramática permiten crear más de una instancia del objeto, cuando incorporan reglas de producción disyuntivas (de tipo *O*). Por lo regular, la mayoría de los modelos basados en gramáticas caen dentro de los modelos composicionales.

La subdivisión de modelos basadas en gramáticas es de mayor interés para esta tesis, puesto que la descomposición de elementos a partir de reglas dadas por una gramática sugiere la idea de objetos que tienen una estructura jerárquica que puede reconocerse desde elementos simples y sus relaciones. Aunque esta idea es compartida con los enfoques de relaciones entre las partes y de métricas de la estructura de grafos, la diferencia con los modelos de relaciones es que cada parte puede tener propiedades, y la diferencia con los modelos basados en grafos es que un objeto puede ser descrito con reglas de producción, generando variantes del objeto (o diferentes explicaciones del objeto) creando una representación del objeto más compacta. Después de haber realizado esta división en la siguiente sección se definirán qué son los modelos composicionales, al estar muy ligados a los modelos basados en gramáticas.

4.3. Modelos composicionales

Los modelos *composicionales* parten de una idea estructural. La composicionalidad[1] se refiere a la habilidad del ser humano de representar entidades u objetos, como una jerarquía de partes. Por ejemplo, un rostro se puede ver como una jerarquía de partes más simples, como la nariz, o un ojo. A su vez, un ojo se puede ver como una jerarquía de partes más sencillas, como iris o pupila. De la misma manera, una pupila se puede ver como un círculo con cierta textura. Una jerarquía es un modelo donde los elementos u objetos tienen una estructura de grafo dirigido acíclico (GDA). De esta manera, se conoce bien que el hombre tiende a realizar una composición de lo que percibe del mundo, lo que ve, oye, habla e incluso lo que piensa. La interpretación de la *composicionalidad*, normalmente se expresa de abajo hacia arriba. Partiendo de cosas simples, el hombre construye elementos más complejos. Como ejemplo, se tiene el lenguaje: a partir de letras, construye palabras, y a partir de palabras, construye frases. Sin embargo, esta idea de composicionalidad es posible interpretarla también de arriba hacia abajo: cuando pensamos y analizamos de que está hecho algo, se busca descomponerlo en partes más sencillas, a fin de lograr el entendimiento. Cuando la composicionalidad es interpretada de arriba hacia abajo, suele llamarse también modelo jerárquico.

Hoy en día sabemos que la composicionalidad, de acuerdo a los trabajos que inició Noam Chomsky [16, 17], resultó esencial para comprender el lenguaje humano. Este enfoque de composicionalidad se describió formalmente mediante gramáticas. Las gramáticas en textos se pueden ver como relaciones entre elementos en una dimensión (los elementos aparecen solamente a la izquierda o a la derecha de otros). Los trabajos que relacionan esta composicionalidad con el análisis de imágenes para tareas de reconocimiento de objetos, parten de la idea de que es posible aplicar un enfoque similar de composición de elementos simples para construir objetos complejos (como el ejemplo anterior del rostro). El enfoque ahora utilizado extiende el concepto a dos dimensiones, pues los elementos no solamente pueden aparecer a la izquierda o la derecha, sino arriba o abajo también. En este sentido, los trabajos que se analizan a continuación, se encuentran vinculados a

[1]Para ver un enfoque más filosófico acerca del principio de composicionalidad referirse a: `http://plato.stanford.edu/entries/compositionality/`

esta idea.

4.4. Trabajos en modelos composicionales

Es términos generales, los trabajos composicionales no tienen una división clara sobre el método que usan, ya que lo que tienen en común estos trabajos es el enfoque de composición en alguna parte de sus modelos, mientras que el resto puede ser bastante diverso. A manera de tratar de dividir de alguna manera los trabajos previos, se propone una separación (no estricta) en tres pequeños bloques: trabajos basados en una jerarquía, trabajos que ocupan alguna gramática como ayuda en la composición, y trabajos que utilizan un grafo como ayuda para representar la jerarquía.

4.4.1. Trabajos apoyados en una jerarquía.

Estos trabajos definen una especie de jerarquía que puede ser descrita manualmente o generada al momento de aprender el modelo.

Trabajos de Ales Leonardis. Los trabajos de Leonardis [36, 34, 35] están concentrados en la elaboración de un diccionario visual jerárquico de manera automática, a partir de familias de filtros Gabor. Su enfoque parte de construir los elementos más simples del diccionario entrenándose con imágenes de objetos naturales, a fin de obtener los grupos de bordes a cierta orientación más comunes. El objetivo es tener un amplio diccionario visual a distintos niveles de composición (elementos de bajo y alto nivel) a fin de lograr aprendizaje usando algún clasificador con los elementos de más alto nivel que fueron detectados. El algoritmo es relativamente rápido para procesar una imagen y se comparan usando la base de datos Caltech [29] para categorías de objetos. En particular, la principal aportación de estos trabajos es que construyen un modelo jerárquico en donde hay comunicación a diferentes niveles entre los elementos que componen dicha jerarquía. No obstante, un inconveniente de esta conceptualización, es que no es claro el nivel de granularidad que tendrán los elementos de alto nivel para realizar tareas de reconocimiento. Adicionalmente, incorporar evi-

dencia a distintos niveles resulta inconveniente en términos de representación del conocimiento: la noción de composición en el modelo es más compleja. Un diagrama de este modelo se ilustra en la Fig. 4.1.

Trabajos de Joachim Buhmann. Los trabajos de Buhmann [87, 88, 89, 90] presentan un modelo composicional para tareas de categorización de objetos. Estos trabajos parten de obtener regiones de las imágenes, empleando extracción de características de diversa índole (bordes, descripción de puntos característicos con histogramas locales, etc.) desde alguna base de datos de imágenes determinada. Posteriormente se forma un diccionario que se construye agrupando mediante *k-medias* los elementos terminales más similares. Este proceso se repite para construir composiciones de composiciones. Posteriormente se construye una pequeña red bayesiana que permite hacer inferencia a partir de un entrenamiento con estas características encontradas. Un inconveniente de este modelo es que la construcción de las características están orientadas a mejorar la tasa de reconocimiento, antes que la construcción de un modelo jerárquico. Adicionalmente, aunque se trata de un modelo jerárquico, el mecanismo de inferencia no tiene una estructura análoga a la jerarquía obtenida entre los objetos simples y su composición. Hay relaciones espaciales de adyacencia, aunque éstas se compactan en un nodo de la red bayesianas de estructura fija. Un diagrama de este modelo se ilustra en la Fig. 4.2.

Trabajos de Maximilian Riesenhuber. Riesenhuber [101, 107] presenta en sus trabajos un modelo jerárquico lo más apegado posible a un modelo bioinspirado. Este modelo, llamado HMAX, logra realizar tareas de reconocimiento de objetos inspirado en una jerarquía de neuronas encargadas del proceso de visión en los primates [51]. Estas neuronas son selectivas a la orientación de la luz, el modelo puede construirse con neuronas tipo Simple (S) y tipo Complejo (C). A partir de los elementos de luminosidad por orientación, los cuales se emulan por un filtro de Gabor [39], y construyendo hacia arriba elementos más complejos se pueden realizar las tareas de reconocimiento con ayuda de algún clasificador en la última fase. Sin embargo, por lo anterior, deja de lado la idea de representar el conocimiento de una manera más clara o expresiva y no considera la incorporación

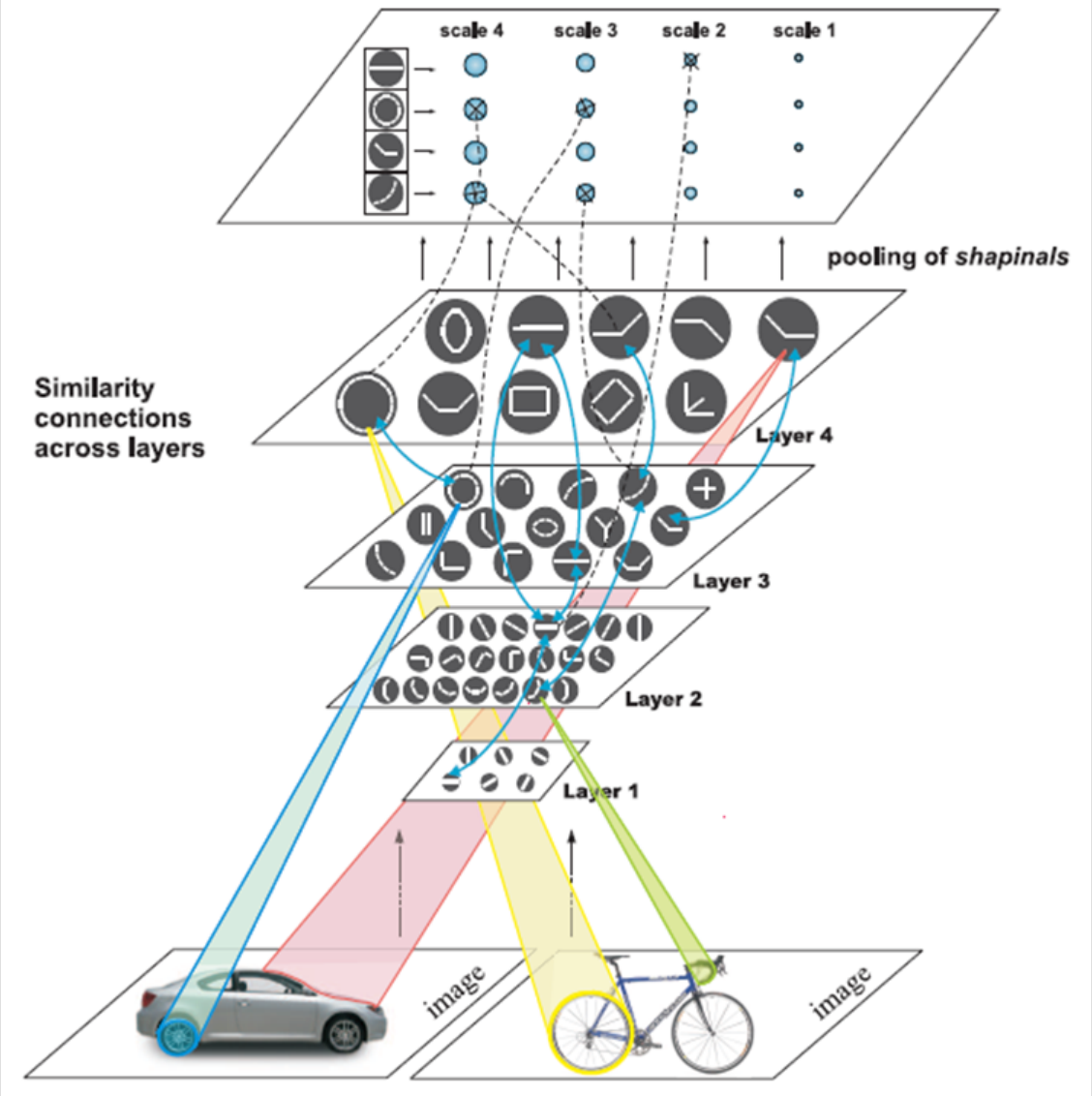


Figura 4.1: Modelo propuesto por [35]. Los elementos detectados en la imagen pueden ser pasados a distintos niveles del modelo jerárquico. Adicionalmente los elementos de cada capa se comunican entre sí e incluso con otras capas (flechas entre capas), a fin de ganar en invarianza. Al final los elementos de más alto nivel (capa superior) que hayan sido detectados son pasados a un clasificador para realizar el reconocimiento de categorías de objetos.

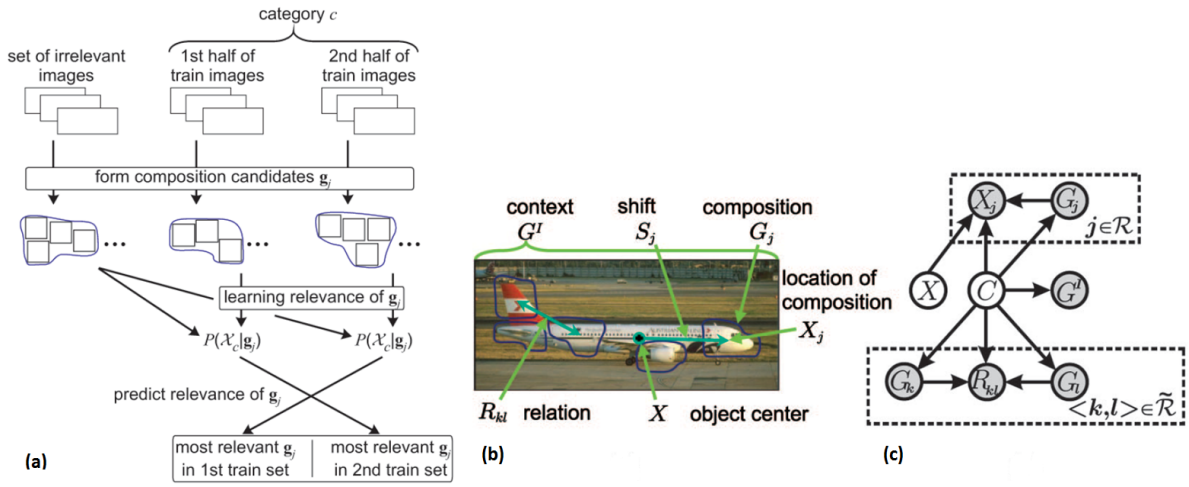


Figura 4.2: (a) Modelo de Ales Leonardis, et-al [90]. Este modelo parte de elementos terminales (átomos) obtenidos de conjuntos de imágenes de entrenamiento tanto relevantes como no relevantes a la categoría que se desea aprender. Posteriormente componen grupos de átomos y se seleccionan aquellos que resulten más relevantes en el entrenamiento. En (b) se incorporan algunos tipos de relaciones más información del contexto (como el fondo) y esta información es evaluada en una red bayesiana de estructura fija (c), que permite hacer inferencia para reconocer la categoría deseada.

de alguna gramática o formalismo que permita describir su proceso de aprendizaje o inferencia. Un diagrama de las capas del modelo se muestra en la Fig. 4.3.

Trabajos de Sinisa Todorovic. El enfoque en los trabajos de Sinisa Todorovic [115, 1, 117, 116] consiste en crear un modelo jerárquico orientado a la segmentación. Una de las ventajas obtenidas es que se tiene un modelo que puede aprender diversas categorías visuales (escenas, aunque no de objetos), así como realizar este reconocimiento en condiciones de oclusión. Un ejemplo de los niveles jerárquicos en la segmentación se muestra en la Fig. 4.4. Aunque este modelo presenta interés en la tarea de segmentación jerárquica, y logra establecer correspondencia del modelo aprendido en otras imágenes, carece de una formalización inicial que describa los elementos a usar o que ayude a una mayor expresividad del modelo. La estructura de inferencia es similar a una red bayesiana, realizada en un árbol con propiedad markoviana. Esto quizás influye en las tasas de reconocimiento, con una tendencia a la aparición de falsos positivos. Por otra parte, la inclusión de reconocimiento multiclase puede afectar la idea de composicionalidad.

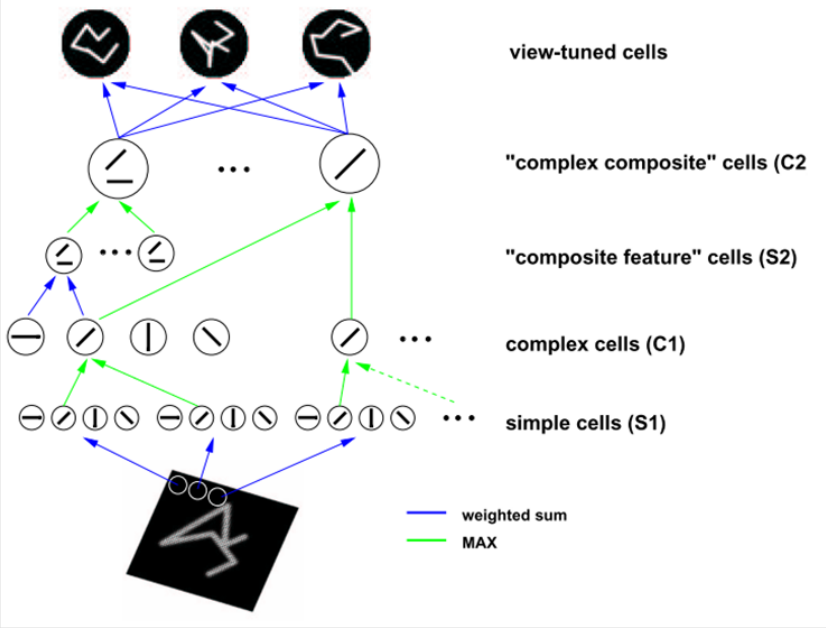


Figura 4.3: Modelo jerárquico bioinspirado propuesto por Riesenhuber, et-al [101]. El modelo se construye a partir de terminales S1 que son sensibles a orientaciones presentadas en una imagen. El modelo compone elementos más complejos en las etapas superiores hasta tener elementos no terminales o *células* que pasan a un clasificador para su entrenamiento. Entre capas S y C el modelo alterna entre sumas pesadas y operadores Max que devuelven el mayor estímulo encontrado.

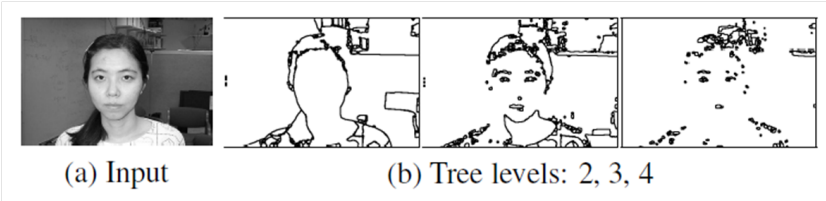


Figura 4.4: En [115] presentan un modelo jerárquico a distintos niveles para aprender escenas. En la imagen se muestran distintos niveles aprendidos, cada nivel aprende segmentos a diferente granularidad.

4.4.2. Trabajos apoyados en grafos

Estos trabajos representan con ayuda de algún grafo la estructura del conocimiento aprendido.

Trabajos de Song Zhu. Los trabajos de Zhu [128, 123, 95, 67, 122] plantean una especie de gramática a través de una representación de grafos *And-Or*. En esta representación, los nodos *Or* apuntan hacia subconfiguraciones alternativas y los nodos *And* son descompuestos en un número determinado de elementos. En estos trabajos se incorpora un diccionario visual, que es descrito por un número de primitivas visuales que al unirlas forman una representación de la imagen. Las primitivas usadas están basadas en esquinas y formas geométricas. Un ejemplo de esto se ilustra en la Fig. 4.5. En estos trabajos, aunque realizan reconocimiento de objetos, se enfocan más en la construcción de la gramática. Uno de los aspectos más interesantes de estos trabajos, es que para la construcción de los elementos terminales que componen el alfabeto visual emplearon ciertas reglas conocidas del agrupamiento perceptual [68]. Si bien estos trabajos son amplios en cuanto a la construcción de una gramática, sus mecanismos de inferencia están basados en cadenas de Markov incorporando ciertas diferencias para adaptarlas al modelo, en el mismo artículo reportan que esta técnica tiene una tendencia a falsos positivos. Para ayudar a reducir la incidencia de falsos positivos, los autores realizan inferencia tanto de abajo hacia arriba como de arriba hacia abajo. En este trabajo la incorporación de evidencia se realiza a varios niveles, de modo que los elementos terminales pueden ser de bajo, medio o alto nivel. Entre sus debilidades, no hay una incorporación explícita de relaciones espaciales entre los elementos que componen a un objeto, en su lugar se describen las relaciones horizontales, dadas entre elementos terminales o no terminales que son adyacentes, al encontrarse en un mismo nivel de la jerarquía construida. (Fig. 4.5).

4.4.3. Trabajos apoyados en gramáticas

Estos trabajos se apoyan en alguna gramática (de tipo transformacional o relacional) y utilizan sus reglas de producción para representar el conocimiento aprendido.

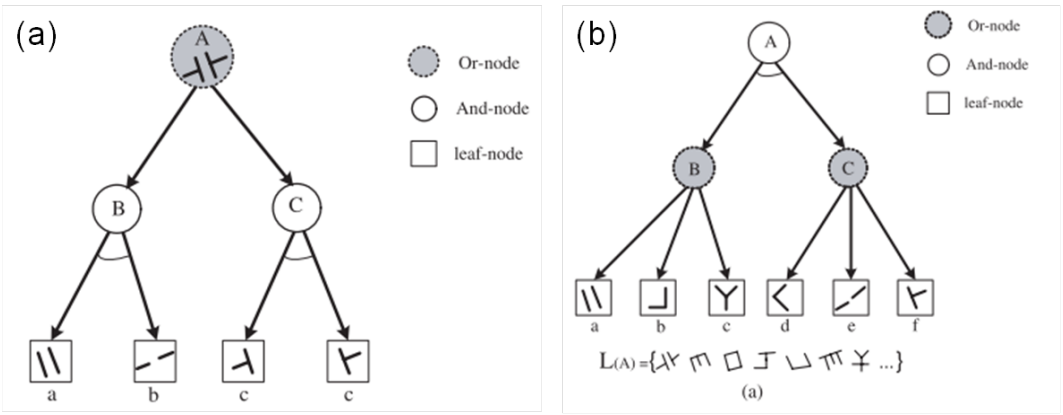


Figura 4.5: Modelo del Grafo *And-Or* propuesto por Zhu, et-al [128]. (a) En la imagen los nodos hoja son terminales de un alfabeto visual. Emplean uniones T, L, barras y regiones homogéneas principalmente. Se obtiene un nuevo nodo no terminal. (b) Combinando nodos *Or* y *And* se obtiene un conjunto de configuraciones que puede adoptar el nodo A. Dichas configuraciones se denominan *Lenguaje* del nodo A.

Trabajos de Stuart Geman. Stuart Geman ha trabajado en el desarrollo de gramáticas de tipo composicional [42, 7, 14] y en el desarrollo de alfabetos visuales para estos modelos [50, 125]. Las gramáticas propuestas buscan describir como se van componiendo objetos simples en una imagen y como van guardando ciertas relaciones de tipo espacial (por ejemplo, líneas cuyas relaciones son dadas con ángulos de referencia o distancias). En estos trabajos, se desarrolla una gramática de tipo composicional que es utilizada en tareas de reconocimiento de objetos, como caracteres alfanuméricos. Estos trabajos parten de una estructura definida de elementos iniciales, como líneas y formas simples de conjuntos de píxeles, los cuales componen el alfabeto visual. En algunos de sus trabajos [58, 59], incorpora un enfoque bayesiano, parecido a una red bayesiana. Sin embargo, el modelo construido no parte del formalismo de una gramática. El enfoque bayesiano está basado en construir un árbol donde cada nodo es un bloque que representan terminales en las hojas y no terminales en el resto. Aquí el alfabeto construido está orientado al reconocimiento de partes de placas de automóvil. Un aspecto discutible en este trabajo es que el alfabeto es construido de acuerdo a la aplicación y puede presentar diversos niveles de detección de elementos terminales (partes de caracteres, letras, uniones T o L, barras de píxeles, conjuntos de varios caracteres, etc.), lo cual implica que las estructuras formadas son heterogéneas. Una vez detectadas estas regiones, se propagan probabilidades sobre el modelo construido. Un diagrama de este modelo se ilustra en

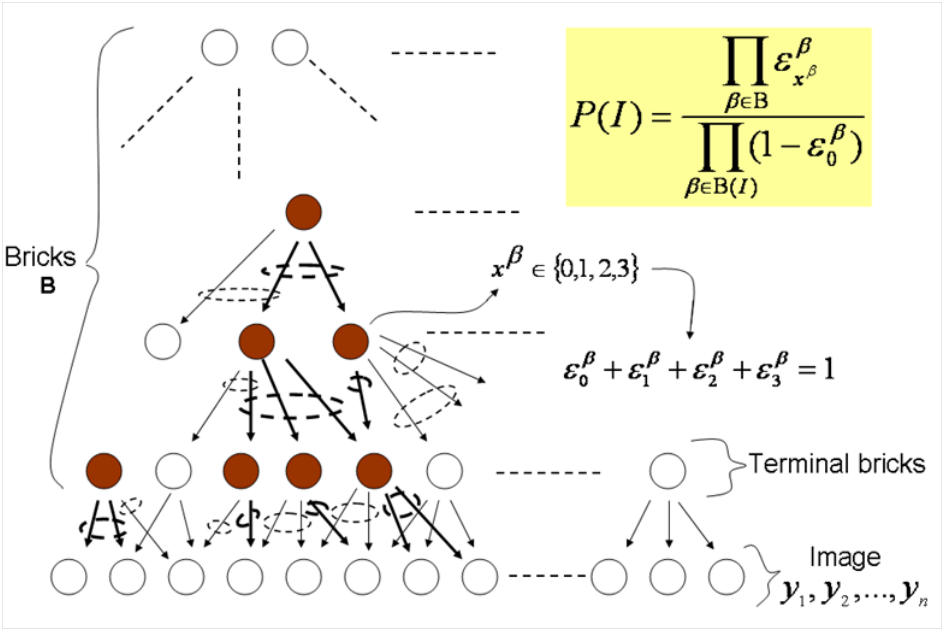


Figura 4.6: Modelo propuesto por Geman, et-al [59]. Los bloques se descomponen hasta llegar a bloques terminales que son detectados en la imagen. Estos bloques pueden ser barras, bordes, formas e incluso regiones de imágenes. Los bloques alcanzan a ser cientos de miles en el modelo. Cada bloque suele tener muchos estados x^β, tantos como combinaciones de hijos activos tengan. La inferencia se realiza a partir de la evidencia de bloques terminales en la imagen. Imagen tomada de [59].

la Fig. 4.6.

Trabajos de Pedro Felzenszwalb. Los trabajos de Felzenszwalb [30, 48] proponen una gramática acíclica genérica en conjunción con una segmentación jerárquica (de tipo árbol) a n niveles de las imágenes. El objetivo es combinar la gramática con los elementos de la imagen para tareas de reconocimiento de objetos. Los autores han probado el modelo en reconocimiento de personas en condiciones de oclusión. La inferencia se realiza mediante un algoritmo de programación dinámica que calcula una puntuación para los elementos encontrados en la imagen. Aunque este modelo tiene una base jerárquica dada por la gramática, el modelo no incorpora relaciones espaciales explícitas, en su lugar definen propiedades de posición en la imagen para los elementos descritos, así como el nivel de composición de dicho elemento (desde bordes con pocos píxeles hasta regiones grandes). Este modelo requiere describir manualmente la gramática. Un ejemplo de los elementos que ocupa esta gramática para reconocer personas se ilustra en la Fig. 4.7.

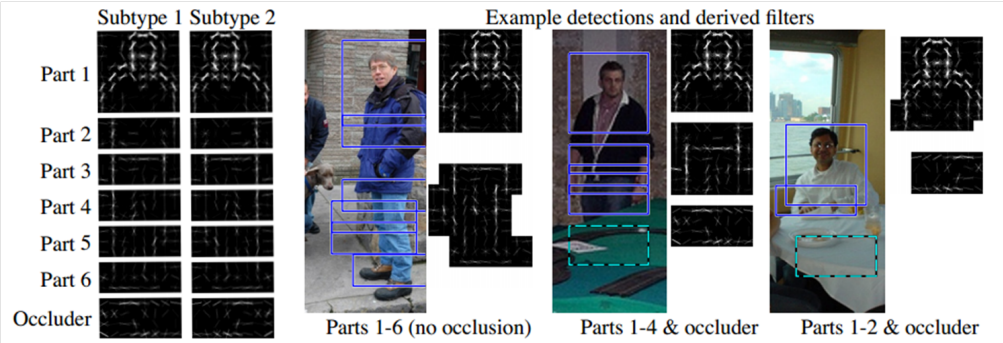


Figura 4.7: Detección de elementos en imágenes [48]. Los recuadros negros denominados subtipo 1 y 2 son los elementos terminales de la gramática. Estos elementos son detectados en las imágenes (recuadros azules en las fotos). Se definen relaciones espaciales a partir de la posición entre unos y otros mediante las coordenadas x,y de la imagen. En la figura se ilustran ejemplos sin oclusión (partes 1-6 sin oclusión) donde pueden reconocerse todos los elementos terminales y también se muestran ejemplos con oclusión (partes 1-4 y partes 1-2 con oclusión). Aquí se detectó el elemento terminal de oclusión sugiriendo que se halló una persona pero no de cuerpo completo. La inferencia se realiza en la gramática mediante el valor máximo de una métrica aplicada en la gramática para los elementos encontrados. Ver mejor en color.

Trabajos de Meléndez. Meléndez [76, 75] presenta un modelo que parte de gramáticas SR para después, de manera manual, convertirlo a un modelo gráfico probabilista (una red bayesiana). Este modelo aprovecha la descripción de la gramática relacional para incorporar relaciones de tipo espacial. El modelo fue aplicado en dominios específicos (rostros de personas) y los elementos terminales obtenidos fueron de alto nivel (por ejemplo, detectores de ojos). Una de las ventajas del modelo es la incorporación de las gramáticas SR que permiten mediante lógica de predicados definir explícitamente las relaciones espaciales usadas. Este trabajo sirvió de base para considerar este tipo de gramáticas y construir un modelo más flexible que permita aplicarse a diversos dominios. En particular, se observaron los siguientes aspectos a mejorar: i) obtener automáticamente un modelo gráfico probabilista a partir de la gramática visual, ii) considerar elementos simples de más bajo nivel, iii) tener un diccionario visual más flexible y, iv) evitar la descripción manual de la gramática mediante un aprendizaje de la misma a partir de ejemplos.

Un diagrama de la red bayesiana usada en este trabajo para realizar inferencia y reconocer rostros se ilustra en la Fig. 4.8.

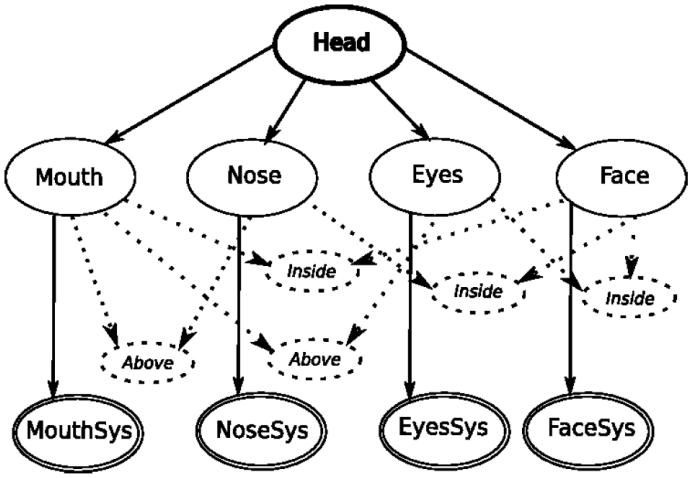


Figura 4.8: Red bayesiana construida a partir de un modelo basado en gramáticas SR. Óvalos en líneas dobles representan los detectores de elementos terminales. Óvalos punteados representan las relaciones espaciales consideradas. La inferencia es realizada sobre la misma red previo entrenamiento a partir de ejemplos. Imagen tomada de [76].

4.4.4. Otros Trabajos

Existen algunos otros trabajos que si bien también caen dentro de los modelos estructurados, su relación con las gramáticas no es tan clara, o bien se enfocan no en construir un modelo de reconocimiento de objetos, sino en una parte de éste. En [124] no aprenden ni definen alguna gramática, aunque consideran la incorporación de un diccionario visual a partir de familias de filtros Gabor que resulta flexible para la tarea que desarrollan. La inferencia se realiza mediante un campo aleatorio de Markov. En [104] la inferencia se realiza mediante optimización evolutiva. Al igual que en el trabajo anterior, los autores no consideran algún enfoque basado en gramáticas. Este enfoque al estar bioinspirado en la corteza visual su objetivo está orientado a construir un modelo adaptado al sistema biológico de la visión. En [127], los autores proponen aprender una gramática de tipo restringido (sólo contiene producciones de tipo *Or*) a partir de tripletas de puntos SIFT. Su inferencia está basada en un CAM. Ninguno de los tres trabajos anteriores considera incorporar relaciones espaciales. En [126] proponen un modelo para aprender estructuras de cierta profundidad. En particular, en el artículo se interesan por analizar la profundidad de las estructuras aprendidas. En este mismo artículo consideran relaciones espaciales restringidas a las coordenadas de los elementos terminales encontrados. Aunque este trabajo guarda una mayor cercanía con los objetivos de es-

Tabla 4.1: Resumen de trabajos relacionados. En general las características más relevantes de los trabajos analizados se pueden agrupar en siete columnas: si incorporan un diccionario visual, si incluyen algún tipo de gramática, si aprenden la gramática, los métodos para realizar inferencia, si utilizan relaciones espaciales explícitas y si se inspiran en algunas estrategias para construir el modelo, como las reglas de agrupamiento perceptual [68], o si es un modelo bioinspirado en la visión [51].

	Diccionario Visual	Tipo de gramática	Aprende gramática	Inferencia	Agrupamiento perceptual	Bioinspirado	Relaciones Espaciales
Stuart Geman	Manual	Composicional	Parcial	Similar a RB	Parcial	No	Parcial
Zhu	Manual	Grafo And-Or	Parcial	Markov	Sí	Parcial	Parcial
Felzenszwalb	Manual	Genérica y Acíclica	No	Métrica	No	No	Posición
Leonardis	Fam-Gabor	No	N/A	Clasificación	No	Sí	No
Buhmann	Autom.	Bolsa de partes	No	RB simple	Sí	No	Parcial
Riesenhuber	Patches Auto	No	N/A	Clasificación	No	Sí	No
Todorovic	Segmentos	Niveles Jerárquicos	Parcial	Métrica DAG	No	No	Parcial
Meléndez	Detectores Débiles	Gramática SR	No	RB	No	No	Sí
Wu	Gabor	N/A	N/A	CAM	N/A	No	No
Ramanan	Recuadros	Transformacional	Parcial	CYK	No	No	Temporal
Tesis	Recuadros	Gramática SR / TR	Sí	RB/RBR/RLM/reglas	Parcial	No	Sí

ta tesis, por sus autores este trabajo también puede estar situado como una extensión alternativa a los trabajos de Zhu y otros [128]. En el trabajo de [93] proponen el uso de una gramática por segmentos, que de alguna forma es una gramática temporal de tipo transformacional. Esta gramática, aunque tiene el objetivo de poder operar sobre vídeos para reconocer acciones, no permite la inclusión de relaciones espaciales, puesto que parte de la idea de que pequeñas sub-acciones se han reconocido previamente. En otras palabras, el elemento más pequeño de la gramática es una sub-acción compuesta de uno o varios cuadros. En contraste, en esta tesis se busca representar un cuadro en términos de sus elementos y su disposición espacial.

4.4.5. Discusión

Dentro de los trabajos analizados, se pueden resumir algunos aspectos de los mismos en la Tabla 4.1. Todos se vinculan con la presentación de un modelo jerárquico y hay alta coincidencia en atacar el problema de inferencia bajo un enfoque bayesiano. También es requerida la descripción de un alfabeto visual, mediante el cual, partiendo de elementos primitivos se puedan ir "componiendo" elementos más complejos. Donde no hay un acuerdo claro es en el enfoque bayesiano a utilizar ni en el alfabeto a definir. Una idea es que para poder preservar la idea de composicionalidad, se debe tener un alfabeto simple, que permita construir con facilidad "expresiones visuales" más complejas. Sería deseable además, que estos elementos primitivos puedan describirse incluyendo sus relaciones espaciales, con claridad y precisión. En el caso del manejo de la incertidumbre esta tesis considera que para lograr una mejor expresividad deben evaluarse los modelos basados en en modelos gráficos probabilistas [62] y con modelos relacionales probabilistas [54, 99, 45], así como sus mecanismos de aprendizaje [55], a fin de buscar una mayor expresividad y un mecanismo de inferencia más adecuado. Por otro lado, también se considera en esta tesis que si se busca construir un modelo que sea expresivo de una representación composicional, una opción consiste en emplear alguna gramática visual que permita realizar este modelado. Esto conlleva a considerar un aprendizaje automático de la misma, puesto que los modelos que presentan gramáticas, aún no consideran la generación automática de éstas, o bien lo hacen con ciertas restricciones.

A partir de los trabajos encontrados, se puede concluir que si bien hay una cantidad muy vasta de trabajos sobre reconocimiento de objetos, aquellos que incluyen un enfoque composicional son pocos, aunque recientemente de cada vez un mayor interés. Este interés resulta de la posibilidad de modelar de manera estructurada componentes simples para obtener estructuras cada vez más complejas. Por otra parte, los modelos que involucran relaciones espaciales también están presentando un aumento, debido a que permiten reducir el exceso de falsos positivos en modelos que se limitan a la detección de características sin considerar una organización o disposición de dichas características encontradas. Algo que se ha observado reiteradamente es el uso de modelos gráficos probabilistas para realizar inferencia y tratar con la incertidumbre propia del reconocimiento de

objetos. Por la parte de las gramáticas si bien se ha visto el interés manifiesto en trabajos previos de incorporarlas, en algunos casos su uso es similar a una caja negra sobre la que el modelo trabaja sin posibilidad de ofrecer un entendimiento para el usuario, puesto que es usual que la descripción quede en un grafo o en una estructura tipo árbol de miles de nodos. No se encontraron trabajos aún que involucren modelos de tipo relacional para reconocimiento de objetos, quizás por su orientación a caracterizar modelos de bases de datos relacionales [37].

Una vez analizado el trabajo relacionado, se pueden detallar las características de esta tesis que permiten diferenciarla del trabajo relacionado. Primeramente la propuesta para construir un modelo expresivo es incorporar gramáticas Simbólico-Relacionales [33], las cuales a diferencia de las gramáticas libres de contexto, incorporan relaciones espaciales mediante lógica de predicados. Estas gramáticas han sido utilizadas para expresar diagramas de flujo en ingeniería de software, describiendo bloques como elementos no terminales y relaciones entre éstos como flechas de adyacencia. Trasladar las gramáticas para describir imágenes es una tarea análoga, puesto que los bloques no terminales son componentes de la imagen (por ejemplo, bordes, regiones homogéneas, esquinas, puntos de interés local, descriptores de forma, etc.) y las relaciones que tienen estos elementos podrían ser de tipo espacial (por ejemplo, *arribaDe*, *dentroDe*). Al igual que con otras gramáticas se provee un enfoque jerárquico. Posteriormente se plantea la definición de un diccionario visual de manera manual empleando algún algoritmo de segmentación y otro basado en recuadros. Este alfabeto visual debe poder utilizarse de manera directa en la gramática, de modo que se propone un *lexicón visual*, que describirá cada elemento de la gramática en términos de dicho alfabeto visual. Dicho de otra manera, el lexicón permite caracterizar cada elemento terminal que utilice la gramática a partir del algoritmo que extraiga la información visual (o diccionario). Esto aportará flexibilidad al modelo, dado que otros modelos de trabajos relacionados están restringidos a un tipo específico de diccionario. En el caso de esta tesis, un conjunto de entrenamiento permitirá crear este lexicón visual, de modo que el lexicón se adapta al objeto que se desea reconocer. En la parte del algoritmo de inferencia, se propone un MGP generado automáticamente a partir de la gramática. En particular se proponen tres casos para comparar: las redes bayesianas [91], las redes

bayesianas relacionales [54] y las redes lógicas de Markov [99], las cuales mediante una descripción de relaciones booleanas y probabilistas de objetos, permiten construir un modelo bayesiano sobre el cual se puede realizar inferencia. Al final, se muestra un análisis comparativo entre cada uno de estos enfoques. Resumiendo las diferencias del modelo propuesto con otros enfoques son:

1. Uso de gramáticas y relaciones espaciales que dan mayor expresividad al modelo. Otros trabajos consideran gramáticas que al no incorporar relaciones, dejan de lado el uso de relaciones espaciales, o éstas se integran en una fase posterior.

2. El uso de un diccionario visual que permita construir un "lexicón" generado automáticamente. El objetivo es tener cierta homogeneidad en la definición de los elementos terminales que contenga la gramática. Los trabajos anteriores por lo regular carecen de la definición del lexicón.

3. La inferencia al realizarse con modelos que manejen incertidumbre permitirá una tratamiento a las cuestiones de oclusión y a suavizar las reglas de producción de la gramática. La comparación con los modelos relacionales probabilistas para observar las potencialidades de los mismos en áreas de representación de conocimiento visual no se ha realizado antes. Para el caso de inclusión de relaciones temporales se incluyó una primera aproximación de inferencia usando reglas.

4. Una propuesta de extensión de gramáticas visuales hacia gramáticas que incorporen explícitamente las relaciones temporales. El uso explícito es para poder dar un tratamiento separado a las relaciones espaciales y a las temporales. Este enfoque es una alternativa más flexible a las gramáticas basadas en segmentos temporales [93].

Con lo anterior la propuesta de tesis queda definida en sus puntos principales. En el siguiente capítulo se detallará el modelo general propuesto en esta tesis. Este modelo toma en cuenta los enfoques estudiados en los trabajos relacionados y su ubica dentro de un área no explorada previamente: la representación de conocimiento visual mediante gramáticas visuales y su posterior aprendizaje

e inferencia incorporando incertidumbre vía modelos gráficos probabilistas para poder mostrar el potencial de reconocer objetos aprendidos con esta representación.

Capítulo 5

Modelo de Representación Visual de Objetos basado en Gramáticas Visuales y Redes Bayesianas

5.1. Introducción

En este capítulo se dan los detalles de la construcción de nuestro modelo para poder realizar tareas de representación visual y reconocimiento de categorías de objetos. El modelo propuesto se puede dividir en dos etapas, una etapa de entrenamiento y una etapa de inferencia. Un diagrama general de la etapa de entrenamiento del modelo propuesto se encuentra en la Fig. 5.1. La gramática visual se describe en la etapa de entrenamiento, mientras que en la etapa de inferencia, ilustrada en la Fig. 5.2, se realiza la tarea de reconocimiento en los conjuntos de imágenes con el MGP aprendido. En las siguientes secciones se detallan las fases que conforman la metodología del modelo propuesto.

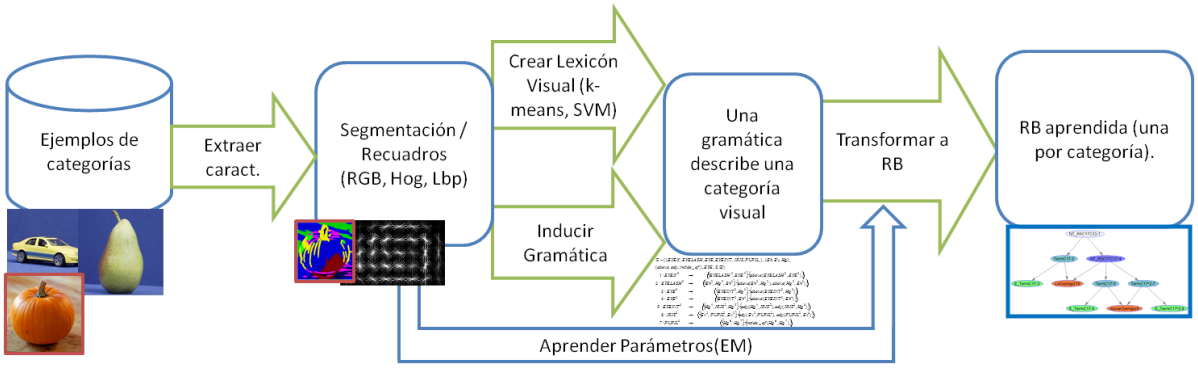


Figura 5.1: Diagrama general de la etapa de entrenamiento del modelo. A partir de un conjunto de imágenes de entrenamiento, se segmentan (o generan recuadros) describiendo sus características. Se genera un Lexicón Visual (manual o automático) que permite describir los elementos terminales de la gramática. La gramática se construye o se aprende automáticamente a partir de ejemplos y se transforma a una red bayesiana a fin de poder realizar inferencia con manejo de incertidumbre.

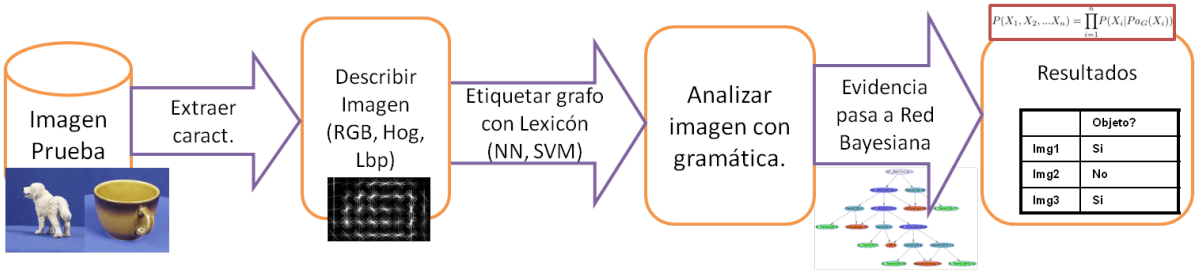


Figura 5.2: Etapa de Inferencia del modelo. Dada una nueva imagen de consulta se describe en características con el diccionario visual y se realizan las correspondencias de las regiones o recuadros encontrados con el Lexicón visual. A partir de ahí se utiliza un algoritmo que evalúa aquellos subconjuntos de regiones o recuadros de la imagen con sus relaciones espaciales que son candidatos para ser evaluados en el MGP. Se realiza inferencia sobre el modelo y se obtienen pruebas del mismo. Las pruebas determinan si hay o no el objeto aprendido por la gramática en la imagen.

5.2. Alfabeto Visual

Partiendo de las imágenes es como se construye el modelo de reconocimiento de objetos. Para ello, es necesario una etapa de procesamiento que represente de alguna forma el conocimiento que se encuentra en las imágenes. Si se desea trabajar con gramáticas simbólico-relacionales, se requiere trasladar la información visual de las imágenes en información que sea procesable por las gramáticas y para ello se utiliza un alfabeto visual. El alfabeto visual a desarrollar debe cumplir con ciertas características como:

- Que resulte adaptable al dominio.
- Que tenga un grado de invarianza a cambios en iluminación.
- Que compacte la información visual y tenga poder discriminativo.

Aunque es difícil cumplir los tres, buscar un consenso puede ser una opción. Tener un alfabeto visual adaptable al dominio permite al alfabeto extraer información pertinente al dominio de aplicación, evitando recuperar ruido. La invarianza a la iluminación permite darle cierta independencia con respecto de las cámaras usadas para tomar las imágenes. Compactar la información visual permite que el entrenamiento y la inferencia puedan ser más rápidos. Un alfabeto que trabaje con píxeles aislados tiene poco poder para compactar la información visual. Compactar en exceso le reducirá su poder para discriminar regiones distintas de una imagen. En este trabajo se han diseñado tres tipos diferentes de alfabetos visuales. A partir cada uno de ellos se pueden crear “palabras visuales” que conformen un lexicón para alguna gramática que reconozca alguna categoría de objetos, tal como “árboles” o “rostros”. Para comprender el concepto de lexicón, dos categorías de objetos distintas normalmente no comparten el mismo lexicón, puesto que sus palabras visuales no suelen ser parecidas. Dado que se busca un modelo que aprenda automáticamente la gramática a partir de ejemplos, es de esperarse que el aprendizaje del lexicón más apropiado para cada categoría también sea automático. De los tres algoritmos propuestos para generar los alfabetos visuales, uno de ellos está basado en segmentación, otro está basado en recuadros y otro más que involucra un

algoritmo de aprendizaje automático para escoger los recuadros más representativos en las imágenes. Aunque es posible utilizar aún más descriptores de imágenes y más alfabetos, no es el objetivo realizar una revisión exhaustiva de los mismos, pues dicha tarea se saldría de los objetivos de esta tesis. En su lugar se estudia el funcionamiento de un modelo composicional a partir de gramáticas que ayude en las tareas de reconocimiento. A continuación se describe cada alfabeto visual.

5.2.1. Alfabeto visual basado en segmentación.

Este alfabeto visual fue creado a partir de una segmentación simple con ideas parcialmente bioinspiradas y está basado en detección de bordes y regiones homogéneas. Este alfabeto utiliza orientaciones de bordes regiones homogéneas basadas en características de color sobre los canales RGB. Éste fue el primer enfoque probado en el modelo, y es relativamente sencillo en su poder de descripción. La idea es que si un modelo basado en gramáticas visuales que tenga elementos mínimos en su diccionario visual logra realizar algunas tareas sencillas de reconocimiento (objetos con pocos bordes), entonces al enriquecer este diccionario visual se podrán lograr tareas de reconocimiento con objetos más elaborados. Para la detección de bordes, se utilizó el filtro Gabor [39], el cual es sensible a orientación. Se describen a continuación los pasos para conseguir este alfabeto.

5.2.1.1. Análisis de bordes y regiones homogéneas

Se encontró en trabajos relacionados que algunos modelos de visión que son bioinspirados [113, 38, 101] utilizan el Filtro Gabor [39] para obtener bordes en diversas orientaciones. El filtro Gabor está definido por:

$$F(u_1,u_2) = e^{(\frac{v_1^2+\gamma^2 v_2^2}{2\sigma^2})}\cos(\frac{2\pi}{\lambda}v_1), \tag{5.1}$$

donde:

$$v_1 = u_1 \cos\theta + u_2 \sin\theta \tag{5.2}$$

$$v_2 = -u_1 \sin\theta + u_2 \cos\theta, \tag{5.3}$$

(u_1, u_2) son valores del filtro en el rango $[-\eta, \eta]$, donde η es el tamaño de la ventana del filtro en el sistema coordenado, θ es la dirección, γ es una constante de aspecto, σ es la amplitud efectiva dependiente de la escala, λ es la longitud de onda dependiente de la escala. La escala del filtro Gabor define el tamaño de η y también define a σ y a λ. En la literatura se utilizan de dos a cuatro orientaciones (0°, 45°, 90° y 135°). Los bordes que se encuentren pasan a formar parte de elementos de bajo nivel dentro de la gramática.

La contraparte de los elementos de bajo nivel, son las zonas homogéneas. Estas zonas son extraídas mediante una cuantización a 32 colores de una imagen dada. A dicha cuantización, se procede a eliminar las zonas pequeñas mediante un algoritmo en el que adopta el color de su vecino más grande.

Para poder reconocer las zonas de cambio por orientación se utilizaron ventanas deslizantes que detectan cambios de intensidades de píxel, predefinidos para las orientaciones usadas. Dichos ventanas son equivalentes a las características Haar [66, 120]. Al operar las ventanas sobre la imagen filtrada con Gabor, se generan zonas sensibles a las orientaciones buscadas. El resto serán regiones homogéneas. Desde un enfoque bioinspirado, es un camino alternativo a los filtros Max [101].

Después de ello se fusionan las orientaciones mediante un filtro de suma lógica (operador lógico o) que incluye limpieza de ruido mediante supresión de puntos no máximos (SNM[1]). Un ejemplo sencillo de píxeles etiquetados se ilustra en la figura 5.3. Después de analizar los bordes se procesan las regiones homogéneas. Estas regiones se cuantizan a 32 colores (cinco bits) sobre los canales RGB. Las imágenes de entrada pasan de una representación de 24 a 5 bits, donde se incluyen los 11 colores básicos[2] del lenguaje propuestos por [6]. Aunque se pueden usar otros colores,

[1] Supresión de No Máximos.

[2] Estos colores básicos son: blanco, negro, rojo, verde, amarillo, azul, café, violeta, rosa, naranja y gris.

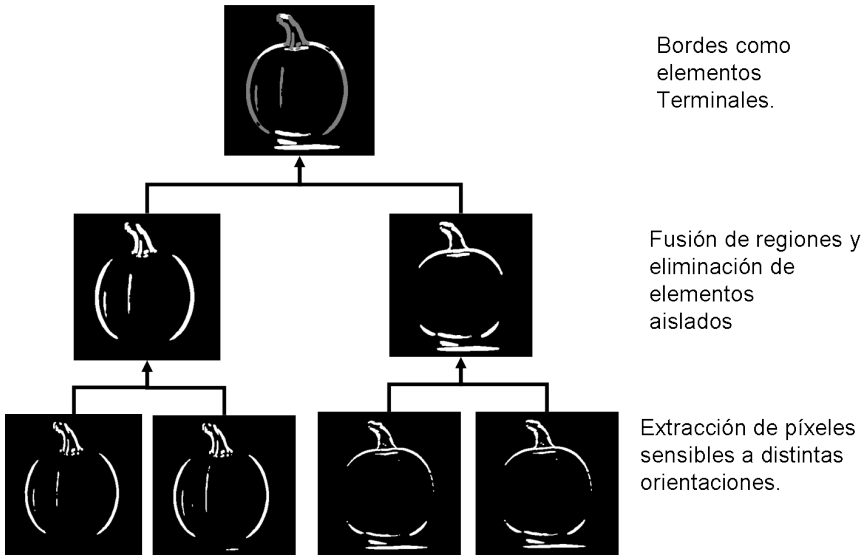


Figura 5.3: Los diversos filtros que extraen los píxeles etiquetados, se fusionan para obtener regiones más gruesas. Una vez fusionados, pasan por un proceso de filtrado para eliminar regiones pequeñas y componer las regiones que serán útiles como elementos terminales.

Figura 5.4: Imagen antes y después de la cuantización a 32 colores y eliminación de regiones pequeñas.

se decidió considerar estos colores debido a que fueron los primeros en ser discriminados por el hombre a partir del lenguaje hablado [6]. Un ejemplo de esta cuantización se ilustra en la Fig. 5.4.

5.2.1.2. Reglas de agrupamiento perceptual

En algunas ocasiones no es claro si el píxel está correctamente etiquetado (si pertenece a una orientación o a otra). Se renombran píxeles con vecinos más cercanos para casos de píxeles aislados. Esto sugirió la consideración de algunas reglas de agrupamiento perceptual [68], pues en ocasiones los bordes obtenidos resultan discontinuos, provocando más regiones de las que deberían ser. En este sentido, si la distancia entre dos regiones con igual etiqueta es inferior a un umbral ($d_{r1,r2} <$

ε_{px}), entonces dichas regiones se fusionan siempre que pertenezcan a la misma orientación. Los huecos en regiones no etiquetadas también se rellenarán. Esto acorde con las reglas de proximidad y continuidad del agrupamiento perceptual [68]. Lo anterior solamente busca reducir el número de regiones discontinuas y cercanas entre sí, para que pertenezcan a un mismo tipo. Estas reglas son aplicadas tanto a bordes como a regiones homogéneas. Finalmente las relaciones espaciales que puede haber entre estas regiones se restringieron a tres básicas: $IzquierdaDe$, $ArribaDe$, $DentroDe$. En los tres casos se exige adyacencia entre las regiones consideradas. Al exigir adyacencia, esta restricción buscó considerar objetos visuales no fragmentados.

5.2.2. Alfabeto visual basado en recuadros

Este algoritmo extrae recuadros de manera uniforme en la imagen, a manera de una estrategia distinta de un algoritmo de segmentación. Aunque los algoritmos de segmentación ayudan a separar regiones borde de regiones homogéneas, suelen perder información tal como la textura en los objetos. En algunos artículos [30, 108], la idea de recuadros es preferida que el uso de segmentación debido a que los recuadros capturan más información de apariencia del objeto, mientras que por el contrario en la segmentación suele ser necesario adaptar la granularidad de las regiones al dominio. Para los propósitos de esta tesis, nuevamente se busca explorar el caso más sencillo del manejo de recuadros, puesto que se desea estudiar el comportamiento del modelo composicional.

Este algoritmo obtiene recuadros a partir de una rejilla con cierto deslizamiento, ello permite que entre dos recuadros cualesquiera existan relaciones espaciales de adyacencia, traslapamiento o ubicación (arriba de, izquierda de, y traslapado a 45°). Cada recuadro se describió utilizando Histograma de Gradiente (HoG) [22] y Patrones Locales Binarios (LBP) [85]. Este tipo de descripción aunque no es tan morfológica (puesto que siempre son recuadros) que la descripción anterior de alfabeto visual basada en segmentación, captura una mayor información de apariencia. También este modelo considera más relaciones espaciales: $ArribaDe$, $IzquierdaDe$, $IzqTraslapado$, $ArribaDeTraslapado$, $Izq\&ArribaTraslapado$, $Izq\&DebajoTrasl pado$. Ejemplos de estas relaciones espaciales se ilustran en la Fig. 5.5.

Para poder generar un alfabeto se parte de recuadros etiquetados de manera positiva si provienen de una imagen que contiene al objeto (aunque no necesariamente el objeto está en el recuadro) y negativos si provienen de imágenes que no contienen al objeto. Con lo anterior se sigue el siguiente algoritmo:

1. Describir recuadros en términos de sus características. Cada recuadro p_i es un vector de la forma $p_i = [v_1, v_2, ...v_{|p_i|}]$. Las $|p_i|$ características están dadas por HoG y LBP.

2. Separar en conjuntos U_p y U_n (recuadros positivos y negativos) los recuadros obtenidos de imágenes de entrenamiento. Entre dos recuadros puede haber un traslape hasta de un 50 %. Si existe información de localidad (recuadro que determine dónde está el objeto en la imagen) se añade un peso a cada $p_i \in U_p \cup U_n$, con peso $v_{|p_i|+1} = 0$ cuando el recuadro no contiene al objeto, $v_{|p_i|+1} = 0.5$ cuando el recuadro p_i se encuentra en la frontera que envuelve al objeto y $v_{|p_i|+1} = 1$ en caso de que el recuadro p_i esté totalmente en el objeto de interés.

3. Agrupar los recuadros similares usando k-medias con k fijo en cada iteración.

4. Para cada grupo determinar el poder discriminativo mediante la fórmula:
   $d = \text{máx}\left(\frac{Freq_p(k_i)}{Freq_n(k_i)+1}, \frac{Freq_n(k_i)}{Freq_p(k_i)+1}\right)$,
   donde $Freq$ devuelve la frecuencia de recuadros asociados al grupo k_i (p cuando estén en U_p y n cuando estén en U_n) .

5. Retener los grupos que superen el umbral $d > |\mathscr{I}|/f$. f es un entero que se define de acuerdo a la prueba realizada. Repetir el paso 3 con los recuadros restantes hasta que no se puedan retener más grupos.

El objetivo del método es quedarse con centroides de aquellos grupos que sus elementos pertenecen en mayor medida al objeto o a la categoría negativa (ausencia del objeto). Los centroides que han quedado conforman el alfabeto visual. Es de notarse que, a diferencia del algoritmo de regiones homogéneas y bordes, este alfabeto se crea de acuerdo a la categoría a reconocer, al igual que el

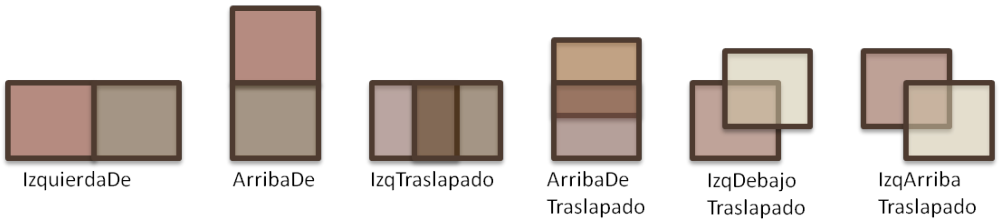


Figura 5.5: Seis tipos de relaciones espaciales utilizadas en el enfoque basado en recuadros. Son dos relaciones espaciales sin considerar traslape y cuatro más considerando traslape.

lexicón (en el algoritmo de regiones homogéneas el alfabeto es independiente de los datos). En ambos casos el lexicón se creará según los datos.

Este método puede describirse como un alfabeto visual obtenido mediante técnicas de agrupación. En la siguiente sección se detalla un método que en su lugar obtiene el alfabeto visual mediante técnicas de aprendizaje automático.

5.2.3. Alfabeto visual basado en recuadros con aprendizaje automático.

Este método surge debido a que el algoritmo que sólo está basado en obtener recuadros de las imágenes de entrenamiento y agruparlos produce con facilidad recuadros que no describen al objeto o que no discriminan correctamente entre imágenes positivas o negativas de un ejemplo dado. En adición a ello, en fechas recientes el uso de modelos basados en apariencia en combinación con estrategias de aprendizaje automático en enfoques supervisados e incluso no supervisados han tenido más éxito en tareas de categorización que los modelos basados en segmentación. Algunos trabajos tales como [71, 108, 60] aprenden los recuadros incorporando estrategias de aprendizaje automático. En este sentido el método realizado es una estrategia similar a los trabajos mencionados, pero adaptada para que la salida del mismo sea un alfabeto visual.

Este algoritmo es una variante del algoritmo *k-medias* que incorpora clasificación de los recuadros de manera iterativa. El objetivo es que se entrenan n clasificadores que se van afinando y haciendo selectivos a cierto tipo de recuadros. Cada clasificador emite una probabilidad para cada recuadro en la imagen. Ejemplos de recuadros recuperados para una palabra visual se ilustran en la Fig. 5.6. Ejemplos de clasificadores o palabras de lexicón ya entrenados para usarse en la gramáti-

ca se ilustran en la Fig. 5.7. El modelo basado en gramáticas aprenderá con estas palabras visuales incorporando composición entre reglas. Este trabajo es una contribución adicional que hasta el momento se diferencia de otros trabajos de la revisión bibliográfica. Se presenta a continuación el algoritmo desarrollado:

1. Describir recuadros en términos de sus características, cada recuadro p_i es un vector de la forma $p_i = [v_1, v_2, ...v_{|p_i|}]$, Las $|p_i|$ características están dadas por HoG y la posición relativa del recuadro en la imagen.

2. Inicializar recuadros de manera aleatoria y formar k grupos usando *k-medias*.

3. Entrenar un clasificador (SVM o naïve bayes por ejemplo) por cada grupo formado.

4. Se prueba cada clasificador contra todos los recuadros de las imágenes y se re-entrena usando los recuadros con mayor y menor puntuación (para aumentar el poder discriminativo).

5. Se prueba cada nuevo clasificador contra todos los recuadros de entrenamiento, de modo que cada clasificador tenga en su poder los recuadros más parecidos (se usa clasificación en lugar de agrupamiento).

6. Se transforma el valor que entregue un clasificador a probabilidad (para el caso de SVMs se usa [94]).

7. Si hay clasificadores que retienen recuadros con baja probabilidad o pocos recuadros, se eliminan.

8. Se añaden recuadros positivos y negativos de otros clasificadores a cada clasificador para reducir el sobreajuste y se vuelve a entrenar. Si un clasificador tiene un número alto de recuadros asociados se divide en dos para evitar clasificadores desbalanceados.

9. Se repiten los pasos 5, 6 y 7.

10. Se refinan cada uno de los recuadros asociados a cada clasificador moviéndolos unos pocos píxeles cada recuadro de manera que mejore la predicción del clasificador.

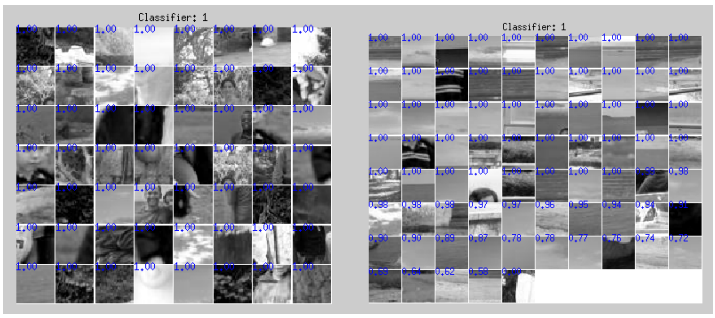


Figura 5.6: Recuadros reconocidos por un clasificador con el algoritmo basado en aprendizaje automático. Este algoritmo es iterativo de modo que puede mejorar en cada iteración los tipos de recuadros que puede reconocer cada clasificador. En la imagen se observan los recuadros detectados para un mismo clasificador en la primera y décima iteración.

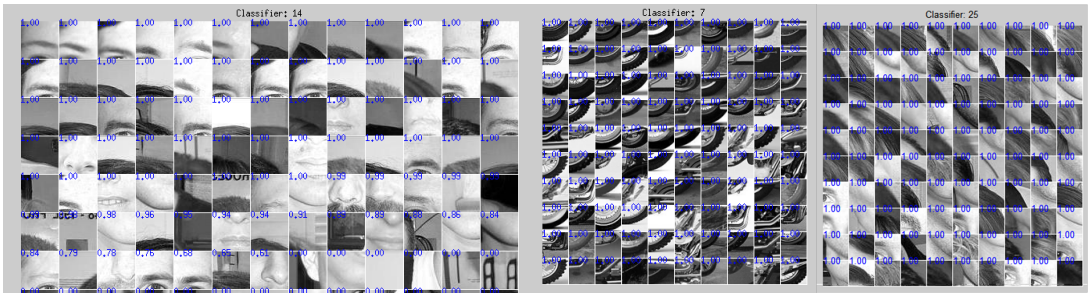


Figura 5.7: Ejemplo de recuadros recuperados por un clasificador como palabra visual. Cada clasificador equivale a una palabra visual del lexicón y a un potencial elemento terminal en la gramática.

11. Se repite desde el paso 8 hasta tener convergencia en los clasificadores o se alcance un número máximo de iteraciones.

Se tomaron todos los clasificadores obtenidos por el algoritmo como elementos del alfabeto visual. Al igual que en el algoritmo anterior, dado que se busca generar el alfabeto visual de forma automática, este alfabeto visual depende de los datos de entrenamiento. En este sentido, aunque se pierde generalidad del alfabeto, hay una automatización de la tarea, evitando la intervención del usuario.

5.3. Descripción de la Gramática Simbólico Relacional

Dado que se incorporan relaciones espaciales entre los elementos dados por un alfabeto, se define primeramente a la relación espacial como predicados de aridad dos, donde el nombre indica la

relación del primer argumento con respecto de segundo. Así, relaciones como A dentro de B, A adyacente de B o A arriba de B son representadas como: $DentroDe(A,B)$, $Ady(A,B)$, $ArribaDe(A,B)$, donde $A,B \in V_N \cup V_T$. Por ejemplo la relación del tronco y el follaje de un árbol se puede representar como:

$$ArribaDe(follaje, tronco),\ follaje \in V_N,\ tronco \in V_N$$

En el presente trabajo se decidió trabajar con las gramáticas Simbólico-Relacionales [33], (descritas en la sección 2.5) dado que permiten describir estas relaciones espaciales de la misma manera, mediante lógica de predicados. Con ello, se adaptan al requerimiento de especificar relaciones espaciales entre regiones locales de una imagen, obtenidas mediante el alfabeto visual. Sin embargo, si se desea transformar un modelo dado por la gramática a un MGP, se debe evitar que la gramática tenga producciones cíclicas, lo cual lleva a incorporar restricciones a la gramática SR.

5.3.1. Restricciones en la Gramática.

De acuerdo a la especificación de las gramáticas SR, y de la misma manera que ocurre en una gramática que opera sobre cadenas de texto, una gramática puede ser "cíclica", es decir, puede volver a contener una línea de producción al elemento no terminal que las produce. De esta manera, se incorpora una restricción para evitar que esto suceda, puesto que al hacerlo, en redes bayesianas no es posible modelar una red "recursiva" o "sin límite estructural".

Así que se agrega una restricción a una gramática SR para que pueda ser transformada en una red bayesiana. Esta restricción elimina las reglas que produzcan no terminales de manera cíclica, por ejemplo las reglas del tipo:

$$1 : A^0 \rightarrow \langle B^2 \rangle$$
$$2 : B^0 \rightarrow \langle A^2 \rangle$$

donde A produce a B y B produce a A. Para ello se define una relación entre no terminales llamada

descendiente.

Supongamos los no terminales X, Y, Z pertenecientes a V_N. Si X por medio de una regla de producción-s produce a Y entonces Y es *descendiente* de X.

La relación *descendiente* tiene la propiedad de ser transitiva, es decir, si Y es *descendiente* de X y Z es *descendiente* de Y, entonces Z es *descendiente* de X.

Una vez definida la relación *descendiente*, la restricción de la gramática-SR queda de la siguiente forma:

Para toda regla de la forma

$$Y^0 \rightarrow \langle \mathbf{M}, \mathbf{R} \rangle$$

y para todo

$$m \in \mathbf{M}$$

se cumple que Y^0 no es *descendiente* de m.

Por otra parte, las gramáticas SR incorporan las producciones R de reescritura que permiten redefinir las relaciones para cada producción de los no terminales (reduciendo los posibles acomodos que pueden tener dos no terminales relacionados). Para efectos de la construcción de un MGP se describirán gramáticas sin este tipo de producciones haciendo $R = \varnothing$ a fin de crear un modelo con relaciones espaciales menos estrictas.

Como ejemplo en la siguiente sección se describe como una imagen previamente segmentada se puede reescribir usando este tipo de gramática visual.

5.3.2. Ejemplo de gramática.

El siguiente ejemplo utiliza una gramática SR para representar el objeto ojo de la figura 5.8.

$G = (\{OJOINIT, CEJA, OJO, OJOINT, IRIS, PUPILA\}, \{Eh, Ev, Hg\},$

$\{arribaDe, ady, dentroDe\}, OJOINIT, S, \varnothing)$.

Donde S contiene las siguientes producciones simbólicas:

$$
\begin{aligned}
1: OJOINIT^0 &\rightarrow < \{CEJA^2, OJO^2\}, \{arribaDe(CEJA^2, OJO^2)\} > \\
2: CEJA^0 &\rightarrow < \{Eh^2, Hg^2, Eh^3\}, \{arribaDe(Eh^2, Hg^2), arribaDe(Hg^2, Eh^2)\} > \\
3: OJO^0 &\rightarrow < \{OJOINT^2, Hg^2\}, \{arribaDe(OJOINT^2, Hg^2)\} > \\
4: OJO^0 &\rightarrow < \{OJOINT^2, Eh^2\}, \{arribaDe(OJOINT^2, Eh^2)\} > \\
5: OJOINT^0 &\rightarrow < \{Hg^2, IRIS^2, Hg^3\}, \{ady(Hg^2, IRIS^2), ady(IRIS^2, Hg^3)\} > \\
6: IRIS^0 &\rightarrow < \{Ev^2, PUPILA^2, Ev^3\}, \{ady(Ev^2, PUPILA^2), ady(PUPILA^2, Ev^3)\} > \\
7: PUPILA^0 &\rightarrow < \{Hg^2, Hg^3\}, \{dentroDe(Hg^2, Hg^3)\} >
\end{aligned}
$$

En la gramática descrita se parte del elemento inicial ($OJOINIT$) para ir descomponiendo este elemento en otros cada vez más sencillos. Cada uno de estos elementos se va relacionando con otros a través de relaciones espaciales descritas explícitamente en cada regla de producción. Para la primera regla, el elemento $OJOINIT$ se descompone en dos elementos ($CEJA$ y OJO) los cuales están relacionados mediante $arribaDe(v^2, w^2)$, es decir, $CEJA$ se encuentra encima de OJO. Como se recuerda en la sección 2.5, los exponentes cero son un formalismo para las variables del lado izquierdo de las reglas de producción y los exponentes mayores o iguales a dos describen la instanciación de los elementos en que se descompone cada regla. Los elementos terminales se componen de dos letras (Eh, Ev, Hg). Las relaciones espaciales pueden ocurrir entre elementos no terminales y terminales. Por la restricción que se ha impuesto en la sección 5.3.1, las producciones circulares no están permitidas.

Figura 5.8: a) Imagen de un ojo b) segmentación aplicando el primer alfabeto visual (sección 5.2.1). A partir de esta segmentación se construyó la gramática mencionada en el texto. Se recomienda ver esta imagen en color.

5.4. Aprendizaje de lexicón

Un lexicón tanto puede ser manualmente construido (decidir que elementos integren los terminales de la gramática) como puede ser aprendido de manera automática a partir de ejemplos. Se tienen tres tipos de lexicón propuestos en el modelo, el primero es más expresivo, pero presenta la desventaja de requerir una descripción manual, la cual para ejemplos con muchos elementos terminales, podría resultar muy laboriosa. El segundo lexicón es una versión simplificada del primero, que permite asignar un tipo de terminal de manera automática a cada región descrita en la primera fase de nuestra metodología. El último lexicón es un podado del anterior, ya que en algunos casos las palabras obtenidas son redundantes y se considera que mientras se obtenga un lexicón más limpio (un lexicón que evite tener palabras que no corresponden al objeto) los resultados pueden ser mejores (para reducir falsos positivos).

5.4.1. Lexicón descriptivo

Este lexicón se utiliza en conjunción con el alfabeto obtenido por segmentación. Este lexicón visual difiere de aquellos utilizados en textos [102], de modo que en este caso se ha dividido en dos partes. Primeramente se construye una tabla de características diversas basadas en la morfología y color de las regiones (histogramas de color, orientación dominante, color dominante, área, convexidad, etc.) de los elementos encontrados por el algoritmo de segmentación. Un ejemplo de esta tabla se describe en la Tab. 5.1. Posteriormente se define un archivo de descripción asociado a la gramática SR, que describe los terminales en términos de las características concentradas en la tabla. Se propuso la admisión de diversos operadores lógicos y relacionales (y ($\wedge$), *o* ($\vee$), *menor que* ($<$), *igual* (=), etc.). Un ejemplo de este lexicón se ilustra en la Fig. 5.9. Un terminal debe satisfacer todas las restricciones que aparezcan, por ejemplo, tener un área mayor a 300, exigir que sea de cierta orientación dominante, o que admita dos colores dominantes (que el terminal pueda ser rojo o verde, pero no amarillo ni azul). Como se ha mencionado el inconveniente de este lexicón es que es poco escalable la escritura manual del mismo para casos más complejos, por ejemplo para una

Tabla 5.1: Descripción de los segmentos o regiones encontradas en la imagen por el algoritmo de segmentación usado para generar el diccionario visual.

Imagen Segmentada				
# región	Color	Tipo	Área	Orientación
Reg(1)	Azul	Homogénea	500	-
Reg(2)	Naranja	Homogénea	250	-
Reg(3)	-	Borde	20	90°
Reg(4)	-	Borde	25	0°

```
% Archivo de Definiciones para Terminales
% Operadores Válidos: ==, >, <, >=, <=, ~=, &&, ||
% Features: REGION_NUMBER, COLOR, TYPE, AREA, ORIENTATION
% Las features pueden variar si el diccionario visual, cambia.

Hg1 := ((COLOR == YELLOW) || (COLOR == GREEN)) && (AREA > 0.2);
Bv1 := (ORIENTATION == 90) && (AREA < 60);
Bx1 := TYPE == 'Border';
```

Figura 5.9: Archivo lexicón asociado a la gramática SR que describe los elementos terminales de la misma. Se admiten diversos operadores. En el ejemplo el terminal definido como $Hg1$, puede tener un color dominante amarillo o verde y su área debe ser mayor al 0.2 de la imagen. $Bv1$ debe tener estrictamente una orientación de 90° y un área menor a 60 píxeles.

mayor variedad y cantidad de elementos en el alfabeto.

5.4.2. Lexicón simplificado generado automáticamente

Este lexicón propuesto es una versión más simple del anterior. De manera análoga contiene una tabla que mediante diversas características describe cada una de las regiones encontradas por el algoritmo de segmentación (o de recuadros, primera fase de nuestro método). Sin embargo, la diferencia radica en que ahora la creación de los elementos terminales es automática. Para ello se requiere de un conjunto de entrenamiento compuesto tanto de ejemplos positivos y negativos del objeto a aprender. A partir de las características de las regiones se aplica un algoritmo de agrupamiento (*k-medias*) y se asignan como elementos del lexicón los centroides de los grupos que a partir de un umbral arbitrario aparezcan más en los conjuntos de entrenamiento positivos. Un elemento del lexicón $p \in L$ es un vector de la forma $p = [f_1, f_2, ...f_n]$ tal que:

$$\sum_{I \in \mathscr{I}_p} freq(p,I) > \upsilon \tag{5.4}$$

donde $\upsilon = r \times |\mathscr{I}_p|$, I es una imagen, $\mathscr{I}_p$ es el conjunto de imágenes de entrenamiento positivas, r es un umbral entre 0 y 1, y $freq$ está dada por:

$$freq(p,I) = \begin{cases} 1 & p \in I \\ & \\ 0 & p \notin I \end{cases}, \tag{5.5}$$

Cuando se utilizan recuadros la generación es más directa, de manera que igualmente queden como terminales las palabras visuales que más aparezcan en el conjunto positivo de imágenes de entrenamiento. La idea es retener una fracción del tamaño del alfabeto visual. Al usar este tipo de lexicón se busca construir un modelo que automatice tanto la cantidad de elementos terminales así como seleccionar las palabras más discriminativas y frecuentes. Una vez aprendidos los terminales, éstos se usan para definir una gramática visual de forma manual o automática.

5.4.3. Lexicón reducido por espacialidad y eliminación de palabras redundantes.

El objetivo de esta estrategia es eliminar del lexicón aquellas palabras terminales que no aporten información significativa o que añaden ruido que se traduce en falsos positivos para el modelo. Este método se aplicó para el diccionario visual basado en recuadros. La técnica consistió en analizar la espacialidad de los recuadros encontrados para cada palabra visual. Si los recuadros asociados a una palabra visual tienen una espacialidad dispersa, se sugiere que es una palabra visual ruidosa: una palabra visual es ruidosa cuando ésta se compone de recuadros que provienen de diversas regiones de las imágenes de entrenamiento. Si por el contrario, dicha palabra acumula frecuencia en una localidad, es candidata a ser una palabra visual que discrimina con mayor precisión una región de las imágenes de entrenamiento. Esto es especialmente útil para bases de datos con imágenes de estructura rígida. El resultado es tener un lexicón con menos palabras ruidosas. En otras palabras, si

$\mathscr{M}_p$ representa el mapa de espacialidad de la palabra visual p, se aceptan las palabras que cumplen:

$$M_o(\mathscr{M}_p) \geqslant m, \tag{5.6}$$

donde $M_o(\mathscr{M}_p)$ es una función que devuelve la moda mayor de la frecuencia espacial en el mapa $\mathscr{M}_p$ y

$$m = \frac{\left(\max\limits_p \left(M_o(\mathscr{M}_p)\right) + \min\limits_p \left(M_o(\mathscr{M}_p)\right) \right)}{2}, \tag{5.7}$$

Un mapa espacial $\mathscr{M}_p$ es una matriz cuadrada de tamaño fijo que contiene la frecuencia de aparición de la palabra visual p en cada posición relativa para todas las imágenes de entrenamiento. Ejemplos de estas matrices $\mathscr{M}_p$ se ilustran en la Fig. 5.10 donde muestran la espacialidad de seis palabras visuales: las primeras acumulan frecuencia en una localidad determinada (en blanco) y las otras tienen una frecuencia que se distribuye en todas las coordenadas de las imágenes (muy grises).

Después de seleccionar un subconjunto más pequeño del lexicón, se realiza una tarea análoga a la selección de atributos, para quitar palabras ruidosas que, aunque tengan alta frecuencia en una localidad determinada, sean ruidosas para clasificar correctamente la categoría visual. También, si dos palabras tienen un comportamiento similar, son consideradas como una sola para efectos de palabra terminal en la gramática. Ambas cosas se pueden realizar mediante las siguientes fórmulas.

Si se cumple que para la k-ésima palabra del lexicón:

$$\left| \sum_{I \in \mathscr{I}_p} freq(p_k, I) - \sum_{I \in \mathscr{I}_n} freq(p_k, I) \right| < \xi, \tag{5.8}$$

donde $\xi = r \times |\mathscr{I}|$, entonces dicha palabra (p_k) se elimina, puesto que tiene poco poder discriminativo. El valor r es un real entre 0 y 1. La función $freq$ es como la definida en la Ec. 5.5.

Para el caso de fusión de palabras, si $\sum\limits_i misma(p_j, p_k, \mathscr{I}_i) = |\mathscr{I}|$, entonces se fusiona la palabra visual p_k, haciendo $p_j = p_j \cup p_k$. La función *misma* está definida por:

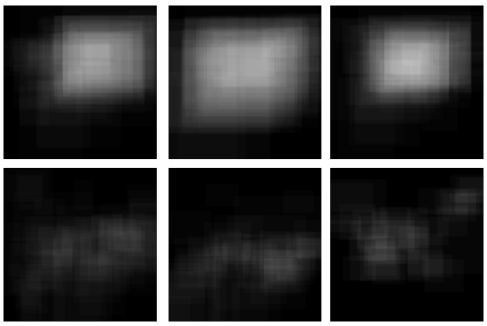


Figura 5.10: Ejemplos de la espacialidad de los recuadros encontrados por palabra visual. En algunos casos pueden incorporar ruido. Arriba: tres palabras visuales con recuadros de *localidad invariante.* Abajo: tres palabras visuales de localidad dispersa, por tanto son candidatas a eliminarse.

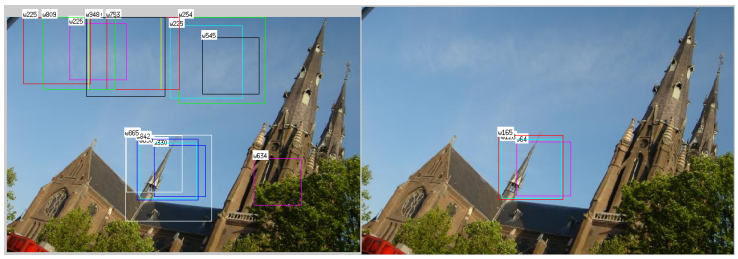


Figura 5.11: Izquierda: palabras encontradas con el *primer lexicón*. Derecha: palabras encontradas con el *segundo lexicón*. El uso de este segundo lexicón más compacto, permite eliminar aquellas palabras muy ruidosas que no aportan a la detección del objeto. Eliminar palabras ruidosas puede ayudar a mitigar el problema de falsos positivos.

$$misma(p_j, p_k, \mathscr{I}_i) = \begin{cases} 1 & \text{si } freq(p_j, \mathscr{I}_i) = freq(p_k, \mathscr{I}_i) \\ 0 & en\,otro\,caso \end{cases} \tag{5.9}$$

En esta técnica el objetivo es reducir el tamaño del lexicón. Como se observará más adelante, utilizar un lexicón más pequeño es preferible para reducir el tiempo de evaluación de un ejemplo, ya que los algoritmos de análisis gramatical buscan todas las configuraciones de palabras que encajen en la gramática. Si se tiene un lexicón más pequeño el tiempo de búsqueda es menor. Como es búsqueda de un patrón en una gramática, se puede ver como una búsqueda de un patrón en un árbol (la gramática propuesta genera un árbol a través de sus reglas de producción) De modo que si se busca el patrón a través del árbol generado, dicha búsqueda esta en P, a diferencia de la búsqueda de un patrón en un grafo (donde éste tenga ciclos) que está en NP-completo [40]. Un ejemplo del podado de lexicón con este método se ilustra en la Fig. 5.11.

5.5. Escritura de la gramática

Una manera de escribir la gramática es dado un conjunto de elementos terminales L, se procede a escribir reglas que, tal como un experto en el dominio, se consideren como útiles para describir un objeto visual. Sin embargo, este método no resulta escalable para aprender muchas categorías de objetos incluyendo variantes de los mismos. En este sentido, se desea construir las gramáticas de una forma lo más automática posible: un camino es tratar de aprender la gramática del modelo a partir de ejemplos. En esta sección se describe un algoritmo base, y una variante del mismo que permite, a partir de un conjunto de entrenamiento, construir una gramática con reglas de producción de manera automática.

5.5.1. Algoritmo base de aprendizaje de la estructura de la gramática

Este algoritmo obtiene de manera voraz reglas candidatas a conformar la gramática.

1. Se contabilizan todas los pares de relaciones con iguales elementos terminales (L) a lo largo de todas las imágenes de entrenamiento. Ejemplo: $ArribaDe(r_x, r_y)$.

2. Se define la frecuencia de aparición de la relación espacial como:

$$Freq_{c_i}(Rule_r(T_a, T_b)) = \sum_{I \in C_i} F(Rule_r(T_a, T_b), I) \tag{5.10}$$

   donde F está definida como:

$$F(Rule_r(T_a, T_b), I) = \begin{cases} 1 & \text{si } Rule_r(T_a, T_b) \in I \\ 0, & \text{si } Rule_r(T_a, T_b) \notin I \end{cases} \tag{5.11}$$

   C_i es la etiqueta de la categoría. Para problemas binarios es solamente la categoría positiva del objeto a reconocer. Para problemas multiclase, es cada una de las categorías de objetos a reconocer. $T_a, T_b \in L$ son elementos del lexicón.

3. Se toma la regla más frecuente y discriminativa de la categoría, siguiendo la fórmula

$$NewRule_{c_i} = \underset{T_a,T_b,r}{\arg\max} \left(Freq_{c_i}\left(Rule_r(T_a,T_b)\right) + \underset{c_x \in C, c_x \neq c_i}{\max} \left(dist\left(Freq_{c_i}\left(Rule_r(T_a,T_b)\right), Freq_{c_i}\left(Rule_r(T_a,T_b)\right)\right)\right)\right) \tag{5.12}$$

donde *NewRule* es un nuevo elemento no terminal pesado por su frecuencia de aparición en la categoría a reconocer y por su poder discriminativo contra las otras categorías (la categoría negativa si solamente son dos clases).

4. Obtenida la nueva regla se añade al conjunto R_{c_i} de reglas candidatas y a V_N mediante:

$$R_{c_i} = R_{c_i} \cup \left\{ NT^0_{r.a.b} \rightarrow \left\langle \left\{ {T_a}^2, T_b^2 \right\}, \left\{ Rule_r({T_a}^2, T_b^2) \right\} \right\rangle \right\} \tag{5.13}$$

donde el nuevo no terminal formado se añade a V_N:

$$VN_{c_i} = VN_{c_i} \cup \{NT_{r.a.b}\} \tag{5.14}$$

5. Después de ello, utilizando un umbral, se abstraen estas relaciones y se convierten en elementos no terminales. Los elementos que conformaban la relación, se sustituyen por el nuevo elemento no terminal.

6. Se repite este proceso de manera iterativa (pasos 3 a 5) abstrayendo incluso terminales con no terminales, hasta llegar al criterio de paro:

$$Freq_{c_i}\left(Rule_r(T_a,T_b)\right) \leqslant U_{c_i} \tag{5.15}$$

siendo U_{c_i} un umbral fijado a $r \times |\mathscr{I}|$. En otras palabras, el umbral es sensible al tamaño del conjunto de imágenes de entrenamiento.

7. Finalizado el proceso anterior, se escriben de manera inversa las reglas de producción a partir

del no terminal que contenga la cadena más larga de producciones.

8. Con las reglas de producción escritas, se agregan las definiciones de elementos no terminales y terminales y se escribe la gramática.

El umbral U_{c_i} permite que no se produzcan reglas que sólo cubren un ejemplo del conjunto de datos. Así también se evita que la gramática resulte "muy grande". Un ejemplo de una gramática generada se muestra a continuación:

$$\begin{aligned} NT3^0 &\rightarrow \left\langle \left\{NT2^2, Tx^2\right\}, \left\{ArribaDe(NT2, Tx)\right\}\right\rangle \\ NT2^0 &\rightarrow \left\langle \left\{NT1^2, NT1^3\right\}, \left\{IzqDe(NT1^2, NT1^3)\right\}\right\rangle \\ NT1^0 &\rightarrow \left\langle \left\{Tx^2, Ty^2\right\}, \left\{DentroDe(Tx^2, Ty^2)\right\}\right\rangle \end{aligned}$$

donde $Tx, Ty \in V_T$, y $NT1, NT2, NT3 \in V_N$. La interpretación que se obtiene de esta gramática se puede ir leyendo de abajo hacia arriba: en la tercera regla se tienen dos regiones Tx y Ty donde Tx está dentro de Ty. Una posible interpretación visual de esto es un disco[3] (Tx) contenido dentro de otro más grande (Ty). Posteriormente en la segunda regla se observan que los dos discos ($NT1$) están uno al lado del otro con la relación $IzqDe$. Finalmente, en la primera regla se indica que estos dos discos juntos ($NT2$) están arriba de una región pequeña ,Tx.

5.5.2. Variante del algoritmo

Al algoritmo anterior de generación de reglas se le realizaron unas modificaciones a fin de obtener conocimiento nuevo y darle flexibilidad al mismo modelo. Lo primero que se realizó fue incluirle reglas que no se encontraron en el conjunto de entrenamiento pero que pudieran ser útiles en prueba. Para flexibilizar el modelo se añadió la regla de disyunción "Or", que en la gramática se representa escribiendo dos reglas de producción con el mismo elemento no terminal del lado

[3]Puede tener forma de disco o forma cuadrada, tener textura o no. Dependiendo del lexicón, los elementos terminales pueden ser iguales o no. En esta tesis, cada terminal es similar a otro, pero en general, no son de igual forma o apariencia.

izquierdo. Estas dos modificaciones permitieron recuperar más ejemplos y hacer que la gramática aprenda variantes del objeto visual.

5.5.2.1. Aprendizaje de reglas nunca vistas

En algunas ocasiones se pueden aprender dos reglas en un conjunto de entrenamiento del tipo:

$$NT_1 \rightarrow \langle \{w_2, w_8\}, \{ArribaDe(w_2, w_8)\}\rangle \tag{5.16}$$

$$NT_2 \rightarrow \langle \{w_2, w_5\}, \{ArribaDe(w_2, w_5)\}\rangle \tag{5.17}$$

dado que se sostiene la misma relación espacial con w_2 como elemento en común, existe la probabilidad de que w_5 y w_8 sean sinónimos, o bien, variantes de pose, (por ejemplo w_2 represente el torso de una persona y w_5 pies juntos, mientras que w_8 representan pies más abiertos. Por lo anterior, si w_5 o w_8 vuelven a aparecer en otra regla, se debe tener en cuenta que son variantes de pose. De esta manera, si se encuentra una regla como $NT_5 \rightarrow Left(w_8, w_1)$, entonces por sinonimia visual, también debe existir $NT_6 \rightarrow Left(w_5, w_1)$, o al menos también podría ser una regla válida aunque esta última *nunca* haya aparecido como regla en el conjunto de entrenamiento. El objetivo de esto es enfrentar el problema de cola larga[4] [129], ya que sucede que ejemplos positivos "raros" no suelen estar en el conjunto de entrenamiento. La idea es inferir posibles ejemplos que, aunque no estén en el conjunto de entrenamiento, se puedan deducir a partir de las otras reglas, por sinonimia visual.

5.5.2.2. Aprendizaje de reglas alternativas (Reglas *Or*)

La generación de reglas se realiza mediante un algoritmo que descubre reglas frecuentes y discriminativas, una por una. Siguiendo el algoritmo anterior no se crean reglas de producción

[4]También llamado problema 80-20. Significa que los ejemplos comunes comprenden el 80 % de la totalidad de los mismos mientras que el 20 % restante son ejemplos raros, pero muy variados entre sí. Es muy difícil que este 20 % aparezca completo en un conjunto de entrenamiento.

disyuntivas (de tipo *Or*). Para incorporar estas reglas alternativas se usa un algoritmo voraz que busca grupos de reglas que cubren el conjunto de entrenamiento. Los grupos de reglas encontrados se consideran como reglas alternativas (*Or*) en la gramática. El algoritmo consiste en los siguientes pasos:

1. Inicializa $k = 1$.
2. Busca la regla más frecuente R_{c_i} en el el conjunto de entrenamiento $\mathscr{I}$.
3. Aprende esta regla y la asigna a un conjunto k de reglas $R^k = R^k \cup R_{c_i}$
4. Quita los ejemplos cubiertos por la regla anterior del conjunto de entrenamiento $\mathscr{I} = \mathscr{I} - \{cobertura(R_{c_i})\}$, donde *cobertura* devuelve todas las imágenes $I_i \in \mathscr{I}$ tales que la regla R_{c_i} está presente en ellas.
5. Repite el proceso haciendo $k = k + 1$ hasta cubrir todos los ejemplos.

Lo anterior genera k particiones del conjunto de entrenamiento. En cada partición se siguen buscando reglas frecuentes y discriminativas que describan a esa partición del conjunto de entrenamiento. El resultado es que se aprende una gramática con reglas *Or* en su primera producción de manera automática. Cuando se transforma una gramática aprendida de esta manera a una red bayesiana, los nodos *Or* quedan como hijos del nodo raíz. Un ejemplo de estos nodos *Or* se ilustra en la Fig. 5.12.

5.6. Transformación de la gramática a una red bayesiana

Si se desea realizar inferencia sobre el modelo se debe tener en cuenta que la naturaleza de los modelos basados en un enfoque estructurado, precisan considerar la posibilidad de que un objeto deba ser detectado a pesar de que no contenga todos los elementos descritos en la gramática. Un ejemplo de esto, son los casos de oclusión parcial del objeto. En estas tareas de oclusión, conviene involucrar manejo de incertidumbre. Dado que las gramáticas SR solamente trabajan con reglas de producción fijas, resulta necesario incorporar técnicas o estrategias que permitan modelar el

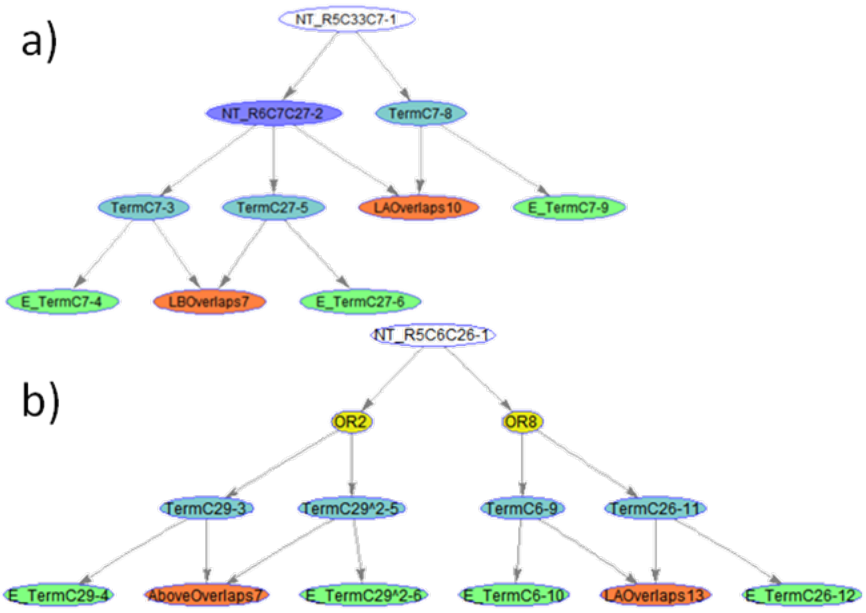


Figura 5.12: a) Red bayesiana que no tiene un nodo Or. Esta red tiene un nivel de composición. b) Red bayesiana con opción a escoger una regla de dos disponibles. Los nodos *Or* (en amarillo) se ubican inmediatamente debajo del nodo raíz. Se ve mejor en color.

problema trabajando con incertidumbre. Lo anterior lleva a incluir modelos gráficos probabilistas para manejarla. Lo que se busca es convertir de una manera automática la gramática SR a un MGP. Por lo anterior, se desarrolló un algoritmo que transforma la gramática en una red bayesiana.

5.6.1. Transformación de gramática SR a red bayesiana

Este algoritmo convierte la descripción de un objeto de gramática SR a una red bayesiana (RB), aprendiendo la estructura de la misma. Primeramente el nodo raíz de la RB es el primer elemento de la gramática, y los nodos hijos se irán construyendo acorde a las producciones S de la gramática. Cada nodo tiene dos estados, uno que define la presencia del objeto (elemento terminal o región de la imagen) y otro estado que define la ausencia del mismo. El algoritmo propuesto también tiene una variante que considera tener varios estados en los nodos de la red, cambiando el aprendizaje paramétrico. A continuación se describen ambos casos.

Algoritmo 5.1 Transformación de la gramática SR a una red bayesiana

Data: $G(V_N, V_T, V_R, S, P, R), Nr$; /* Nr = Nodo Referencia */
Result: Bn

if $Nr = S$ **then**
  Fijar S como nodo raíz en Bn
foreach $p_i \in P$ ***donde*** $Y^0 = Nr$ **do**
  // p_i tiene la forma $l : Y^0 \rightarrow \langle \mathbf{M}, \mathbf{R} \rangle$
  foreach $m \in \boldsymbol{M}$ **do**
    Añadir p_i como hijo de Nr
    if $p_i \in V_N$ **then**
      ConvertSRGtoBN(G, p_i); /* Recursión */
    if $p_i \in V_T$ **then**
      Añadir p_{i_E} como hijo de p_i
  foreach $r_i \in \boldsymbol{R}$ **do**
    // r tiene la forma $r(X,Y)$
    Añadir nodo r_i como hijo de X y Y.

5.6.1.1. Algoritmo base de transformación gramática SR a red bayesiana

Este algoritmo realiza la transformación de la gramática SR a una red bayesiana con su estructura aprendida. Esto se muestra en el algoritmo 5.1. De manera breve, el algoritmo de conversión realiza los siguientes pasos:

1. Fijar el nodo raíz.

2. Para cada regla $producción - s$, donde el término de la izquierda es el nodo Referencia Nr (siempre es $Nr \in V_N$) y para cada símbolo p_i definido en cada producción (lado derecho de la regla), Añade p_i como hijo de Nr. Si $p_i \in V_N$, realizar una llamada recursiva del algoritmo haciendo $Nr = p_i$. Si $p_i \in V_T$, añade p_{i_E} (nodo evidencia) como hijo de p_i.

3. Para cada relación $r(X,Y)$, añade el nodo como un hijo de sus padres X y Y.

Para el caso del ejemplo de gramática que describe un ojo (sección 5.3.2), se ilustra la red bayesiana que genera este algoritmo en la Fig. 5.13.

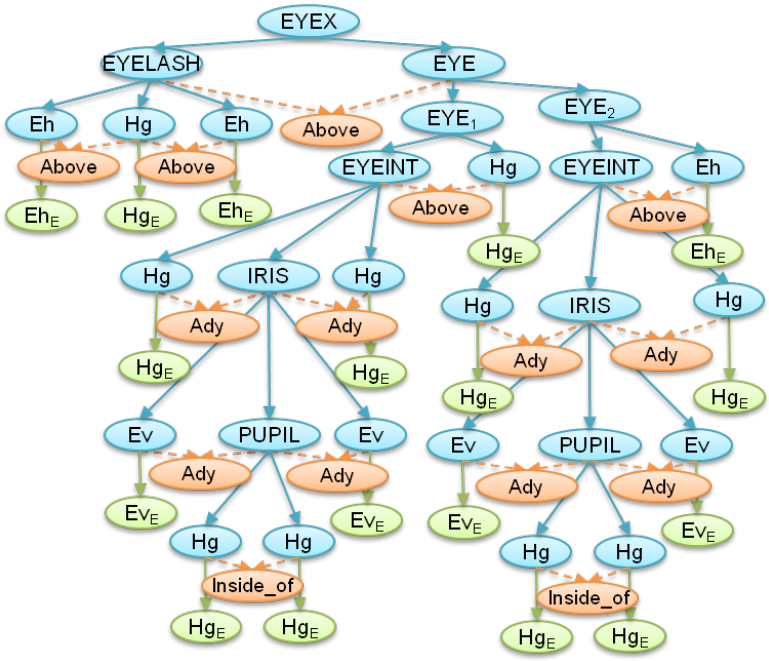


Figura 5.13: Red bayesiana generada por el algoritmo de transformación. Los nodos hoja siempre son nodos de paso de evidencia o nodos de relaciones espaciales. Los nodos con subíndice (EYE_1, EYE_2) son nodos tipo *Or*. El nodo raíz es el símbolo inicial de la gramática.

5.6.1.2. Variante de transformación: reglas con multiestados

Es posible que existan elementos terminales parecidos entre sí, pero no por ello signifique que sean elementos terminales redundantes, más bien son palabras que *espacialmente* pueden aparecer cerca de otras palabras pero que no están pareadas. En estos casos se analizó la posibilidad de manejar varios estados en la red bayesiana, en lugar de los dos únicos estados que manejan en sus nodos hoja (presencia/ausencia de la región o recuadro). El objetivo es tener una distribución de probabilidad asociada a varias palabras visuales en lugar de solamente una palabra visual. El cambio se aplica únicamente en las tablas de probabilidad condicional de los nodos hoja de la red bayesiana. El método propuesto para ello es el siguiente:

1. Para cada imagen del conjunto de entrenamiento $\mathscr{I}$,

   a) para cada palabra visual p_i,

      1) buscar aquellas otras palabras visuales p_j que cumplan con $dist(p_i, p_j) < \varepsilon$, es decir, una proximidad espacial menor que el umbral ε.

	pos	neg
pos	α	$1-\beta$
neg	$1-\alpha$	β

	pos	neg
p_1	μ_1	ν_2
p_2	μ_2	ν_2
p_3	μ_3	ν_3
p_4	μ_4	ν_4

Tabla 5.2: Ejemplo de transformación a una TPC a multiestados. Las tablas se aplican a los nodos que se corresponden con los elementos terminales de la gramática. Se cumple que $\sum_k \mu_k = 1$ y $\sum_k \nu_k = 1$. Cada $m_k, \alpha, \beta \in [0 \dots 1]$.

2) para cada p_j que cumpla el enunciado anterior hacer: $M_{i,j} = M_{i,j} + 1$,

2. Se toman los n valores más grandes de la matriz M para cada renglón i como estados para la palabra visual p_i. En pruebas se usó $n = 3$ y $n = 4$. En casos donde no haya palabras que satisfagan el umbral, la palabra p_i se queda como nodo binario. En la Tabla 5.2 se muestra un ejemplo del cambio de los valores de TPC de dos estados a TPC de varios estados (cuatro estados). Para el llenado de las TPC, ver la sección 5.6.1.3.

5.6.1.3. Aprendizaje paramétrico de la red

Una vez que se ha aprendido la estructura de la red bayesiana se deben aprender los parámetros. Se realizaron dos formas de aprendizaje: en la primera se obtienen las probabilidades condicionales *a priori* a través de un experto. En la segunda se utiliza aprendizaje estadístico.

- Red bayesiana con aprendizaje subjetivo: este método se utilizó cuando los nodos de la red son todos de dos estados. En los nodos hoja de la red colocaron probabilidades de acuerdo a ejemplos usando un estimador de máxima verosimilitud. En el resto de los nodos se utilizó un aprendizaje subjetivo en función de la aparición o no aparición de los elementos no terminales. Este tipo de aprendizaje es simple y se puede incluir en el algoritmo de transformación explicado anteriormente en la sección 5.6.1.1.

- Red bayesiana con aprendizaje estadístico: este método se puede aplicar para nodos con dos o más estados. En este método se inicializa cada nodo con probabilidades a partir de ejemplos (vía estimador de máxima verosimilitud), y una inicialización uniforme en nodos donde no

se pueda estimar (se consideró como nodos no observables a los no terminales). Después de ello se realizó una estimación EM [23] para aprender los parámetros de las variables. En lugar de utilizar las imágenes de entrenamiento se utilizó un conjunto de validación, distinto del de entrenamiento, para prevenir sobreajuste.

5.7. Inferencia del modelo

Hasta este momento se han descrito los pasos para entrenar el modelo. La fase de inferencia consiste en evaluar si un ejemplo recibido puede ser aceptado o no por la gramática. Como en el modelo propuesto la inferencia se realiza en la red bayesiana y no en la gramática, se necesita generar un algoritmo que pase adecuadamente la evidencia encontrada en una imagen I hacia una red $\mathscr{R}$, de modo que el proceso de inferencia se realice de manera correcta. Como nota, la consulta sobre si el objeto se ha hallado en la imagen se encuentra siempre en el nodo inicial de la red.

El algoritmo propuesto busca exhaustivamente todas las configuraciones aplicables. Aunque esto en principio puede resultar lento de evaluar, siempre se encuentra en P, dado que el problema es equivalente a búsqueda de subgrafos en un árbol. Para acelerar el proceso, en secciones anteriores se habló de la conveniencia de podar el lexicón y de producir reglas en la gramática usando un umbral, para evitar la escritura de gramáticas que superen ese umbral. El algoritmo parte de una imagen I descrita en términos del lexicón y a partir de ahí hace una transformación hacia un grafo G etiquetado en donde los nodos son regiones o recuadros de la imagen y las aristas son las relaciones espaciales. La configuración aceptada es:

$$\arg\max_{r_I, r_R} Expandir(r_I(a_I, b_I), r_R(a_R, b_R), \mathscr{R}, G) \tag{5.18}$$

donde $r_I(a_I, b_I)$ es una relación espacial que se encuentra en G, y $r_R(a_R, b_R)$ es una relación espacial que se encuentra en la red $\mathscr{R}$, donde $a_R, b_r \in V_T$. Visto de manera gráfica una relación $r(a,b)$ se puede interpretar como dos nodos con etiquetas a y b que sostienen una arista con etiqueta r. La función *Expandir* es recursiva y busca el subgrafo (o subconfiguración) más grande de G que se

corresponde con una subestructura de la red $\mathscr{R}$. *Expandir* devuelve la probabilidad de hallar el objeto dada la evidencia observada (el subgrafo de G). Como el subgrafo de G está etiquetado y el grafo G es un isomorfismo de la imagen I, es posible también indicar *dónde* está el objeto en la imagen. El algoritmo 5.2 describe la función *Expandir* para hallar estas subconfiguraciones. Lo que realiza este algoritmo es lo siguiente:

1. Toma un par de elementos visuales a_I, b_I de la imagen I que sostienen una relación espacial r_I.

2. Evalúa este par de elementos para ver si son aceptados por la red $\mathscr{R}$ (debe haber un par $a_R, b_R \in \mathscr{R}$ que acepten a a_I, b_I).

3. Si son aceptados:

   a) Busca otro par de elementos visuales en la imagen dentro de la *vecindad* del par anterior.

   b) Si encuentra otro par, repite el paso 2 con este nuevo par (usando recursión).

4. Si no son aceptados o no se encuentra otro par:

   a) Devuelve todos los pares agrupados hasta el momento (un subgrafo de G, o bien una región de I) y devuelve la evidencia de todos estos pares sobre la red bayesiana, haciendo posible la inferencia.

A manera de ejemplo si en una imagen I se aplica una gramática $\mathscr{G}$ que detecta caballos, después de la inferencia con el algoritmo 5.2, la Ec. 5.18 devuelve la configuración más probable y el lugar donde se encuentra el objeto aprendido por la gramática (Fig. 5.14).

5.8. Resumen

En este capítulo se describió el método a seguir para construir el modelo de reconocimiento de objetos. El modelo presentado sigue una estructura por capas donde cada capa admite ciertos

Algoritmo 5.2 Algoritmo de búsqueda de subconfiguraciones en la imagen que correspondan en la gramática.

Data: **Red**, **Imagen** ; `/* Imagen segmentada */`
Result: **Subconfiguraciones** ; `/* con correspondencia en la Imagen */`

foreach *Relación aplicable en* ***Imagen*** **do**
  foreach *Relación de V_T's en la* ***Red*** **do**
    Llamar a **Expandir** con **Red** e **Imagen**
      Marcar Evidencia en Relaciones y Nodos
      Buscar Relaciones en **Imagen** que coincidan con
      relaciones adyacentes en la **Red**
      `// a los lados, hacia abajo y hacia arriba`
    if *No hay relaciones en* ***Imagen*** *o la* ***Red*** *no tiene mas nodos* **then**
      Aceptar **Subconfiguracion** y terminar
    if *hay n relaciones en* ***Imagen*** **then**
      Llamar a **Expandir** por cada relación (Recursión)

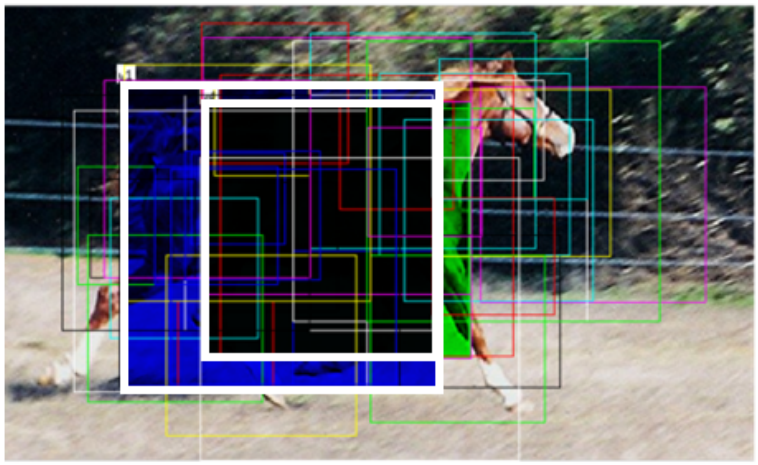


Figura 5.14: Ejemplo devuelto por el algoritmo de análisis gramatical. El algoritmo devuelve el lugar donde se ubica la configuración más probable (en la imagen es el par de recuadros blancos). Los otros recuadros en la imagen se corresponden con otras configuraciones con menor probabilidad.

cambios sin afectar el resto. Por ejemplo, se pueden usar diversos tipos de alfabetos visuales dependiendo de las aplicaciones. También, el lexicón puede admitir variantes en los que se modifique manualmente o se pode a fin de eliminar aquellas palabras ruidosas al modelo. La gramática se puede escribir manualmente o aprender automáticamente. Adicionalmente, se puede restringir la gramática para que admita o rechace la generación de reglas disyuntivas (reglas Or) aprendidas también de manera automática. La gramática visual aprendida es transformada a una red bayesiana para realizar el proceso de reconocimiento mediante inferencia en la red. En el siguiente capítulo se analizará y se presentará el método para utilizar un modelo relacional en lugar de la red bayesiana. Se mostrará como se puede migrar la gramática en particular a dos modelos relacionales: redes bayesianas relacionales [54] y redes lógicas de Markov [99].

Capítulo 6

Modelo de Reconocimiento de Objetos basado en Gramáticas Visuales y Modelos Relacionales Probabilistas

6.1. Introducción

En el capítulo anterior se describió el método para reconocer objetos basado en gramáticas simbólico relacionales y redes bayesianas. En este capítulo se verá como construir una variante del mismo, que en lugar de construir una red bayesiana genera un modelo estadístico relacional, los cuales son una generalización de los MGP en combinación con lógica. Buena parte de la motivación de utilizar un modelo relacional en lugar de la red bayesiana radica en los siguientes aspectos:

- Los modelos relacionales al ser una extensión de los MGP, tratan mejor el problema de lidiar con objetos que cambian su comportamiento según las relaciones que tenga éste con otros objetos del dominio: las expresiones de relaciones espaciales de regiones en imágenes actúan como relaciones entre objetos en los modelos relacionales.

- Un modelo relacional captura la estructura general de un objeto (o de un conjunto de objetos) incluyendo incertidumbre. Esta generalización suave es un beneficio adicional sobre utilizar

inferencia basada en reglas, siendo este último un enfoque más estricto.

- Se desea comparar el modelo *ad-hoc* de red bayesiana propuesto contra las instancias de modelos gráficos que generan los modelos relacionales. Se desea analizar la complejidad estructural y eficiencia en ambos modelos.

- Los modelos relacionales trabajan con bases de conocimiento como información de entrada. La información que provee una gramática visual es transformable mediante una función biyectiva a una base de conocimiento de manera directa y con costo bajo.

- Los modelos relacionales tienen hasta el momento relativamente pocas aplicaciones en la literatura (aunque en aumento). La mayoría de ellas se centra en tratar de realizar inferencia sobre bases de datos relacionales sintéticas. En otros casos se utilizan en experimentos llamados de cajas de arena que permiten mostrar sus ventajas pero con relativamente poca utilidad práctica. Hace falta descubrir el potencial de tales modelos en aplicaciones más realistas.

- Los modelos relacionales tienen problemas de eficiencia con su aprendizaje e inferencia (NP) y se desea conocer la escalabilidad con aplicaciones más realistas. No se busca mejorar el coste computacional de estos algoritmos, solamente se explora salir de los experimentos pequeños y saltar hacia una aplicación con mayores datos que los vistos en los trabajos relacionados.

- Si se representa una imagen en términos de un modelo relacional probabilista, se abre una puerta a convertir cualquier tipo de meta-datos (audio, imágenes, video) en estructuras relacionales probabilistas. La extracción del conocimiento en estas estructuras parece promisorio a futuro puesto que gran parte del conocimiento en la web podria ser representado de manera relacional.

Como no hay trabajos previos que realicen una actividad similar, estos modelos serán comparados contra el método de inferencia con redes bayesianas. En las siguientes secciones se describirán los

cambios que se realizan a nuestro método base para integrarle la transformación hacia modelos relacionales.

6.2. Representación del conocimiento con Modelos Relacionales Probabilistas

Para construir un modelo relacional probabilista, como se vio en la sección 3.3, se necesita la descripción del modelo y una *BC* que represente al ejemplo a evaluar. Por ello se necesita representar el conocimiento en términos de una *BC*. Considerando que el modelo en la etapa de inferencia transforma una imagen de entrada en un grafo mediante una función biyectiva, se puede análogamente crear una función que, para cada relación espacial encontrada en la imagen, escriba un predicado de la forma $rel(nA, nB)$; con $rel \in V_R$ y con $nA, nB \in V_T$. Dependiendo del modelo relacional a usar, la interpretación en predicados puede ser sutilmente distinta; en RBR se debe construir un S^D, mientras que en RLMs basta con la *BC* tal cual. Ninguna transformación quita o añade información a la generada por una imagen representada en términos del lexicón y sus relaciones espaciales, de modo que la generación de la *BC* (o el S^D) es transparente y directa.

6.3. Gramática y lexicón

En este modelo se usaron dos tipos de alfabetos visuales, uno basado en detectores de mediano nivel, utilizando el algoritmo de Viola y Jones [120] y otro basado en regiones homogéneas (explicado en la sección 5.2.1). En el primer algoritmo se consideraron los siguientes detectores: Boca, Nariz, Ojos, Cabeza (que detecta la piel del rostro). Estos detectores aplicados sobre las imágenes generan recuadros donde hay probabilidad de encontrar al objeto descrito. Este tipo de alfabeto visual está basado en los trabajos de [75]. El método seguido en esta parte para generar el lexicón y la gramática es equivalente al usado en nuestro método general. El cambio ocurre al momento de transformar la gramática hacia la red bayesiana. En lugar de trasladar la gramática hacia la red,

la gramática se transforma en un modelo relacional. En las siguientes secciones se describen los dos casos elaborados: transformación hacia redes bayesianas relacionales y hacia redes lógicas de Markov.

6.3.1. Transformación hacia una red bayesiana relacional

En el caso de una red bayesiana relacional, se transforma cada relación espacial que esté escrita en la gramática en una fórmula probabilista, que además se irá descomponiendo de acuerdo a la estructura dada por la misma gramática. En RBR se forma una composición de fórmulas probabilistas que además aprenden sus distribuciones de probabilidad de acuerdo a los ejemplos de validación. Los átomos unarios tales como $RegionRoja(r_x)$, $RecuadroTipoX(r_x)$ no se consideran probabilistas, y sólo devuelven valores de verdadero/falso. Este tipo de predicados son muy útiles para restringir la búsqueda en todo el universo de los datos donde se realiza la inferencia, pues se evita una explosión de combinaciones. Considerar estos predicados de manera probabilista puede elevar el costo computacional tanto del entrenamiento como de la inferencia.

De esta manera la transformación propuesta sigue dos pasos. En el primero, cada relación espacial del tipo $r(v_l, v_r)$ se reescribe en un *RRSM* usando la siguiente fórmula:

$$r([v_l]x_l, [v_r]x_r) \quad = \quad Is_{v_l}(x_l) : (Is_{v_r}(x_r) : (R(x_l, x_r) : (p_{r^+}, p_{r^-}), 0.0), 0.0) \tag{6.1}$$

donde cada Is_* es una fórmula predefinida (no estocástica) que evalúa los tipos de los elementos (se trata de forzar que x_l y x_r tengan los mismos tipos que v_l y v_r, por ello cuando la fórmula predefinida Is_* es falsa, ésta devuelve 0.0). p_{r^+} y p_{r^-} son valores de probabilidad asociados a la relación espacial $R(v_l, v_r)$ los cuales se obtienen de las fórmulas:

$$p_{r^+} = P(Obj_{\mathscr{G}}(S_{\mathscr{I}})|SpatialRel(S_{\mathscr{I}}, R(v_l, v_r))) \quad (6.2)$$

$$p_{r^-} = P(Obj_{\mathscr{G}}(S_{\mathscr{I}})|\neg SpatialRel(S_{\mathscr{I}}, R(v_l, v_r))) \quad (6.3)$$

donde *SpatialRel* en la Ec. 6.2 significa la probabilidad de la relación R en el conjunto de entrenamiento $S_{\mathscr{I}}$. $\neg SpatialRel$ en la Ec. 6.3 es la probabilidad cuando la relación R no ocurre. $Obj_{\mathscr{G}}(S_{\mathscr{I}})$ es la probabilidad de encontrar el objeto descrito por la gramática $\mathscr{G}$ en el conjunto de entrenamiento $S_{\mathscr{I}}$.

Para el segundo paso se realiza una composición usando el símbolo inicial de la gramática:

$$Obj_G([t_1]v_1, ...[t_n]v_n) = R_1([t_1]v_1, [t_2]v_2) : R_2(...), p_1) \quad (6.4)$$

y la última relación se escribe como:

$$R_j([t_l]v_l, [t_r]v_r) : (p_j, p_{j-1}) \quad (6.5)$$

donde las probabilidades están dadas por:

$$p_j = P(Obj_{\mathscr{G}}(S_{\mathscr{I}})|SpatialRel(S_{\mathscr{I}}, R_j^*(v_l, v_r))) \quad (6.6)$$

donde $R_j^*(v_l, v_r)$ es la *concatenación conjuntiva*[1] de las reglas $R_1, R_2, \ldots R_j$ que aparecen en las producciones de tipo P en la gramática. La última regla significa que todas las relaciones de la gramática estuvieron presentes, es decir, se encontraron todas las partes con sus relaciones espaciales del objeto en la imagen.

[1]También se le puede llamar anidación de las reglas $R_1, R_2, \ldots R_j$, vistas como funciones anidadas una dentro de otra.

Para ejemplificar lo anterior, si se tiene una regla ($producción - P$) en una $gramática - SR$ como:

$$1 : FACE^0 \quad \rightarrow \quad \left\langle \{eyes^2, mouth^2\}, \{above(eyes^2, mouth^2)\} \right\rangle \tag{6.7}$$

entonces su transformación hacia la estructura aleatorio relacional (*RRSM*), siguiendo la Ec. 6.1, se genera de la siguiente forma:

$$above([Rg]x_l, [Rg]x_r) = IsEyes(x_l) : (IsMouth(x_r) : (above(x_l, x_r) : \ (p_{r^+}, p_{r^-}), 0.0), 0.0); \tag{6.8}$$

donde x_l y x_r admiten posibles regiones que cumplan con la regla. p_{r^+} y p_{r^+} se obtienen de las fórmulas 6.2 y 6.3. Las restricciones *Rg* reducen el número de variables que se evalúan en la regla anterior. Las funciones $IsEyes$, $IsMouth$ son predefinidas, de modo que devuelven un valor de verdadero. En este ejemplo, la relación probabilista *above* se evalúa si y sólo si ambas son verdaderas. En caso contrario la probabilidad devuelta por esta estructura es 0.0 (lado derecho de la regla). Posteriormente, para la composición se utiliza la Ec. 6.4. Debido a que en este ejemplo de una sola regla no hay anidación, se usa directamente la Ec. 6.5 y queda como:

$$FACE([IsEyes]v_1, [IsMouth]v_2) = above([IsEyes]v_1, [IsMouth]v_2) : (0.8, 0.43); \tag{6.9}$$

Lo anterior concluye la formación de la estructura aleatoria relacional. Para el proceso de inferencia, se transforma un nuevo ejemplo en una S^D. Esta transformación es simple y solamente define el dominio de los objetos y las relaciones:

$$Dom = \{v_i | v_i \in \mathscr{I}\} \tag{6.10}$$

$$Rel = \{R_i(v_l, v_r) | R_i(v_l, v_r) \in \mathscr{I}, \{v_l, v_r\} \subseteq Dom, R_i \in V_R\} \tag{6.11}$$

Se ilustra un ejemplo de un típico S^D a continuación:

```
DOMAIN: B1 B2 B3 B4 C1 C2 C3 C4 N1 N2 N3 N4 O1 O2 O3 O4;
RELATION: Above/2 \{(11,0) (8,3) (10,2) (12,0) (15,0) (14,2)\};
RELATION: IsMouth/1 \{0 1 2 3\};
```

En este ejemplo, el dominio son todas las regiones encontradas en la imagen. La relación espacial *Above* tiene aridad dos y la relación *IsMouth* tiene aridad uno. Los números entre llaves corresponden a índices respecto del dominio (sólo por razones de simplicidad en la representación). En estos modelos también se permiten aridades superiores a dos, aunque no se realizaron ejemplos para estos casos. En adición a este ejemplo, se ilustra un caso de un *RRSM* y su S^D asociado con estas fórmulas en la Fig. 6.1.

6.3.2. Transformación a una red lógica de Markov

Realizar la transformación a una red lógica de Markov es más directo. Primero se necesita declarar las fórmulas. Esta declaración es similar al caso de RBRs: $R_i(v_l, v_r)$ (ver Ec. 6.11).

De la misma manera, sólo se consideran fórmulas con aridad dos. Ejemplos de este tipo de relaciones para un caso de reconocimiento de rostros a partir de elementos terminales de mediano nivel:

```
aboveEM(eyes,mouth), aboveNM(nose,mouth),
withinEH(eyes,head), withinNH(nose,head),
withinMH(mouth,head), FaceENMH(eyes,nose,mouth,head)
```

Estas fórmulas se obtienen de V_R de la gramática $\mathscr{G}$. Después de ello, se requiere declarar el dominio (equivalente a S^D en la RBR). El dominio en RLM queda definido igual que en el caso

a)

```
EyesWithinHead([Rg]a, [Rg]b) =
        (sformula(IsEyes(a)):
                (sformula(IsHead(b)):
                        (sformula(Within(a,b)):
                                0.75, 0.37),
                0.0),
        0.0);
[...]
```

b)

```
IsFace([IsEyes]E, [IsNose]N, [IsMouth]M, [IsHead]H) =
(NoseAboveMouth(N,M):
        (EyesAboveMouth(E,M):
                (EyesWithinHead(E,H):
                        (NoseWithinHead(N,H):
                                (MouthWithinHead(M,H): 0.98, 0.86),
                                0.81),
                0.74),
        0.66),
0.48);
```

c)

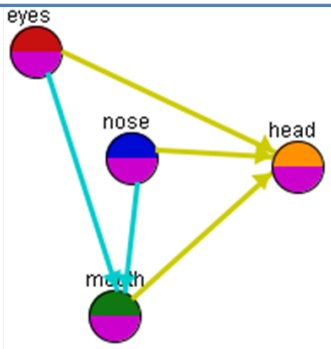


Figura 6.1: a) Fórmulas relacionales como "EyesWithinHead" se obtienen de aplicar la ecuación 6.1. Cada relación espacial debe estar definida (en la imagen solamente se muestra una de ellas). b) Función principal del RRSM. La ecuación 6.4 permite construir esta función desde el nodo raíz (para este ejemplo, *IsFace*) el cual se define en términos de las anteriores fórmulas de manera anidada. c) representación gráfica del S^D. Los nodos son regiones o segmentos de una imagen, el color de los nodos son las relaciones de aridad-uno (o atributos) aplicados a cada nodo (objeto, tal como *IsEye*) y las flechas representan las relaciones de aridad dos (tal como *Above*). Números son ejemplos de los valores p_{r^+}, p_{r^-} y p_j con propósitos ilustrativos.

anterior (Ec. 6.10).

A manera de ejemplo, para el caso del problema de reconocimiento de rostros un ejemplo de un dominio se define como:

```
eyes={E1,E2,E3,E4}; nose={N1,N2,N3,N4}
mouth={M1,M2,M3,M4}; head={H1,H2,H3,H4}
```

A continuación se tienen que escribir las fórmulas de primer orden con su peso asociado. En este caso se reescribe cada regla de la forma:

$$NT_i \quad \rightarrow \quad \langle \{v_l, v_r\}, \{R(v_l, v_r)\} \rangle \tag{6.12}$$

en una fórmula pesada admitida por la RLM:

$$\mathtt{F_i} \quad = \quad w_i \mathtt{NT_i}(\mathtt{v_l}, \mathtt{v_r}) \vee \mathtt{R}(\mathtt{v_l}, \mathtt{v_r}) \tag{6.13}$$

Para calcular w_i se utiliza un conjunto de validación y se define su valor como:

$$w_i \quad = \quad P(Obj_{\mathscr{G}}(S_{\mathscr{J}}) | SpatialRel(S_{\mathscr{J}}, R(v_l, v_r))) + \xi_M \tag{6.14}$$

donde $S_{\mathscr{J}}$ es el conjunto de entrenamiento y $\xi_M \in \Re^+$ es un parámetro de ajuste. Un valor más alto de ξ_M reduce las diferencias de peso entre distintas fórmulas (cada fórmula tendrá la misma importancia relativa que las demás).

Si $\xi_M \rightarrow \infty$ se sigue que $w_{F_i} \rightarrow \infty$ para todas las fórmulas, perdiendo el enfoque probabilista (si una regla no se cumple, la probabilidad se vuelve cero). Por el contrario, si $\xi_M = 0$ algunas fórmulas tendrán un peso muy pequeño y reducirán su importancia frente a otras en la RLM. Este parámetro permite tener un compromiso entre evitar pesos iguales y evitar pesos cercanos a cero.

Tabla 6.1: Ejemplo de una BC para una RLM. Todos los símbolos y predicados que se usen en la BC deben ser previamente definidos en el dominio de la RLM. Los predicados no admiten sobrecarga, de modo que la diferencia entre aboveNM y aboveEM es que sus argumentos son de distinto tipo. La BC puede representarse gráficamente tal como en la Fig. 6.2.

aboveNM(N4,M1)	withinEH(E3,H3)	withinNH(N2,H2)
aboveEM(E3,M3)	withinEH(E1,H1)	withinMH(M3,H3)
aboveEM(E4,M1)	withinEH(E1,H2)	withinMH(M1,H4)
aboveEM(E1,M1)	withinEH(E1,H4)	withinMH(M2,H1)
aboveNM(N3,M3)	withinNH(N3,H3)	withinMH(M2,H4)
aboveNM(N1,M4)	withinNH(N4,H4)	

$R(v_l, v_r)$ se obtiene del lado derecho de la fórmula F_i y *SpatialRel* se calcula como en la Ec. 6.2.

En el caso de las reglas de *producción disyuntiva* (*Reglas-Or)* también se pueden escribir en la RLM. Si se tiene la gramática $\mathscr{G}$ compuesta de las siguientes reglas *Or*:

$$1 : FACE^0 \rightarrow \left\langle \{eyes^2, mouth^2\}, \{above(eyes^2, mouth^2)\} \right\rangle \tag{6.15}$$

$$2 : FACE^0 \rightarrow \left\langle \{nose, mouth^2\}, \{above(nose^2, mouth^2)\} \right\rangle \tag{6.16}$$

Se transforman en una RLM de acuerdo a las Ec. 6.12 y 6.13, quedando de la siguiente manera:

```
1.58 FaceENMH(e,n,m,h) V aboveEM(e,m)
```

```
1.67 FaceENMH(e,n,m,h) V aboveNM(n,m)
```

Debido a que el formalismo de las redes lógicas de Markov no permite sobrecarga de predicados (o dos predicados de igual nombre, pero distintos argumentos), se han añadido letras a los predicados: *aboveEM* y *aboveNM* dado que tienen diferentes argumentos. Teniendo la estructura de la RLM se debe describir también la base de conocimientos asociada a los ejemplos que se ejecutarán con la RLM. En la tabla 6.1 se ilustra una *BC* sintética que después será procesada por una RLM. El gráfico de la Fig. 6.2 es análogo a la tabla.

Para la realización de la inferencia, se traslada esta estructura en una red de Markov. Para el ejemplo anterior, se ilustra el ejemplo de la red en la Fig. 6.3. Cada ejemplo primero se transforma

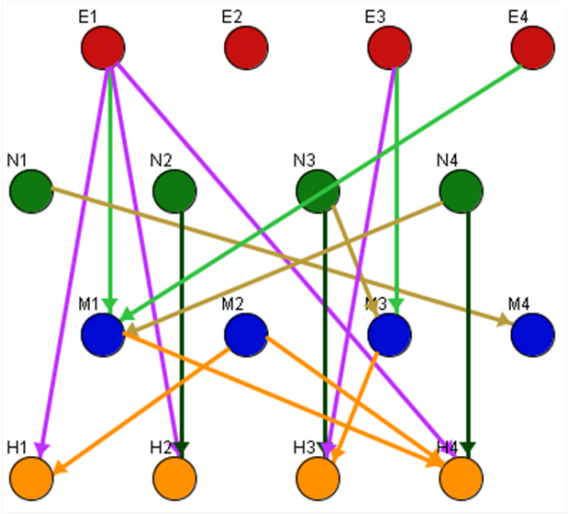


Figura 6.2: Versión gráfica de la *BC* de la tabla 6.1.

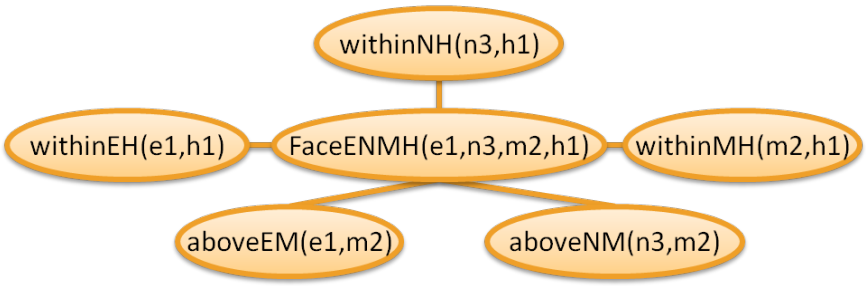


Figura 6.3: Ejemplo de una red de Markov instanciada con un ejemplo. Para el caso de esta imagen, la red es una instanciación con las cuatro regiones: e1, n3, m2, y h1; las cuales fueron obtenidas de una base de conocimiento *BC*.

en una *BC* de primer orden. la RLM al leer la *BC* generará las instancias de redes de Markov (la de la Fig. 6.3 es una de esas instancias) donde se puede realizar la inferencia. En esta tesis se utilizaron dos formas de procesar la inferencia. Un caso es utilizar la inferencia total siguiendo la fórmula 3.7 mencionada en el capítulo 3. Sin embargo, como esta inferencia puede ser lenta y requerir mucho espacio en memoria con una base de conocimiento muy grande, se utilizó el muestreo de Gibbs. En ambos casos la implementación se puede realizar utilizando la aplicación Alchemy[2] [25]. Si adicionalmente se desean visualizar las instanciaciones de las redes de Markov, una opción utilizada en esta tesis fue utilizar la implementación[3] provista por Jaeger en [15].

[2]Este software puede obtenerse de `https://alchemy.cs.washington.edu/`
[3]Esta aplicación puede descargarse en: `http://people.cs.aau.dk/~jaeger/Primula/`

6.4. Resumen

En esta sección se describieron los pasos para trasladar la descripción de una gramática visual a dos modelos relacionales: redes bayesianas relacionales y redes lógicas de Markov. El primer modelo es más flexible en cuanto a la descripción de parámetros (las probabilidades asociadas a las relaciones espaciales) mientras que el segundo busca la simplificación en la representación del conocimiento, aunque los parámetros, al ser manejados como pesos, son un poco menos representativos para describir la incertidumbre. Este capítulo puede verse como una modificación al método propuesto en esta tesis en su última capa. En el siguiente capítulo se mostrará una variación para la primera etapa del modelo propuesto: añadir relaciones temporales a la gramática visual, construyendo una gramática que permita manejar no solamente información espacial provista en imágenes, sino información temporal entre *secuencias* de imágenes.

Capítulo 7

Modelo de reconocimiento de secuencias de imágenes basado en Gramáticas Temporal-Relacionales.

7.1. Introducción

Con el objetivo de proponer una migración de las gramáticas visuales a entornos donde se consideren secuencias de imágenes (por ejemplo, en vídeo), en esta tesis se ha propuesto también una extensión a las gramáticas Simbólico-Relacionales que permiten definir explícitamente relaciones temporales, a fin de dotarle de expresividad para representar información no solamente espacial sobre una imagen, sino también considerar secuencias de imágenes para determinar si dicha secuencia es aceptada o no por una gramática. Puesto que esta extensión es un primer planteamiento hacia el entendimiento de la estructura subyacente en secuencias de imágenes, se propone también un algoritmo que analiza secuencias de imágenes basado en reglas que permite decidir si una secuencia es aceptada por la gramática. El aprendizaje de la gramática a partir de una sola secuencia de imágenes es trivial, sin embargo aprender de varias secuencias, es aún un problema abierto, al igual que aprender una gramática SR de manera general [33].

Para entender mejor el concepto de temporalidad en unión con la espacialidad, podemos imaginar una obra de teatro: hay elementos visuales, personajes, objetos y hay actos en donde cambian la disposición de los personajes y objetos entre uno y otro acto. Una *historia* puede ser contada a partir de describir la disposición espacial y temporal de dichos elementos. Resultaría útil en términos de representación del conocimiento, poder abstraer esta información y compactarla mediante una gramática que describa lo ocurrido, tanto espacialmente como temporalmente. También es deseable no solamente representar este conocimiento, sino poder analizar si un nuevo ejemplo es parecido a lo aprendido a la gramática o no, e incluso más allá: si el nuevo ejemplo no se parece a lo aprendido, poder describir "qué le falta" para parecerse más, por ejemplo: "Al principio de la secuencia de imágenes, la gramática tiene a dos personas paradas y el nuevo ejemplo a analizar sólo tiene a una". Una aplicación inmediata que surge es la capacidad de describir vídeos. Para poder aplicar esta idea, se debe considerar un conjunto de elementos escalable y un conjunto de relaciones espaciales y temporales que se utilizarán para representar el conocimiento. Querer describir la espacialidad a nivel píxel y la transición temporal entre píxeles en un video a través de los distintos cuadros es, aunque posible, una tarea poco escalable por su nivel de granularidad tan fino. Para darse una idea, si se considera un vídeo en alta definición (HD) se obtendrían más de 25 millones de predicados de transición temporal en un segundo. Una opción más manejable sería utilizar regiones de cientos o quizás miles de píxeles (una granularidad menos fina) y hacer las transiciones temporales cuando solamente ocurren cambios en estas regiones y no en los píxeles. Lo anterior podría reducir el número de predicados temporales a, por poner un ejemplo, menos de diez predicados para el mismo segundo.

En este capítulo se muestra un modelo que deriva del método presentado en esta tesis, que incorpora relaciones temporales de manera explícita a la definición dada previamente de gramáticas visuales y que logra explicar, con cierto nivel de detalle, errores en la gramática aprendida. Para ello este modelo integrará una máquina de inferencia basada en reglas, en lugar del manejo de incertidumbre anteriormente usado basado en redes bayesianas. Se detallará la representación del conocimiento, la definición de la gramática visual temporal-relacional y las reglas usadas por la

máquina de inferencia.

7.2. Gramática Visual Temporal-Relacional

El objetivo de este formalismo de gramática es describir secuencias de imágenes a través de relaciones espaciales y temporales de manera explícita. La gramática propuesta admite su utilización en entornos donde se maneje tanto información espacial como temporal, tales como secuencias de vídeos. Aunque en la literatura hay trabajos que utilizan gramáticas visuales [64, 72, 79, 21], éstos no consideran relaciones temporales de manera explícita, puesto que están enfocados en imágenes o diagramas estáticos. Otros trabajos describen gramáticas temporales como en [26, 93]. Sin embargo, estas gramáticas temporales[1] no incluyen todas las posibles relaciones temporales que puede haber de acuerdo a las relaciones espacio-temporales propuestas por Allen [2]. Esta imposibilidad radica en que, al no ser relacionales, sólo admiten producciones temporales a la izquierda o derecha de sus símbolos, impidiendo especificar si la relación temporal es "inmediatamente seguido por", "seguido por (con traslape)" o "seguido por (sin traslape)". Por otra parte, aunque se afirma en [33] que las gramáticas simbólico–relacionales pueden incluir relaciones de tipo temporal, esta escritura podría ser confusa debido a que no se definen de manera explícita en las gramáticas SR cuáles predicados se corresponden con relaciones espaciales y cuáles con relaciones temporales. En contraste, en el formalismo propuesto en este capítulo el manejo de las relaciones temporales está separado de las relaciones espaciales, haciendo más simple el proceso de escritura de la gramática y en particular, de su inferencia usando reglas.

Se requieren ciertos cambios en las gramáticas SR para poder describir el formalismo de gramáticas Temporal-Relacionales. El objetivo de crear este nuevo formalismo radica en tratar de describir explícitamente relaciones de tipo temporal en una gramática ya que, aunque las gramáticas SR pueden incluir relaciones temporales, no incluyen una línea de tiempo asociada a la ocurrencia de los eventos, además de que al momento de analizar gramaticalmente un ejemplo, las gramáti-

[1] En el caso de [26], la gramática temporal implica restricción en las producciones P, lo cual se aleja un poco del objetivo de representar transiciones temporales entre los elementos terminales o no terminales de la gramática.

cas SR darían tratamiento igual a todas las relaciones encontradas. A continuación se presenta la definición formal de una gramática visual temporal-relacional (o gramática TR, para abreviar).

Definición. Una gramática TR es una tupla: $\mathscr{G}_T = (V_N, V_T, V_R, S, P, R)$. La definición de $\mathscr{G}_T$ es similar como en las gramáticas SR. Las gramáticas TR incluyen las relaciones temporales en V_R. Sin embargo, tiene un diferente formalismo en las reglas de producción. En este sentido, P es un conjunto finito de reglas etiquetadas llamadas producciones s de la forma:

$$l : Y^0 \rightarrow \langle \mathbf{M}, \mathbf{R} \rangle$$

donde:

- l es un entero que etiqueta de manera única cada producción de tipo s.
- $\langle \mathbf{M}, \mathbf{R} \rangle$ es una sentencia sobre V_R y $V_N \cup V_T$
  - **M** es un conjunto de *s-ítems* de la forma (v, t, i) con $v \in V_N \cup V_T$, t es un número natural para describir el cuadro visual al que pertenece el símbolo v y el número natural i es usado para distinguir diferentes instancias de un mismo símbolo en un mismo cuadro visual.
  - **R** es un conjunto de *r-ítems* de la forma $r(X_m^i, Y_n^j)$, con $X_m^i, Y_n^j \in \mathbf{M}$ y $r \in V_R$
- También se cumple que $Y \in V_N$, $Y_0 \notin \mathbf{M}$

Es de observarse que ahora hay dos índices asociados para cada *s-ítem* (símbolo). Los superíndices se usan para definir instancias de un mismo símbolo que está en $V_T \cup V_N$, de igual manera que en gramáticas SR; y los subíndices que se usan para saber en que parte de la línea de tiempo se ubica el símbolo. Si en alguna aplicación, no ocurren dos instancias de objetos iguales, se pueden omitir los superíndices, ganando claridad de lectura. Casos como dos personas o más en una imagen de la secuencia sí exige los superíndices, en el entendido de que ambos elementos pertenecen al tipo *persona*. A manera de ejemplificar un caso donde no hay dos instancias iguales, se muestra

el caso de prendas de ropa usadas por una sola persona. En este caso, como no ocurre que una persona se ponga dos camisas iguales (o dos pantalones iguales) los superíndices se podrán omitir. Por ejemplo: $A^0 \rightarrow Next(Shirt_1^2, Jacket_2^2)$ se escribirá como: $A \rightarrow Next(Shirt_1, Jacket_2)$, donde una camisa (de una persona) pertenece al cuadro "1" de la secuencia y una chamarra pertenece al cuadro "2" de la secuencia visual. La añadidura de relaciones temporales puede combinarse en una misma regla, aunque para clarificar la construcción de las reglas sólo se tendrán reglas que incluyan relaciones temporales y reglas que solamente incluyan relaciones espaciales. Las reglas de reescritura, al igual que para el modelo de visión, no son usadas, de modo que: $R = \varnothing$.

En este formalismo también se admiten reglas disyuntivas de tipo *Or* para explicar opciones de objetos en un cuadro determinado de la secuencia: siguiendo con el ejemplo de prendas de ropa, si imaginamos una secuencia donde una persona toma algún tipo de abrigo antes de salir, esa persona puede vestir o una *chamarra* o un *suéter*. Se trata de una regla de tipo *Or* pues tiene dos alternativas.

El subíndice adicional permite manejar las relaciones temporales de manera separada. La composición puede ser operada al nivel de elementos terminales o al nivel de elementos no terminales (en meta-reglas). De acuerdo a los experimentos realizados, hay mejor nivel de detalle en la explicación de las secuencias cuando la composición temporal se realiza a nivel de elementos terminales. Para ejemplificar lo anterior, se muestran dos casos de composición: el primero ocurre en meta-reglas y el segundo alcanza mayor nivel de detalle, a nivel de elementos terminales. Tomando el ejemplo de prendas vestir, se tiene una gramática TR $\mathscr{G}_t$ definida por:

$$\mathscr{G}_T = (\{Seq, First, Second\}, \{tshirt, poloshirt, jeans\},$$
$$\{above, aligned, Next\}, Seq, S, \varnothing).$$

donde S está dado por las siguientes reglas de producción:

$$1 : First \rightarrow \langle \{tshirt_1, jeans_1\}, \{above(tshirt_1, jeans_1), aligned(tshirt_1, jeans_1)\} \rangle$$
$$2 : Second \rightarrow \langle \{poloshirt_2, jeans_2\}, \{above(poloshirt_2, jeans_2), aligned(poloshirt_2, jeans_2)\} \rangle$$
$$3 : Seq \rightarrow \langle \{First_1, Second_2\}, \{Next(First_1, Second_2)\} \rangle$$

nótese que los subíndices están definidos de acuerdo a la definición TR. Las dos definiciones de pantalones de mezclilla ($jeans_1$ y $jeans_2$) que aparecen en este ejemplo están ubicadas en distintos cuadros, de modo que se tratan como objetos distintos. Este ejemplo crea composición de manera jerárquica al nivel de los elementos no terminales (*First* y *Second* se componen en la regla tres). No queda claro cual es el acomodo de las prendas entre el primer y segundo cuadro de la secuencia visual. Para obtener una composición más explícita entre los terminales, se puede reformular las reglas de producción de la gramática TR con más detalle:

$$
\begin{aligned}
1 : First &\rightarrow \langle \{tshirt_1, jeans_1\}, \{above(tshirt_1, jeans_1), aligned(tshirt_1, jeans_1)\}\rangle \\
2 : Second &\rightarrow \langle \{poloshirt_2, jeans_2\}, \{above(poloshirt_2, jeans_2), aligned(poloshirt_2, jeans_2)\}\rangle \\
3 : UpperT &\rightarrow \langle \{tshirt_1, poloshirt_2\}, \{Next(tshirt_1, poloshirt_2)\}\rangle \\
4 : LowerT &\rightarrow \langle \{jeans_1, jeans_2\}, \{Next(jeans_1, jeans_2)\}\rangle \\
5 : Seq &\rightarrow \langle \{First_1, Second_2, UpperT_*, LowerT_*\}, \{\varnothing\}\rangle
\end{aligned}
$$

donde $UpperT$ y $LowerT \in V_N$. Con esta reformulación es más fácil ver en detalle las transiciones entre los objetos de cada cuadro. Como $UpperT$ y $LowerT$ son no terminales que operan entre dos cuadros visuales, se usa el símbolo estrella (*) en el subíndice en lugar de los cuadros donde ellos operan. La secuencia completa está definida con la aparición de todos los elementos no terminales que aparecen en la regla número cinco (*Seq*, el símbolo inicial). Notar que es más fácil comprender que la playera tipo polo (*poloshirt*) debe ir después de (*Next*) la camiseta (*tshirt*), así como que los dos pantalones de mezclilla situados en los cuadros uno y dos ($jeans_1$ y $jeans_2$), deben coincidir. Esta gramática es la abstracción de una secuencia visual compuesta de dos imágenes observada en la Fig. 7.1.

Figure 7.1: Ejemplo de una secuencia visual de una persona con un sólo cambio en las prendas de ropa: una playera tipo polo se coloca después de una camiseta. No hay cambios en pantalón, lo cual también es correcto.

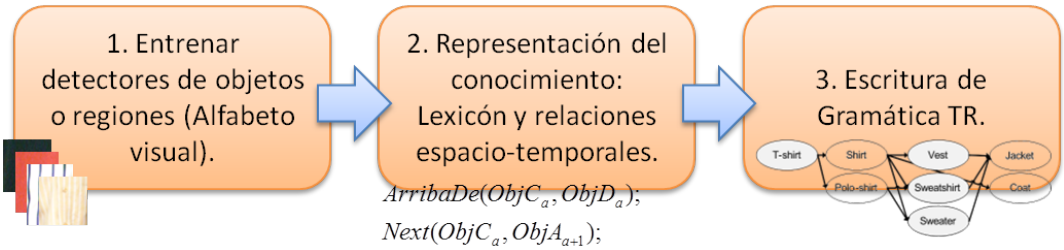


Figure 7.2: Esquema general de la fase de entrenamiento. En la primera etapa el modelo extrae características visuales de objeto del dominio y entrena los detectores mediante aprendizaje automático. En la segunda etapa se representan las relaciones espaciales y temporales encontradas en las secuencias visuales. Finalmente se escribe una gramática que incluye todas las formas correctas de una secuencia.

7.3. Reconocimiento de secuencias visuales con gramáticas TR

El modelo general de reconocimiento de secuencias visuales difiere un poco del modelo general de visión del capítulo 5. El modelo propuesto de secuencias visuales, a diferencia del modelo general de visión, no incluye una etapa que maneje incertidumbre. Los diagramas generales de entrenamiento e inferencia se ilustran en las Figs. 7.2 y 7.3. De manera resumida se realizan los siguientes pasos en el modelo:

1. Se entrenan detectores de objetos de mediano o alto nivel que constituyen el alfabeto visual.

2. Se construyen predicados que representan el conocimiento espacial y temporal asociado a las secuencias de imágenes, por ejemplo: “objeto o región *a* aparece arriba de y alineada con respecto de objeto o región *b*, donde *a* y *b* son objetos que pueden estar en un mismo

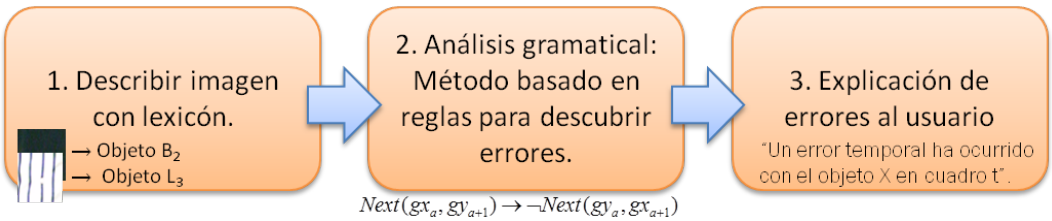


Figure 7.3: Esquema de la fase de prueba. En la primera etapa cada imagen se describe en términos del lexicón generado en la gramática. En la segunda fase un método de análisis gramatical o inferencia basada en reglas que descubre errores en cada secuencia (si los hay). Finalmente se transforma el resultado dado por el método en una respuesta que da el sistema al usuario.

cuadro visual o en dos cuadros consecutivos t y $t+1$. La representación es análoga a la usada en lógica de predicados, el enunciado anterior puede quedar como: $arribaDe(a_1, b_2) \wedge alineado(a_1, b_2)$.

3. Se escribe una gramática que explica las secuencias correctas. En este paso de escritura se incluye el lexicón con el que trabaja la gramática. Escribir la gramática a partir de una sola secuencia visual es automático a partir de la representación visual del paso anterior. Para construir una gramática a partir de más de dos secuencias, se realiza de forma manual. Aunque en esta tesis se dio un algoritmo para aprender automáticamente una gramática visual a partir de ejemplos, no se propuso un algoritmo de aprendizaje de gramáticas TR a partir de ejemplos.

4. En la etapa de inferencia, se procesan las secuencias de imágenes en términos del lexicón asociado a la gramática. Después se utiliza un método basado en reglas para decidir si la secuencia es aceptada por la gramática o no. Cuando no es aceptada, se detecta el error y éste es indicado al usuario. Con el sistema de inferencia basado en reglas se pueden detectar tres tipos de error:

   a) "Hay un error temporal en la secuencia porque el objeto o región G_a aparece antes que el objeto o región G_b" (error temporal).

   b) "Hay un error espacial porque el objeto G_a aparece *arribaDe* el objeto G_b y esto no está señalado en la gramática" (error espacial).

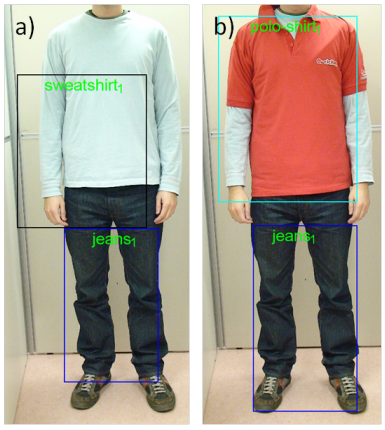


Figure 7.4: Ejemplo de error temporal: una playera normalmente no va después de una sudadera. La gramática debe aceptar el caso opuesto: una sudadera sí puede ir *después de* una playera tipo polo.

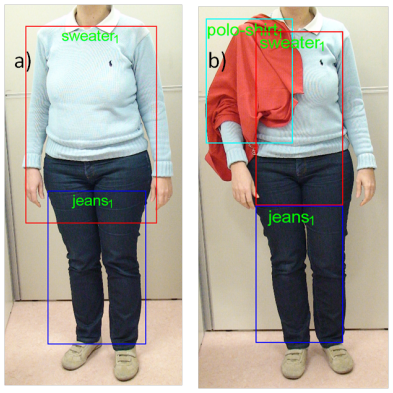


Figure 7.5: Ejemplo de error de tipo espacial: la ropa está parcialmente puesta. En este caso la gramática no debe tener reglas en que dos prendas coexistan en un cuadro de la secuencia.

c) "El objeto G_a pertenece a un tipo de objeto no reconocido por la gramática" (error de lexicón).

El primer error involucra error al observar los predicados temporales. El segundo error equivale a no encontrar una relación espacial en una sola imagen. El último error equivale a encontrar un objeto *no esperado* o *no descrito*, por la gramática TR.

Utilizando el ejemplo de secuencias para prendas de vestir, ejemplos de los errores temporales y espaciales se ilustran en las Figs. 7.4 y 7.5. Un error de lexicón es sencillo: los detectores visuales pudieron haber encontrado un rostro, pero éste no se definió en la gramática. A continuación se presenta el método en mayor detalle.

7.3.1. Alfabeto visual

Inicialmente el método a usar para detectar regiones, puede ser el utilizado en la sección 5.2.3, es decir aprender regiones u objetos mediante recuadros que utilizan aprendizaje automático. Sin embargo, dada la naturaleza de que las aplicaciones en secuencias visuales, es posible que se sepa, de manera previa, algo acerca de los objetos a reconocer en la misma secuencia. Si se desea analizar secuencias de personas caminando por un pasillo, pueden usarse detectores de personas, mientras que si se busca crear una gramática acerca del movimiento permitido de vehículos en una avenida se pueden manejar detectores de vehículos desde un principio. Como normalmente se desea reconocer el objeto en todas las imágenes de la secuencia visual, los detectores deben ser más bien específicos. Por lo regular para la detección de objetos específicos se utilizan modelos basados en puntos de interés local [69], o una combinación de bordes, forma, textura o características globales [77, 86, 22]. En este sentido el método de aprendizaje de recuadros propuesto se puede adaptar según la aplicación requerida.

7.3.2. Representación del conocimiento

La manera de representar el conocimiento visual y temporal es similar a la presentada en la sección 5.3, añadiendo el caso de predicados que involucran temporalidad.

Al igual que en el modelo de reconocimiento de objetos aquí se busca representar la información visual en lógica de predicados de modo que se pueda trasladar esa información a una gramática (en este caso, la gramática TR propuesta). Si se tiene el ejemplo de la Fig. 7.6 se desea señalar por un lado la espacialidad de los objetos en cada imagen de la secuencia así como señalar la temporalidad que existe entre los objetos para todos los cuadros (para este ejemplo, sólo dos cuadros). Expresando esta información en términos de relaciones espaciales se pueden escribir las siguientes relaciones en lógica de predicados:

$Above(shirt_1, jeans_1)$: dos objetos que sostienen la relación arriba de en el primer cuadro.

$Next(shirt_1, jacket_2)$: dos objetos que sostienen una relación temporal de traslape o suplantación entre dos cuadros consecutivos.

Figura 7.6: Ejemplo de una secuencia visual y sus relaciones espaciales y temporales asociadas. La representación del conocimiento del lado derecho permite compactar la información visual y procesar esta información por una gramática de tipo TR.

$Above(jacket_2, jeans_2)$: nuevamente dos objetos en un sólo cuadro sosteniendo una relación espacial.

También se puede definir el lugar espacial que ocupa el objeto con respecto de la imagen. Lo anterior se puede realizar con predicados de aridad uno, por ejemplo: $IsUpper(Shirt_a)$ donde a es el número del recuadro, *Shirt* es el nombre del objeto e $IsUpper$ es un predicado unario que describe la posición del centro del recuadro con respecto de la imagen.

Para el caso de las transiciones se usa una sola relación temporal llamada *siguiente* (*Next*). Esta relación significa la sustitución de una prenda por otra en dos recuadros sucesivos. En la sección 2.4 se analizaron brevemente estas relaciones temporales. La relación *Next* tiene la forma: $Next(A_f, B_{f+1})$, donde A y B son prendas ubicadas en dos recuadros sucesivos. Para definir esta relación a partir de las imágenes se utiliza una razón de traslape entre los objetos A y B, se debe cumplir que: $(A_f \cap B_{f+1})/(A_f \cup B_{f+1}) \geq \varepsilon$, donde $\varepsilon \in [0 \ldots 1]$. Esta razón de traslape se utiliza ya que las secuencias de imágenes y los pacientes tienen un ligero desplazamiento, haciendo que la superposición de prendas en la secuencia no sea exacta. En otras palabras, usar un umbral más alto (cercano a uno) sugiere que la cámara ni los pacientes se mueven entre una puesta de ropa y otra. La intersección se da por el posicionamiento de las imágenes en la secuencia.

7.3.3. Construcción de la Gramática Temporal-Relacional

Las gramáticas SR pueden admitir predicados que sugieran una relación temporal entre los símbolos. Sin embargo esta aceptación implícita pierde posicionalidad de los objetos, haciendo más difícil el proceso de análisis gramatical para evaluar si una secuencia es aceptada por la gramática o no. A diferencia, la gramática TR es más explícita: permite distinguir fácilmente relaciones espaciales y temporales. También se pueden usar las relaciones de tipo *Or* para mostrar alternativas en la construcción de la gramática. Ejemplos como los siguientes son válidos:

$$
\begin{aligned}
G &\rightarrow \langle \{shirt_a, vest_{a+1}\}, \{Next(shirt_a, vest_{a+1})\}\rangle, \\
G &\rightarrow \langle \{shirt_a, sweater_{a+1}\}, \{Next(shirt_a, sweater_{a+1})\}\rangle, \\
G &\rightarrow \langle \{shirt_a, sweatshirt_{a+1}\}, \{Next(shirt_a, sweatshirt_{a+1})\}\rangle, \\
&\ldots
\end{aligned}
$$

en donde los subíndices a, $a+1$, como se ha mencionado antes, indican que los objetos se encuentran en imágenes distintas pero sucesivas. Este ejemplo particular sugiere que *después* de vestir una camisa, una persona puede vestir un chaleco, o un suéter, o una sudadera, de manera alternativa.

Para el caso de las relaciones espaciales, también hay producciones alternativas, en el siguiente ejemplo se muestra un caso:

$$
\begin{aligned}
G &\rightarrow \langle \{tshirt_a, trousers_a\}, \{Above(tshirt_a, trousers_a)\}\rangle, \\
G &\rightarrow \langle \{shirt_a, jeans_a\}, \{Above(shirt_a, jeans_a), Aligned(shirt_a, jeans_a)\}\rangle, \\
G &\rightarrow \langle \{sweater_a, jeans_a\}, \{Above(sweater_a, jeans_a), Aligned(sweater_a, jeans_a)\}\rangle, \\
&\ldots
\end{aligned}
$$

donde todas las relaciones espaciales se consideran en el mismo recuadro o imagen. Estos ejemplos

fueron recuperados de actividades de vestimenta consideradas correctas. A partir de estas reglas, se construyó de manera manual la gramática que comprende las combinaciones correctas de vestirse, para la base de datos utilizada [74].

7.3.4. Análisis gramatical de la secuencia por la gramática TR

El objetivo es no solamente detectar errores sino también explicar a un usuario el tipo de error encontrado. Hay tres tipos de error básicos que se han considerado. Estos son:

1. Error temporal: el orden de aparición de los objetos en la secuencia es equivocado o no acorde a la gramática.

2. Error espacial: en un cuadro determinado de la secuencia, hay un objeto que tiene una disposición espacial errónea con respecto de otro.

3. Error de lexicón: en un cuadro de la secuencia un objeto no es reconocido por el lexicón.

Los errores temporales pueden explicarse mediante las relaciones temporales. Los errores espaciales pueden explicarse mediante una combinación de relaciones espaciales y temporales. Los errores de lexicón se pueden explicar a través de un problema con la detección visual de un objeto o región que no se espera.

7.3.4.1. Manejo de error temporal

La gramática detecta esta falla con solamente las relaciones temporales: si el ejemplo a evaluar tiene una relación temporal que no aparece en la gramática, entonces el algoritmo fallará e indicará un error de tipo temporal. En otras palabras no se esperan reglas "opuestas" a las observadas en la gramática. Visto desde el sistema basado en reglas el modelo ejecuta reglas como la del siguiente ejemplo:

$$Next(Objx_a, Objy_{a+1}) \rightarrow \neg Next(Objy_a, Objx_{a+1}), \tag{7.1}$$

donde a es el recuadro donde el analizador está operando.

7.3.4.2. Manejo de error espacial

Este error se maneja usando información temporal y espacial. Primero se evalúa si el objeto o región está correctamente ubicado. Esto puede hacerse usando predicados de aridad uno. Si una regla de la gramática describe lo siguiente:

$$G \rightarrow \langle \{Objx_a, Objy_a\}, \{Above(Objx_a, Objy_a), isUpper(Objx_a), isLower(Objy_a)\} \rangle, \quad (7.2)$$

donde $isUpper$ e $isLower$ son predicados de aridad uno que exigen que los objetos se ubiquen en cierta posición relativa del cuadro. La regla que marca un error de posicionamiento del objeto es:

$$isUpper(Obj_a) \rightarrow \neg isLower(Obj_a). \quad (7.3)$$

Otros errores espaciales se pueden manejar utilizando las siguientes reglas:

$$Next(Objx_a, Objy_{a+1}) \rightarrow \neg Next(Objx_a, Objx_{a+1}) \quad (7.4)$$

$$Next(Objx_a, Objy_{a+1}) \rightarrow \neg Left(Objx_{a+1}, Objy_{a+1}), \quad (7.5)$$

donde la regla 7.4 significa que no se esperan objetos traslapados y la regla 7.5 significa que el objeto x debe desaparecer en cuadro $a+1$ y no lo hizo, puesto que está quizás en oclusión o colisión con el objeto y.

7.3.4.3. Manejo de error de lexicón

Este problema identifica objetos que aparecen en la secuencia visual pero que no se esperan en la gramática. Se maneja con la regla: $\exists Objx | r(Objx, Objy) \in Sec$ donde $Objy \in V_T$ pero $Objx \notin V_T$

y *Sec* es la secuencia visual que se está analizando por la gramática $\mathscr{G}_T$. Desafortunadamente, este problema puede estar más relacionado con un fallo en los detectores usados en el alfabeto visual, es decir, la secuencia no tiene un elemento no esperado, sino que hay un falso positivo por parte de los detectores.

7.4. Resumen

En este capítulo se describió un modelo de representación del conocimiento visual en secuencias de imágenes, incorporando relaciones temporales en una gramática denominada de tipo temporal-relacional. Este método permite reconocer secuencias de imágenes aceptadas por la gramática. Se puede ver como una extensión al modelo descrito en el capítulo 5, donde el primero descubre objetos en una imagen, mientras el propuesto aquí descubre secuencias de objetos sobre varias imágenes. Ambos modelos utilizan una gramática para representar el conocimiento, así como un alfabeto visual y una definición de lexicón. La diferencia radica en que aquí se propuso una variante para poder añadir relaciones temporales de forma explícita a la gramática visual. La construcción de la gramática solamente es automática si se construye a partir de un ejemplo. La inferencia se realizó mediante reglas, a diferencia del caso anterior, donde se utilizó manejo de incertidumbre con redes bayesianas. El método presentado en este capítulo se probará en el capítulo siguiente en un caso particular de análisis de vestimenta correcta. Esta aplicación muestra cómo puede simplificarse el conocimiento visual en predicados estructurados a través de la gramática visual.

Capítulo 8

Experimentos y Resultados

8.1. Introducción

En este capítulo se describen los experimentos realizados con cada uno de los tres métodos descritos en los capítulos 5, 6 y 7. Para el primer caso se hicieron pruebas utilizando los diferentes alfabetos visuales, así como como el mejoramiento del lexicón reduciendo su dimensionalidad y haciendo una comparación con trabajos composicionales que buscan reconocer objetos. El objetivo de esto es mostrar que la representación generada por una gramática ha sido útil para reconocer objetos, de modo que dicha representación es en efecto una abstracción del objeto que se desea aprender. Al hacerlo de esta manera se prueba por un lado la capacidad de la gramática de representar información de un objeto visual y por otro lado, se muestra la utilidad del algoritmo que aprende gramáticas a partir de ejemplos. En el segundo caso se hicieron pruebas con dos modelos relacionales probabilistas a fin de validar el procedimiento de transformación de la gramática a redes bayesianas observando si hay o no pérdida de información al hacer tanto el aprendizaje estructural como paramétrico. Finalmente, para el método que propone una gramática temporal-relacional, se hicieron pruebas en una base de datos que contiene secuencias visuales que describen acciones de vestimenta correctas o incorrectas (es correcto vestir un abrigo después de una playera, pero es incorrecto ponerse una camiseta después de una sudadera o una chamarra). En estas

pruebas se comparó lo logrado por la gramática TR con un trabajo que utilizó etiquetas RFID en la ropa para poder detectar estas fallas. En las siguientes secciones se detallan más cada uno de los experimentos realizados.

8.2. Entorno experimental

En esta sección se describen las bases de datos que se utilizan, las medidas empleadas en los experimentos realizados y las comparativas con otros trabajos. [116, 65, 32, 22, 74]. En la selección de las bases de datos hubo un interés mayor por aquellas que presentasen objetos con cierto poder de descomposición.

8.2.1. Bases de datos utilizadas

Para realizar los experimentos se utilizaron las siguientes bases de datos:

- Caltech 101 [29]. Es una base de datos que contiene 101 categorías de objetos obtenidas de Internet. Imágenes como aviones, rostros, autos de lado, bonsais, delfines, jarrones y otras son contenidas en esta base de datos. Las imágenes tienen un tamaño, por lo regular, alrededor de los 320 x 240 píxeles. Estas imágenes, al ser obtenidas de Internet, tienen fondo diverso. En ellas se pueden representar objetos naturales o artificialmente creados.

- Caltech 256 [49]. Esta base de datos contiene 256 categorías de objetos diversos obtenidas de Internet. Algunas categorías son tomadas de la base de datos Caltech 101, mientras que la mayoría son nuevas. Se incluyen otras categorías de animales y objetos de uso común por humanos, tales como computadoras, discos compactos y otros. En algunos casos la categoría se repite pero incluyendo más ejemplos.

- Base de Datos ETH [65]. Este conjunto de datos contiene 8 categorías de objetos visuales en diferentes puntos de vista. Son objetos naturales en donde se tiene un fondo azul uniforme. Estas categorías pueden englobarse como frutas del tipo manzana o pera y juguetes de

plástico como coche o vaca.

- INRIA Personas [22]. Esta base de datos contiene imágenes de peatones. El fondo es diverso debido a que son tomadas en exteriores.

- INRIA Caballos [31]. Esta base de datos contiene imágenes de caballos vistos de lado en su mayoría. Estas imágenes son a diversas resoluciones, aunque usualmente no más allá de alta definición (HD).

- Poses humanas [10]. Esta base de datos contiene imágenes de personas en diversas poses con etiquetas que describen la posición de algunos puntos de interés de las personas, tal como hombros, rodillas, pies, ojos, etc.

- Monitoreo de Actividad de Vestimenta [74]. Esta base de datos contiene secuencias de imágenes de personas con diferentes prendas de vestir. Cada secuencia tiene información sobre si la secuencia es correcta o incorrecta y el tipo de error que tiene (prenda puesta al revés, dos prendas en orden inverso o prenda puesta a medias).

Como nota en algunas bases de datos (como Caltech), no se usaron en su totalidad, solamente un subconjunto de las categorías de cada una. En otras, como la base de datos de actividades de vestimenta, sólo se usó la información visual, ignorando información adicional de metadatos.

8.2.2. Medidas de Evaluación

Para el análisis cuantitativo se utilizaron las siguientes medidas:

Matriz de confusión. En su caso más sencillo, la matriz de confusión de clasificación binaria (dos clases o categorías) es un recuadro compuesto de cuatro valores los cuales cuantifican los aciertos y errores de las predicciones modelo con respecto del valor de la realidad. Si una imagen fue clasificada positivamente y es la clase correcta, se habla de un *verdadero positivo* o *VP*. Si fue clasificada negativamente y también fue la clase correcta, se habla de un verdadero negativo o VN.

Cuando ocurren errores, hay dos casos: se predice clase positiva pero en realidad es clase negativa, se habla de *falso positivo* o FP. El segundo caso es cuando se predice clase negativa pero en realidad es clase positiva, se habla de un *falso negativo* o FN. Se puede visualizar como: $\begin{bmatrix} VP & FP \\ FN & VN \end{bmatrix}$

En su caso más general la matriz de confusión describe errores para varias categorías, de modo que la diagonal principal contiene los aciertos de las predicciones y cualquier número de la matriz ubicado en una posición $m_{i,j}$ indica que hubo un error de clasificación: clasificó erróneamente a un ejemplo como de la categoría j, cuando en realidad pertenece a la categoría i.

Precisión. Se define como la razón de verdaderos positivos sobre la suma de verdaderos positivos con los falsos positivos (precisión = $\frac{VP}{VP+FP}$). Esta medida determina el porcentaje de ejemplos correctamente clasificados como positivos con respecto del total que fueron etiquetados así por el sistema.

Recuerdo. Se define como la razón de verdaderos positivos sobre la suma de verdaderos positivos con falsos negativos (recuerdo = $\frac{VP}{VP+FN}$). Esta medida determina la razón de verdaderos positivos devueltos por el sistema con respecto del total de ejemplos positivos que hay en el universo.

Exactitud. Se define como la razón de verdaderos positivos y negativos sobre todo el universo de ejemplos (exactitud = $\frac{VP+VN}{VP+FN+FP+VN}$). Esta medida ayuda a determinar la cantidad de ejemplos correctamente clasificados con respecto del total.

Medida F. Esta medida trata de establecer un compromiso entre la precisión y el recuerdo, evitando caer en una sobre representación de alguno de los dos. Se define como:

$$MedidaF = \frac{2(precision)(recuerdo)}{precision + recuerdo}. \tag{8.1}$$

Curva de característica operativa del receptor[1]. Es una gráfica que permite visualizar la sensibilidad (o recuerdo) contra la especificidad, según se varíe un umbral de clasificación. Esta curva tiene uso sobre todo en sistemas de clasificación binaria donde la etiqueta de pertenencia a la clase o no, es un valor de probabilidad. Se ha usado esta curva dado que se utiliza un valor de probabilidad devuelto por la red bayesiana que se se transforma en una etiqueta de pertenencia o no a la clase aprendida una vez considerado un umbral de aceptación. Dado que este tipo de curva sugiere que se tiene un mejor clasificador cuando en el gráfico aparece más pegada hacia arriba y hacia la izquierda, a veces se mide la integral de dicha curva, a este indicador se le conoce como valor AUC[2] de la curva ROC. Aunque en los resultados mostrados en esta tesis no se realiza el cálculo de la curva AUC, la intención es mayormente ilustrativa.

Según la aplicación puede ser deseable tener un modelo que tenga mejor precisión que recuerdo, o a la inversa, o un compromiso entre ambos. Por ello se han utilizado cada una de estas medidas para evaluar los resultados obtenidos de modelo propuesto en esta tesis. Debido a que en algunos casos los experimentos realizados dependen de los datos de entrenamiento que se tomen, en los experimentos realizados que así son, se realiza una replicación cruzada de cinco segmentos, de modo que se replica el experimento cinco veces, tomando diferentes ejemplos de entrenamiento cada vez. Finalmente estas pruebas fueron realizadas en un equipo con procesador intel core i7-2630QM y 8GB de memoria RAM sin ninguna optimización en código. Los modelos se implementaron en MatLab.

8.3. Experimentos usando Redes Bayesianas

En esta sección se presentan los resultados obtenidos usando el método propuesto con redes bayesianas (Cap. 5). Por un lado, se ha construido un modelo muy básico que a partir de una gramática escrita manualmente y un lexicón descriptivo construye un modelo basado en redes bayesianas que hace inferencia para reconocer objetos. Este es el caso de prueba de concepto del modelo propues-

[1] También denominada como curva *ROC*, del inglés Receiver Operating Characteristic.

[2] Del inglés area under curve, es decir, el área bajo la curva, que equivale a la integral de la curva ROC.

to. Por otro lado, se construyó un modelo que permite a partir de un conjunto de entrenamiento, aprender una gramática y su lexicón de forma automática. Este experimento mostró que el modelo es capaz de representar conocimiento visual a partir de una gramática y además, usar este conocimiento para inferencia. En otros experimentos se mostrarán escalamientos algo más amplios.

8.3.1. Experimento 1: factibilidad de reconocimiento visual

Hipótesis del experimento. Una gramática visual descrita manualmente sirve para detectar un objeto, previa transformación a una red bayesiana

Objetivo del experimento. Realizar una prueba de concepto para observar la factibilidad del enfoque propuesto en cuanto a representación y reconocimiento de objetos visuales empleando elementos mínimos para la construcción de la gramática.

Condiciones del experimento. En este experimento se construyó un modelo de reconocimiento de objetos empleando los siguientes elementos:

1. Gramática visual descrita manualmente: un ojo.

2. El alfabeto visual fue basado en segmentación (sección 5.2.1).

3. Se utilizó un lexicón descriptivo (sección 5.4.1) de solamente tres tipos de terminales: bordes verticales (0° de orientación Gabor), bordes horizontales (90°) y regiones homogéneas, sin considerar información de color dentro de la descripción del lexicón. La escala de los bordes fue 11 y 13 (sección 5.2.1.1).

4. Se consideraron tres tipos de relaciones espaciales: $IzquierdaDe(v,w)$, $ArribaDe(v,w)$ y $DentroDe(v,w)$.

5. Se utilizó el algoritmo de transformación de la gramática a una red bayesiana descrito en la metodología del capítulo 5.

Tabla 8.1: Valores de probabilidad de reconocimiento del ojo en cada imagen evaluada. Se añaden las probabilidades cuando se eliminó la información de las relaciones espaciales. Se incluyen los tiempos de búsqueda de todas las subconfiguraciones que correspondieron con la imagen.

Clase	Imagen	Sin relaciones espaciales	Con relaciones espaciales	Tiempo
Positivas	Ojo 1	0.70	0.74↑	6.34s
	Ojo 2	0.72	0.76↑	4.3s
	Ojo 3	0.8	0.86↑	7.2s
	Ojo 4	0.95	0.94↓	9.3s
	Ojo 5	0.76	0.81↑	7.5s
Negativas	Vaca	0.72	0.68↓	2.8s
	Casa	0.71	0.69↓	3.6s
	Bicicleta	0.83	0.76↓	2.3s

6. Se utilizó un aprendizaje paramétrico vía estimación subjetiva

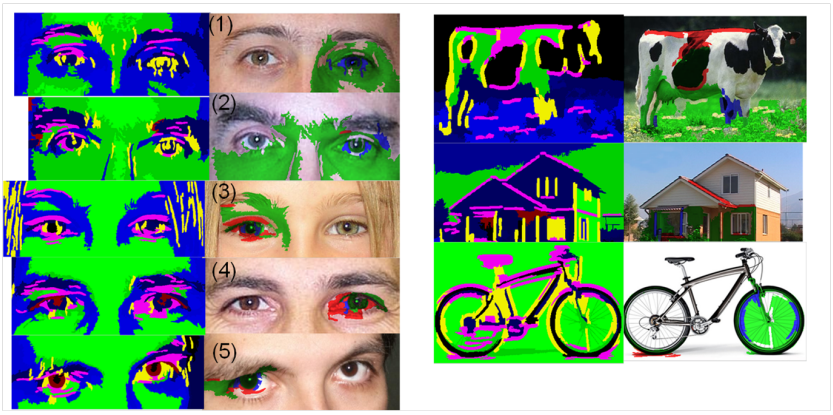
7. Se utilizó el algoritmo de obtención de subconfiguraciones válidas para la red bayesiana obtenida.

8. Se realizó inferencia sobre la red.

9. No se utilizó replicación con los datos.

Resultados. Éste fue el experimento más simple con respecto del resto debido a que fue una prueba de concepto. Por ello no se evaluó en alguna base de datos amplia. Los resultados se condensan en la tabla 8.1. Si se fija un umbral de .75 se obtiene un falso positivo y un falso negativo para esta prueba con ocho ejemplos. Si se quitan las relaciones espaciales y se deja la detección solamente con los terminales, se obtendrían dos falsos negativos en lugar de uno, con el mismo umbral. Adicionalmente se observan probabilidades menos separadas entre los ejemplos positivos y negativos. Las imágenes descritas en la tabla de resultados, se muestran en la Fig. 8.1.

Discusión. La conclusión que se obtiene de este ejercicio es que las relaciones espaciales juegan un papel importante para ayudar a discriminar mejor ante la detección de muchos elementos terminales, puesto que la añadidura de relaciones espaciales ayudó a separar mejor los resultados de la inferencia. A pesar de tener un lexicón muy pequeño, se logró capturar en cierta medida la estruc-



(a) Positivas (b) Negativas

Figura 8.1: Imágenes evaluadas con la gramática del ojo. Las subconfiguraciones mostradas son las que presentaron probabilidad más alta. El algoritmo de segmentación devolvió las imágenes de la izquierda para cada ejemplo. La inferencia devolvió las regiones marcadas en las imágenes del lado derecho. El color resaltado representa las zonas que detectó como un ojo con el valor de probabilidad dado en la tabla 8.1. Los falsos positivos (aunque con baja probabilidad) se podrían eliminar con ayuda de un diccionario visual menos ruidoso.

tura de un ojo. Los resultados de falsos positivos son consecuencia de usar un lexicón muy simple (esta prueba es la que motivó a utilizar un lexicón más amplio y que se aprenda automáticamente). En el caso de la imagen de la vaca las manchas de su piel fueron capturadas en parte como la forma de un ojo. En el caso de la bicicleta se capturó la redondez de la llanta de la bicicleta. Aunque pudiera no ser tan claro en la imagen de una casa, se capturaron los bordes alrededor de una pared. El algoritmo realizar todo el proceso de inferencia en segundos, a pesar de no contar con optimización de código. No se reportan tiempos de entrenamiento debido a que la gramática fue previamente establecida. Al realizar el paso de búsqueda de subconfiguraciones se observó que se detectaban en promedio dos niveles de profundidad de la gramática. Ello sugiere que en la inferencia fue difícil encontrar reglas de gramáticas muy profundas.

8.3.2. Experimento 2: aprendizaje automático de la gramática con ejemplos

Hipótesis. La gramática puede aprenderse de forma automática a partir de ejemplos.

Objetivo. Analizar el grado de aprendizaje de una gramática únicamente (sin incertidumbre) a partir de ejemplos. Los ejemplos son imágenes de prueba que contienen el objeto aprendido.

Condiciones del experimento. Este método probó una variedad más amplia de relaciones espaciales que en el expeerimento anterior:

- Relaciones topológicas: *IzquierdaDe*, *AdentroDe*, *EncimaDe*, *ConectadoCon*, *DisjuntoDe*, *AbarcandoA*, *InvadiendoA*, *CubriendoA*.
- Relaciones difusas: *LejosDe*, *CercaDe*.
- Relaciones de orden o dirección: *LadoIzquierdoDe*, *ArribaDe*.

El alfabeto visual usado está basado en segmentos (de manera similar al experimento 1), con dos orientaciones del filtro gabor, dos escalas (11 y 13), y colores a cinco bits.

Al final se desarrolló un algoritmo que construye reglas de producción de una gramática SR a partir del lexicón visual usando la frecuencia de relaciones sobre el conjunto de imágenes de entrenamiento, mencionado en la sección 5.5.1. Las reglas abstraídas se les fijó el criterio de paro a $U_{c_i} = 2$.

Resultados. Nótese que en este experimento se utilizó de manera más extensa las variantes del método propuesto. Este experimento sirvió para estudiar el tipo de reglas que aprende el modelo de forma automática. Se escogieron las gramáticas que contenían la mayor cantidad de reglas de producción y se observó la cantidad de imágenes sobre las cuales aparecía dicha gramática. Los resultados se resumen en la tabla 8.2. Las imágenes fueron tomadas de la base de datos Caltech-256 [49]. El conjunto de entrenamiento fueron 30 imágenes positivas del objeto y 30 negativas. Las pruebas sobre el conjunto positivo se realizaron sobre 20 imágenes. El conjunto negativo se compuso de la categoría de imágenes de fondo de la misma base de datos, usando 20 imágenes. El tiempo se refiere al costo del aprendizaje de la gramática con los ejemplos de entrenamiento incluyendo la detección de las regiones en las imágenes por el diccionario visual. El número de terminales usados del lexicón fue de aproximadamente cinco (de un espacio de 50 posibles) para los casos evaluados.

Tabla 8.2: Resultados de la detección de gramáticas aprendidas automáticamente a partir de ejemplos.

Gramática	#Reglas	Tiempo	Tasa positivos	Tasa negativos	Precisión	Recuerdo	Terminales usados
Girasoles	5	2.9 min.	5/20	0/20	1	.25	5/50
Velas	4	2.5 min.	5/20	1/20	.83	.25	4/50
Bonsai	6	3.5 min.	4/20	0/20	1	.2	6/50
Aviones	5	2.5 min.	7/20	1/20	.875	.35	5/50

Discusión. Se observa que el modelo sí aprendió gramáticas visuales, aunque con un recuerdo aún limitado. Un ejemplo de una estructura obtenida es:

$G = (\{CNT1, CNT2, CNT3, CNT4, INIT\}, \{CT1, CT2, CT6, CT8, CT9\},$

$\{EncimaDe, ArribaDe, LadoIzquierdoDe, AdentroDe, \}, INIT, S, \emptyset)$.

Donde S contiene las siguientes producciones simbólicas:

$$
\begin{aligned}
1 : INIT^0 &\rightarrow \langle \{CNT1^2, CNT2^2\}, \{ArribaDe(CNT1^2, CNT2^2,)\}\rangle \\
2 : CNT1^0 &\rightarrow \langle \{CNT3^2, CNT4^2\}, \{LadoIzquierdoDe(CNT3^2, CNT4^2)\}\rangle \\
3 : CNT2^0 &\rightarrow \langle \{CT1^2, CT2^2\}, \{EncimaDe(CT1^2, CT2^2)\}\rangle \\
4 : CNT3^0 &\rightarrow \langle \{CT6^2, CT8^2\}, \{AdentroDe(CT6^2, CT8^2)\}\rangle \\
5 : CNT4^0 &\rightarrow \langle \{CT1^2, CT9^2\}, \{LadoIzquierdoDe(CT1^2, CT9^2)\}\rangle
\end{aligned}
$$

donde los terminales ($CT1$, $CT2$, etc.) fueron obtenidos del diccionario visual. La precisión al ser de 1 para un par de casos de esta prueba, explica que la gramática fue capaz de discriminar al objeto aprendido. No obstante, el recuerdo de la gramática, al ser del orden de 0.25, indica que la gramática sí aprende una representación visual aunque deben buscarse mecanismos para mejorar la cobertura. Esto llevó a proponer, por un lado, la inclusión de incertidumbre en el modelo, a fin de ampliar el recuerdo y por otro, un aprendizaje de las reglas gramaticales de tipo *Or* también para ampliar el recuerdo. Las relaciones espaciales difusas y de orden generan muchos predicados en cada ejemplo, de modo que provocan lentitud en el aprendizaje. También se observó que debe hacerse un podado de terminales que no se utilicen en la gramática, o bien, construir un alfabeto

visual más útil para la gramática: si la gramática toma 5 terminales y desecha 45, sugiere que los terminales no utilizados tienen escaso poder discriminativo. Esta prueba al no contar con un modelo de manejo de incertidumbre (lo cual es común en este tipo de problemas) lleva a la necesidad de realizar un experimento más ambicioso: la integración de un MGP que ayude a manejar la incertidumbre además de un aprendizaje paramétrico automático. Adicionalmente, que se integren reglas de tipo *Or* en la gramática, así como realizar una prueba en un dominio más amplio. El siguiente experimento busca cubrir estos aspectos.

8.3.3. Experimento 3: uso de recuadros en base de datos Caltech-256

Hipótesis. El modelo tiene un poder predictivo similar a otro modelo composicional de reconocimiento de objetos .

Objetivo. Evaluar cuantitativamente el poder predictivo de la gramática para reconocer un objeto aprendido sobre varias imágenes de prueba. Hacer una validación concurrente con otros trabajos relacionados que utilicen información estructurada [116], utilizando el mismo repositorio .

Condiciones del experimento. Dentro de las características que se utilizaron del experimento fueron:

- Uso de bases de datos Caltech 101 y Caltech 256.
- El entrenamiento se realizó con 50 ejemplos positivos y 50 ejemplos negativos, para cada categoría.
- El alfabeto visual usado fue el que incorpora aprendizaje automático (sección 5.2.3).
- El clasificador utilizado fue un SVM [20] con kernel lineal, siendo el kernel que presentó mejores resultados en una prueba de validación con tres categorías visuales. Otros clasificadores presentaron un costo computacional mayor o tienen un menor poder discriminativo[3].

[3]Se probaron naïve bayes y redes neuronales.

- La gramática fue aprendida automáticamente (sección 5.5.1) con un umbral de abstracción de $U_{c_i} = 0.05 \times |\mathscr{I}|$. Se incorporaron reglas *Or* (sección 5.5.2.2).

- El lexicón se aprendió automáticamente utilizando también la mejora por reducción del mismo al eliminar palabras redundantes (sección 5.4.3) con un valor de $\xi = 0.1 \times |\mathscr{I}|$.

Resultados. Los resultados muestran que hay un aprendizaje del objeto visual a partir de la gramática y la inferencia muestra resultados que, aunque son cercanos a otros trabajos relacionados, en general hay un contraste en resultados. Los resultados más cercanos en exactitud en comparación con otros trabajos fueron para objetos que presentaron estructuras rígidas. Aquellos objetos que presentan mayor variabilidad en su estructura, presentan resultados más bajos. En algunos casos, el uso de gramáticas que incluyen reglas de tipo *Or*, pueden manejar un poco mejor estos objetos más flexibles.

En este experimento, se evaluaron tres aspectos del lexicón: el primero es un lexicón que directamente se trató de clasificar usando un clasificador bayesiano simple. En el segundo aspecto se añadieron las relaciones espaciales y se transformó el modelo a red bayesiana y ahí se realizó la inferencia. En el tercer caso se redujo el lexicón de acuerdo a la mejora propuesta en el método (reducido por espacialidad y eliminando palabras redundantes), de modo que se pueda observar el impacto que tiene esto en las tareas de reconocimiento. La gráfica de la Fig. 8.2 muestra que hubo una mejora pequeña para las clases que se evaluaron. Las gráficas muestran las curvas *ROC*, de modo que a mayor área por encima de la diagonal principal, mejor compromiso tiene entre precisión y especificidad. Se muestran las curvas para cuatro clases antes y después de añadir la reducción del lexicón.

En la tabla 8.3 se muestran los resultados de aplicar el modelo propuesto en algunas categorías de la base de datos Caltech 256. Se realizaron cinco replicaciones utilizando una segmentación de los datos de cinco pliegues, de modo que pueden generarse gramáticas distintas en cada replicación. El modelo se prueba con las imágenes restantes para cada categoría. La clasificación es binaria: se entrena un modelo separado por cada clase. Un ejemplo de como detecta la gramática partes del

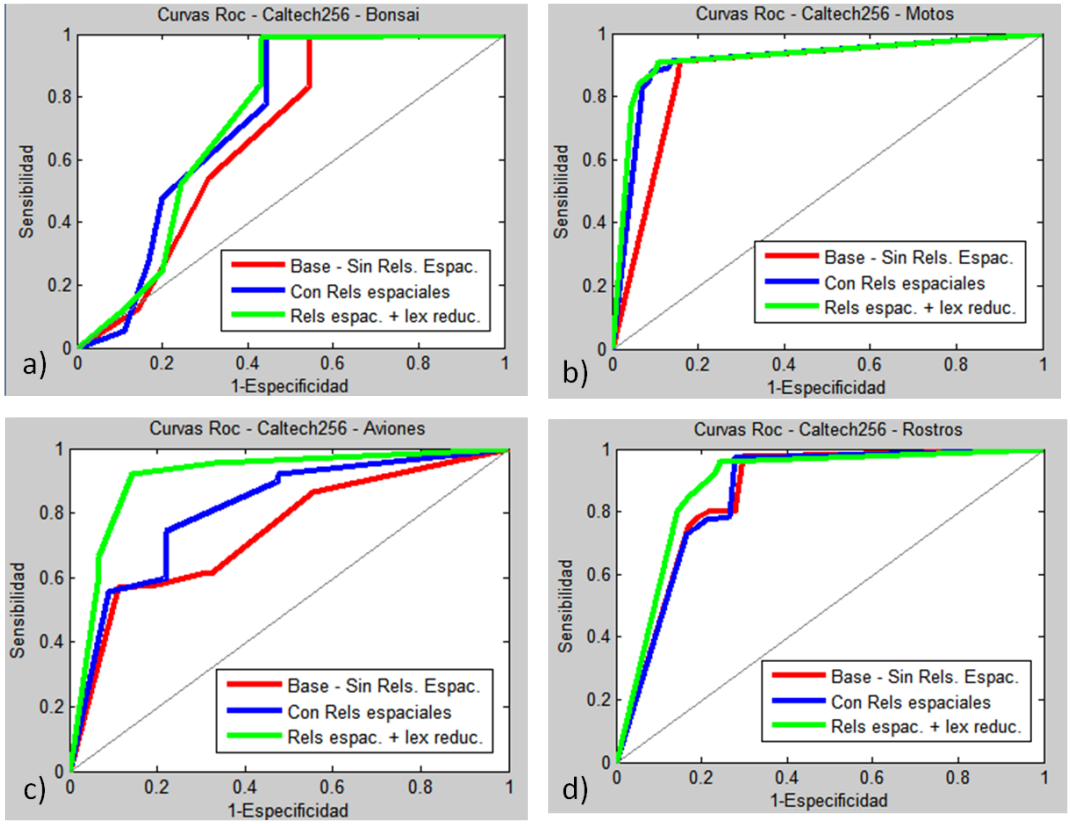


Figura 8.2: Curvas ROC antes y después de la incorporación de relaciones espaciales. Se ilustran tres casos. Línea roja: es el caso base. Solamente utiliza el lexicón sin incorporar la gramática ni las relaciones espaciales. Línea azul: uso de relaciones espaciales con la gramática visual. Línea verde: lexicón reducido con la gramática visual.

Tabla 8.3: Resultados obtenidos en precisión y recuerdo de probar el método propuesto con varias categorías de la base de datos Caltech 256. En negritas se destacan los resultados más altos según la categoría visual.

7 clases	rostros	motos	aviones	autos-atrás	autos-lado	caballos	vacas
Precisión (propio)	79.8 ± 2.7	**89.4** ± 1.8	86 ± 3.2	78 ± 6.7	85.6 ± 2.2	80.5 ± 6.1	81.1 ± 2.0
Precisión [116]	**86.1** ± 1.5	81.2 ± 4.3	**89.9** ± 2.5	80.3 ± 10.1	**89.8** ± 2.3	81.5 ± 7.3	**84.5** ± 1.2
Recuerdo (propio).	**95.7** ± 1.6	89.1 ± 1.6	**92.1** ± 2.3	84.6 ± 5.4	**92.1** ± 1.8	73.3 ± 5.5	82.3 ± 2.1
Recuerdo [116]	89.3 ± 1.1	**91.2** ± 4.3	84.5 ± 2.1	84.7 ± 6.7	89.2 ± 1.5	**78.6** ± 7.6	**86.3** ± 2.2

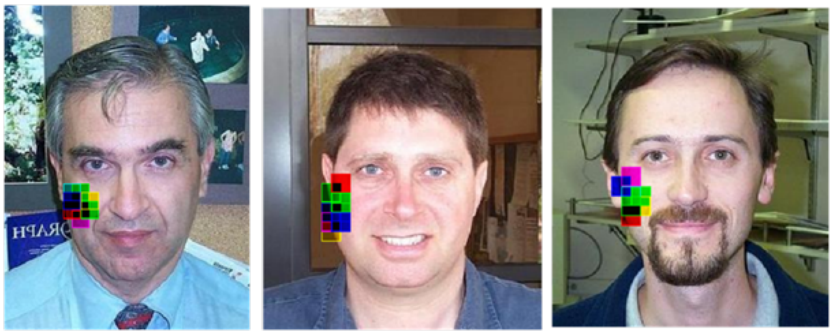


Figura 8.3: Ejemplo de detección de la gramática para el caso de la categoría de rostros. Se detecta la parte más invariante de las imágenes de entrenamiento, en este caso, las mejillas.

objeto en la imagen, se ilustra en la Fig. 8.3.

Discusión. En el caso de la gráfica comparativa de los distintos lexicones visuales, se encontró que un lexicón que ocupe la mayoría de las palabras del alfabeto visual provee un ruido que afecta la precisión del modelo, puesto que aparecerán más falsos positivos. Comparando el caso base (que no utiliza la gramática visual ni las relaciones espaciales) con el modelo propuesto se observa que hay una mejora, lo cual indica que el modelo está ayudando a entregar un conocimiento que sirve para discriminar mejor la categoría visual. La mejora que aporta el lexicón reducido por espacialidad es la reducción de falsos positivos. En cuanto a la comparación con otro trabajo en siete categorías visuales, los resultados indican un comportamiento similar en unas categorías pero no en todas. los ejemplos que contenían ejemplos de objetos más rígidos, tal como rostros, motos o aviones, presentan mejores resultados que el resto. Las categorías con menores resultados son de objetos de estructura un poco más variable, tal como las vacas y los caballos en distintas poses o mayor variación intraclase. Se realizaron pruebas de significancia estadística para cada categoría. La prueba usada fue t de estudiante de varianzas desconocidas diferentes, con un $\alpha = 0.05$ tomando niveles

de significancia de dos colas. Los resultados mostraron que para precisión: la categoría caballos no es estadísticamente significativa. El resto sí lo son, de modo que el modelo propuesto solamente gana en la categoría de motos. Para el recuerdo, la categoría auto-atrás no es estadísticamente significativa, mientras que el resto sí lo son, de modo que el modelo propuesto gana en recuerdo en las categorías de rostros, aviones y autos-lado.

8.3.4. Experimento 4: reconocimiento de objetos en las base de datos ETH, INRIA y poses humanas

Hipótesis. Las reglas *Or* y el aprendizaje de reglas nuevas por sinonima ayuda a incrementar el poder predictivo de la gramática. Un lexicón basado en recuadros ayuda a incrementar el poder predictivo de la gramática.

Objetivo. Evaluar algunos aspectos del comportamiento del aprendizaje de la gramática tales como:

- Aprender a partir de *pocos* ejemplos (menos de 40).
- Evaluar el potencial de la incorporación de reglas *Or* en la gramática.
- Evaluar el potencial del aprendizaje a partir de la generación de nuevas reglas (sección 5.5.2.1).
- Evaluar un lexicón visual compuesto de recuadros a diferentes tamaños.

Condiciones del experimento. La base de datos ETH [65] consta de varias categorías de objetos de estructura rígida en diferentes poses. Un ejemplo aparece en la Fig. 8.4. La base de datos INRIA caballos [31] consta de caballos de diferentes colores aunque de similar estructura. INRIA personas [22] contiene imágenes de peatones en la vía pública, (ejemplos en las Fig. 8.6 y 8.7). También se probó el modelo en una base de datos de prueba conocida como poses de personas [10], que contiene solamente ejemplos positivos, de imágenes de personas en distintas poses (caminando

o sentados). Esta última base de datos tiene etiquetas con regiones del cuerpo, tales como ojos, rodillas, pies, nariz, hombros, entre otros que ayudan a entrenar partes o regiones de la imagen.

Las relaciones espaciales consideradas fueron:

- Topológicas: $IzquierdaDe$, $AdentroDe$, $EncimaDe$, $TraslapadoCon$, $TraslapeIzquierda$, $TraslapeEncima$, $TraslapeEncimaIzq$, $TraslapeDebajoIzq$.
- De orden o dirección: $AlaIzquierdaDe$, $ArribaDe$.

En las relaciones topológicas siempre hay roce o traslape entre regiones, mientras que en las de orden no se tocan las regiones involucradas. El alfabeto visual usado fue el de aprendizaje automático (sección 5.2.3). El lexicón también incluyó reducción de espacialidad (sección 5.4.3) con $\xi = 0.1 \times |\mathscr{I}|$. La evaluación del potencial de las reglas *Or* se realizó en las bases de datos de INRIA caballos y personas.

Resultados. Se describen los resultados encontrados para las pruebas en cada uno de los tres casos.

Base de Datos ETH. El método de evaluación fue una replicación de cinco segmentos de los datos, cambiando los ejemplos de entrenamiento. La tabla 8.4 ilustra los resultados obtenidos. Los ejemplos negativos para cada categoría son el resto de categorías. El entrenamiento se realiza de forma binaria: se aprende una categoría cada vez. Los resultados muestran que si bien se aprende la estructura más invariante que se corresponde con el objeto, en general no se supera el estado del arte. La exactitud promedio para el modelo propuesto (81.5 %) está aún lejos de los trabajos más recientes (94.5 % de [70]). Los trabajos comparados utilizan contornos como características base [65], características de tipo estadístico [70], grafos de segmentos [80]. En esta base de datos se observó que los contornos muestran una mayor variabilidad interclase, de modo que esto ayuda más al método con que se comparó. El método propio utilizó un alfabeto visual *compartido* para todas las categorías basado en recuadros, que si bien tiene un grado de invarianza para separar las categorías, resultó más limitado para esta tarea. En contraparte, los contornos del trabajo comparado

Tabla 8.4: Resultados en base de datos ETH. En negritas se muestran los valores de exactitud más altos, para cada categoría visual. Leibe replicó los experimentos dejando un objeto fuera cada vez, Stasiak y Morales usaron clasificadores y ensamble de clasificadores. La prueba se realizó con 310 ejemplos, cada categoría requirió de dos a tres minutos para la inferencia.

Clase	**Exactitud**	**Precisión**	**Recuerdo**	**Leibe** [65]	Stasiak [112]	Morales [80]
manzana	83.0±1.7	82.59±1.7	83.21±1.7	77.0	89.75	**98.4**
carro	83.50±2.0	80.57±1.8	86.97±2.1	**90.7**	83.9	89.0
vaca	71.41±3.0	67.44±2.9	86.06±2.1	70.7	63.41	**87.1**
taza	91.08±1.6	89.16±1.4	92.08±1.5	86.3	86.34	**99.4**
perro	73.02±4.2	75.24±4.1	67.22±4.3	**81.9**	47.56	69.4
caballo	76.91±4.1	72.02 ±3.6	91.90±4.0	**84.6**	41.95	70.0
pera	85.96±3.6	86.33±3.5	83.93±4.1	99.7	**100**	91.0
tomate	87.68 ±1.8	92.06 ±1.7	82.18±1.7	99.5	92.19	**100**

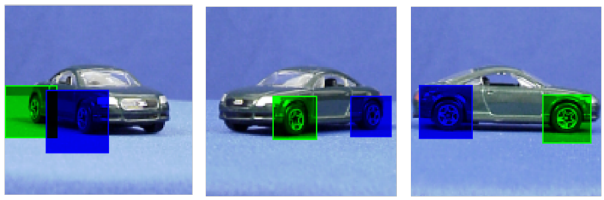


Figura 8.4: Ejemplos de detección de carros con la gramática. Se observa que la gramática no es profunda, puesto que sólo consta de una regla: la detección de un par de llantas.

permitieron retirar el ruido de fondo y trabajar con información más precisa. Por su parte, el lexicón utilizado considera recuadros que incluyen al objeto y al fondo.

Para evaluar el aprendizaje de gramáticas, considerando el caso de pocos ejemplos, se utilizó la medida F sobre un aprendizaje de gramáticas realizado de manera incremental. Los resultados arrojaron que la selección del lexicón puede impactar significativamente en los resultados. La Fig. 8.5 muestra como en general los resultados son similares, desde pocos ejemplos con la excepción del caso para 20 y 60 ejemplos. En estos casos, se observó un lexicón mal seleccionado que al estar compartido, perjudicó a algunas clases sobre otras. Una manera de estabilizar esto es incrementar el número de ejemplos de entrenamiento. La figura muestra resultados un poco más estables cuando se entrena a partir de 100 ejemplos.

Base de Datos INRIA caballos y personas. Para estas bases de datos se utilizó el alfabeto visual basado en recuadros a tamaños variables. El caso para recuadros pequeños fue que generaba

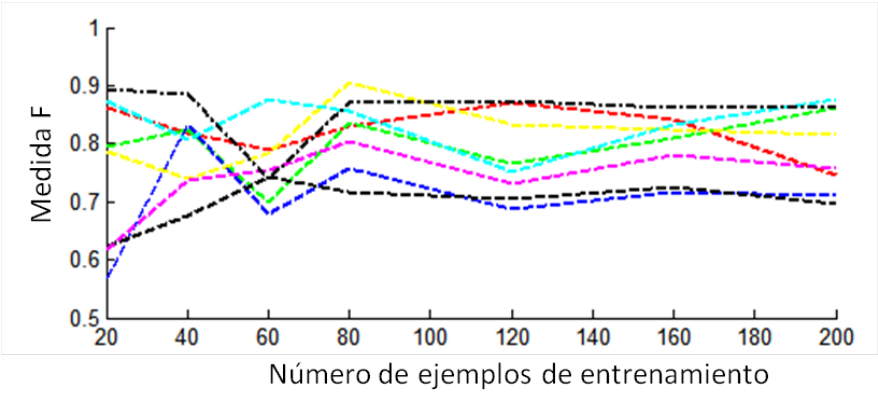


Figura 8.5: Evaluación de la medida F incrementando el número de ejemplos de entrenamiento en la BD de ETH. Cada línea representa los resultados de la medida F de cada una de las ocho categorías visuales de la base de datos.

muchas reglas y mucha composición pero teniendo exactitudes menores a 0.7 para estos casos. En otras palabras, a mayor granularidad mayor composición, pero más ruido y falsos positivos, mientras que a menor granularidad el nivel de composición es menor, aunque hay un mejor compromiso entre la precisión y el recuerdo. Las reglas son menos profundas, aunque son más disyuntivas, es decir se consideran muchas variantes de pose para reconocer a los caballos y a los peatones. Los resultados mostrados en las Tablas 8.5 y 8.6, corresponden al caso de recuadros grandes. Un ejemplo de las reglas encontradas para el caso de peatones es el siguiente:

1. $Spersonas \rightarrow \langle \{TermC_{56}, TermC_{56}^2\}, \{traslapadoCon(TermC_{56}, TermC_{56}^2)\} \rangle$;

2. $Spersonas \rightarrow \langle \{TermC_{28}, TermC_{28}^2\}, \{traslapadoCon(TermC_{28}, TermC_{28}^2)\} \rangle$;

3. $Spersonas \rightarrow \langle \{TermC_{126}, TermC_{164}\}, \{traslapadoCon(TermC_{126}, TermC_{164})\} \rangle$;

4. $Spersonas \rightarrow \langle \{TermC_{174}, TermC_{164}\}, \{traslapadoCon(TermC_{174}, TermC_{164})\} \rangle$;

Es de notarse que incluso se aprenden reglas con una misma palabra visual, tal como las reglas uno y dos del ejemplo anterior, haciendo que incluso la redundancia ayude a dar poder discriminativo y descubrir parte del objeto en la imagen.

Para la base de datos de caballos la comparación se hizo con el número de falsos positivos por imagen, pues éste es el método usado por [32]. El método propuesto halla buena parte de los caballos, aunque no supera los resultados reportados por el método de Ferrari. En el caso de la base de datos de INRIA personas (Tabla 8.6), la medida es con precisión y recuerdo. En esta

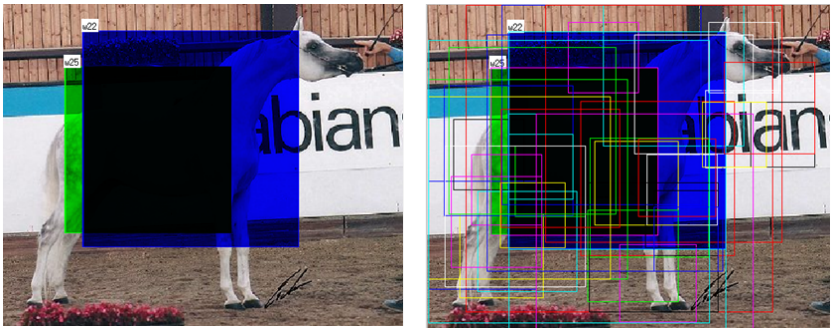


Figura 8.6: Izquierda: regla encontrada que detecta partes del caballo. Los dos recuadros se traslapan casi en su totalidad, por tanto la regla es del tipo $traslapadoCon(w25, w22)$ Derecha: palabras visuales encontradas en la imagen.

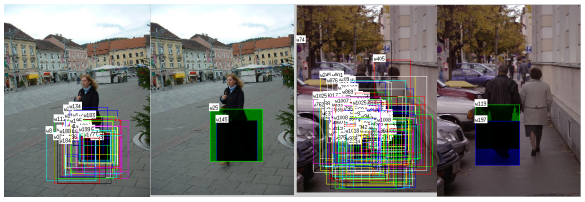


Figura 8.7: Ejemplos de detección de palabras visuales y reconocimiento con la gramática. Automáticamente aprende una determinada parte (en este caso, piernas de peatones).

tabla se incluyen resultados con el método de aprendizaje de nuevas reglas por sinonimia visual. Esta técnica ayudó a mejorar el recuerdo, aunque perdiendo cierto compromiso en precisión. Esta pérdida puede deberse a que el lexicón al ser muy grande promovió la generación de muchas reglas nuevas. Ejemplos de recuadros encontrados para esta base de datos se ilustran en la Fig. 8.7.

Base de datos de poses de personas. Este experimento se realizó con una base de datos que no tiene ejemplos negativos, de modo que solamente se evaluó el recuerdo en la misma. El objetivo es observar el nivel de recuerdo que puede alcanzar la gramática. el máximo recuerdo obtenido fue 94 %. Esta prueba mostró que el aprendizaje de reglas nuevas (sección 5.5.2.1) es útil para generar conocimiento nuevo.

Tabla 8.5: Resultados en bases de datos INRIA caballos. En negritas exactitud más alta, fijando los falsos positivos por imagen.

Clase	**Exactitud**	**FP por imagen**	**Ejemp(+)**	Ejemp(-)	**Ferrari** [32]
Caballos	77.9±0.9	0.20	120	170	**80.0**
Caballos	84±0.8	0.40	120	170	**87.1**

Tabla 8.6: Resultados de bases de datos INRIA personas. En negritas mejores tasas de precisión/recuerdo. Para comparación, se muestran los resultados reportados por Dalal con esta base de datos.

Clase	**Precisión**	**Recuerdo**	**Ejemp(+)**	Ejemp(-)	**Prec/Rec Dalal** [22]
Personas	68.5± 2.6	**84**±3.1	288	453	[**84**/78]
Personas (reglas nuevas)	60.8±1.3	92.2±0.8	288	453	N/A

Figura 8.8: Ejemplos de etiquetas (puntos verdes) basadas en puntos clave para la BD de poses de personas.

Tabla 8.7: Resultados de recuerdo para la base de datos de poses. La segunda columna describe el recuerdo alcanzado por el método desarrollado en la tesis. En la tercera columna se añadió la variante del método que aprende reglas nuevas descrito en la sección 5.5.2.1.

Reglas	Recuerdo	Recuerdo (aprendizaje de reglas)
20 %	0.24	0.24
40 %	0.41	0.46
60 %	0.77	0.81
80 %	0.89	0.93
100 %	0.91	0.94

Discusión. Los resultados mostraron que la gramática visual alcanza a capturar información composicional del objeto visual. También se observó que la composición aprende parte del objeto en lugar del objeto completo. La clasificación se realiza en muchas ocasiones con solamente una o dos reglas encontradas en la gramática, tal como en las Figs. 8.6 y 8.7. En otras palabras, rara vez se detecta el objeto completo acorde a la gramática aprendida. Aún así, aprender estas partes de la categoría del objeto, ayuda a reconocer el mismo, incluyendo los casos de oclusión, puesto que *pocas* reglas implican que puede perderse parte del objeto sin perder recuerdo. Esta estrategia aprendida automáticamente por las gramáticas también ayudó a tratar mejor con el problema de variantes de pose, ya que se aprende la regla o el par de reglas más invariante a estos cambios. Como ejemplo, en el caso de los vehículos se aprende solamente las llantas, evitando aprender el resto, puesto que son los elementos que más sufren a cambios de pose. En el caso de los peatones en vía pública se aprenden el par de piernas, usando relaciones *arribaDe* con *Traslapado,* ya que hay menor invarianza al querer aprender la parte superior de las personas: ocurre que la parte superior son distintas camisas o blusas, mientras que para la parte inferior la mayoría de las personas usa pantalón y además, de color oscuro. Aprender gramáticas con reglas *Or* también fue útil dado que, para el caso de la base de datos de poses de personas, se aprendieron varias configuraciones de reglas para poses de varias partes del cuerpo, por ejemplo: brazo cruzado, piernas del sujeto sentadas o paradas, rostro con hombros, entre otras más. El aprendizaje de reglas nunca vistas ayudó a mejorar resultados en la bases de datos de poses de personas [10]. En casos de bases de datos con pocas variantes de pose, tal como *INRIA personas* [22]. el cambio no es significativo. Se encontró que la principal debilidad del algoritmo se encuentra en los detectores usados (el alfabeto visual). Los resultados, si se ven desde el lado de los modelos de reconocimiento de objetos, no superan en general a otros trabajos relacionados. Por contraste, si se ven desde el lado de representación del conocimiento, hay una capacidad de recuperación de la información visual de manera simplificada y estructurada y que aprende automáticamente objetos, teniendo poca información de ayuda, tal como solamente las imágenes de ejemplos positivos, sin ningún procesamiento adicional.

8.4. Reconocimiento con modelos relacionales

En esta sección se presentan los resultados obtenidos usando el método descrito en el capítulo de Modelos relacionales probabilistas (capítulo 6). En estas pruebas se presenta un análisis comparativo del potencial que tienen los modelos relacionales probabilistas, en áreas de reconocimiento de objetos, las cuales hasta el momento son poco explotadas por los mismos. Las pruebas permiten examinar ciertas ventajas y desventajas que tienen dos de este tipo de modelos y se contratan sus resultados con el modelo general de visión, presentado en el capítulo 5.

Hipótesis. La inferencia con modelos relacionales a partir de una gramática visual para descubrir un objeto en una imagen es una tarea factible.

Objetivo. Evaluar el modelo propuesto de inferencia con redes bayesianas con dos modelos relacionales probabilistas: un modelo con redes bayesianas relacionales y otro con redes lógicas de Markov. Ambos modelos se detallaron en el capítulo 6.

Condiciones del experimento. Se compararon los tres modelos propuestos en diferentes aspectos. Debido a que todos los modelos trabajan con la misma gramática visual como entrada, se deben tener iguales resultados en tasas de reconocimiento, si no es así, habría una pérdida de información en alguno de ellos (y ello implicaría que la traslación de la gramática al modelo que incorpora incertidumbre presenta una falla).

Los datos que se usaron fueron rostros obtenidos de la base de datos Caltech 101 [29]. Para el caso de los ojos se utilizaron imágenes de rostros en mayor resolución. Los ejemplos negativos fueron obtenidos de la categoría de imágenes de fondo de Caltech. Cada imagen de prueba contenía desde 40 hasta 320 regiones como máximo. Este dato es de utilidad para observar el espacio sobre el cual están trabajando los modelos relacionales. Los experimentos se hicieron con 20 imágenes en el caso la BD de ojos y con 50 ejemplos para la gramática de rostros. La estructura de la gramática de rostros fue tomada de [76]. En el caso de la estructura de la gramática del ojo, se siguió la estructura mostrada en la sección 5.3.2 de esta tesis. Los parámetros fueron aprendidos a partir de

Tabla 8.8: Comparación de los modelos relacionales para la gramática de un ojo.

Modelo	Precisión	Recuerdo	Tamaño red	RB creadas	Tiempo	Memoria (MB)
Ojo-RB	76	74	28 (fijo)	44 ± 9.4	<2s	<300
Ojo-RBR	76	74	[1-12]	78 ± 6.3	<1s	<60
Ojo-RLM	76	74	[1-9]	80 ± 10.1	<1s	<60

los datos de entrenamiento siguiendo el método de los capítulos 5 y 6. El modelo propio con RBs se desarrolló en Matlab, corriendo sobre dicha plataforma. En el caso de los modelos relacionales probabilistas se corrieron sobre una máquina virtual de java.

Se evaluaron los siguientes aspectos:

1. Precisión y recuerdo: son usados para verificar que los modelos son equivalentes.

2. Tamaño de la red: cuántos nodos tiene la red bayesiana compilada en promedio por cada imagen.

3. RB creadas: cuántas redes bayesianas se compilan por cada ejemplo. Cuando hay muchas RB creadas se sugiere que el objeto se encuentra o bien partes del mismo han sido detectadas en la imagen.

4. Tiempo y memoria: tiempo de procesamiento (en segundos) y espacio en memoria (Megabytes) cuando se procesa un ejemplo. No se realizó ninguna estrategia de paralelización. Se procesa solamente un ejemplo cada vez.

Resultados. Las pruebas mostraron que los tres modelos son equivalentes en cuanto a los resultados de la inferencia, de modo que más bien en estas pruebas hay más interés en las diferencias con respecto a la transparencia, eficiencia y facilidad de entendimiento de los modelos. Como los modelos fueron pequeños en cuanto al número de nodos de las redes, el costo en tiempo y espacio para la inferencia fue de pocos segundos en general en todos los casos. Las RB creadas tienen una desviación estándar debido a que se pueden crear más redes o menos redes según el ejemplo de prueba que se esté evaluando. Los resultados se resumen en las tablas 8.8 y 8.9.

Table 8.9: Comparación de los modelos relacionales para una gramática basada en rostros.

Modelo	Precisión	Recuerdo	Tamaño Rred	RB creadas	Tiempo	Memoria (MB)
Rostros-RB	86	79	14 (fijo)	314 ± 44	<4s	<400
Rostros-RBR	86	79	[1-9]	691 ± 49	<2s	<150
Rostros-RLM	86	79	[1-6]	753 ± 59	<2s	<150

A continuación se estudia cada caso a manera de comparar las diferencias entre cada uno de los tres modelos.

Red bayesiana. Ejemplo de las redes bayesianas que han sido compiladas a partir del modelo propio se muestra en la Fig. 8.9. La red bayesiana mostrada es fija en tamaño tanto para ejemplos positivos como negativos. La red Ojo-RB tuvo más nodos (tamaño red) con respecto que la red Rostros-RB debido a que la gramática escrita fue más amplia. Esta variación de tamaño siempre es directamente proporcional en los ejemplos instanciados con RBs. las RBs creadas son en general menores a 60 para el caso del Ojo y menores a 400 en el caso de Rostros. Como se ha indicado, el número de redes depende directamente de si en los ejemplos hay regiones que puedan evaluarse en la inferencia. De todas las redes siempre se escoge aquella con el valor de probabilidad más alto al consultar el nodo inicial, tal como se mencionó en la sección 5.7, Ec. 5.18. Los tiempos fueron de segundos para estos ejemplos. En memoria, los modelos construidos en redes bayesianas ocupan un espacio mayor que los relacionales. Aunque en apariencia esto indica un mayor costo espacial, debe tenerse en cuenta que los modelos de red bayesiana incluyen la máquina de Matlab. Al restar el tamaño de ésta, puede decirse que el costo espacial de instanciación es similar entre los tres modelos (entre 60 y 90 megabytes).

Redes bayesianas relacionales. La estructura aleatorio relacional asociada a las RBR siguiendo el método propuesto en la sección 6.3.1 presentó la estructura mostrada en el ejemplo de la Fig. 6.1a y b. La transformación generó el modelo gráfico mostrado en la Fig. 8.10. Una primera observación es que este enfoque elimina los nodos V_T y V_N de la estructura de la red. Esto permite compactar la representación y reducirla a las relaciones espaciales haciendo que éstas tengan mayor importancia para hacer el reconocimiento de la gramática en los ejemplos. Se concluye por tanto que la aporta-

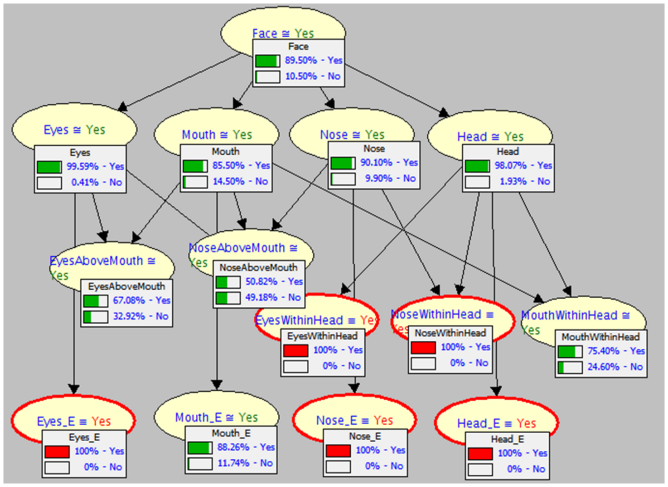


Figura 8.9: Red bayesiana compilada de la gramática de rostros. La estructura fue construida considerando V_N, V_T y V_R de la gramática G como nodos de la RB. La evidencia es pasada sobre los nodos hoja V_T y V_R. Estos nodos hoja son regiones detectadas (nariz o boca) o bien relaciones espaciales (ojos *arriba de* boca). La evidencia se propaga de abajo hacia arriba. La respuesta a la consulta ¿hay un objeto en la imagen? Siempre se contesta en el nodo raíz.

ción de los nodos V_T y V_N en la RBR no es necesaria al no aparecer en la estructura de la red. En otras palabras, las relaciones espaciales son suficientes para la inferencia. Los resultados muestran una estructura que añade nodos auxiliares para poder hacer esta inferencia. Estas estructuras auxiliares sirven como compuertas lógicas que dejan pasar evidencia entre las relaciones espaciales. La estructura luce invertida, de modo que el paso de inferencia aquí es de arriba hacia abajo, en lugar de abajo hacia arriba. La implicación directa que tiene esto es que las TPC de cada nodo tienen más estados. El tamaño de la red es variable y depende del ejemplo que se esté considerando en la inferencia. En la imagen de la Fig. 8.10, se ilustra un caso de un ejemplo positivo. Para ejemplos negativos si *ninguna* relación espacial se cumple solo se crea el nodo del elemento inicial (en este caso, *IsFace*, nodo hoja). En los ejemplos positivos los nodos pueden ir creciendo hasta llegar al máximo de nueve para rostros-RBR y de doce para ojo-RBR, que ejemplifica un verdadero positivo (o un falso positivo, si encuentra todas las relaciones espaciales en él). Esta dinámica en la inferencia permite solamente considerar, de manera similar que en el modelo de inferencia con RBs, la red bayesiana (o redes, si aparece mas de una vez el objeto en la imagen) instanciada de *mayor* tamaño, que equivale a la consulta de la mayor probabilidad que se realiza en RBs. En cuanto al tamaño de las redes, éstas ocupan un tamaño mayor que en RBs. Como los modelos relacionales

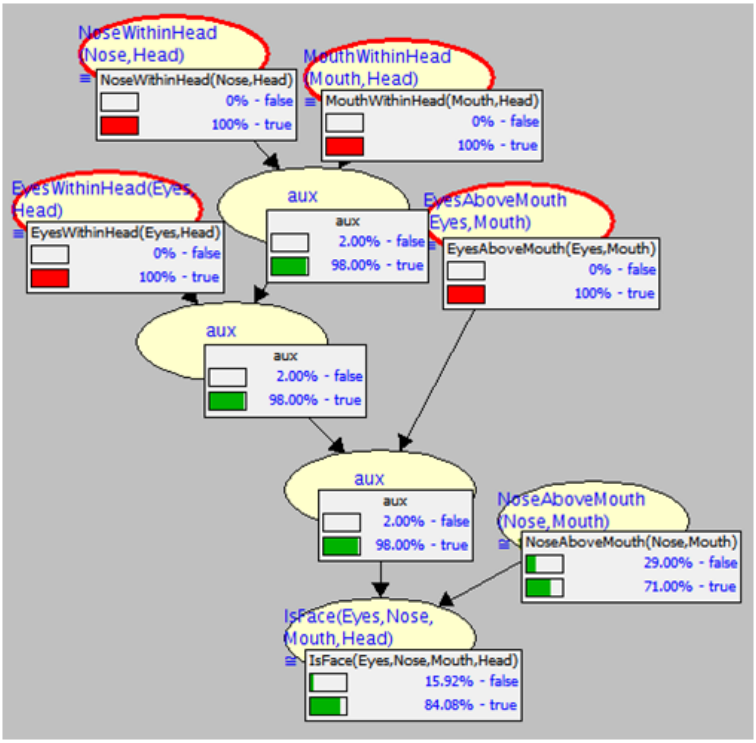


Figura 8.10: Red Bayesiana obtenida de la gramática usando RBR. Se observa una inversión de la estructura de la gramática con respecto de la RB obtenida directamente de la gramática (caso anterior), de modo que la propagación de evidencia en este caso es de arriba hacia abajo. En esta estructura se compacta el número de nodos, solamente los nodos V_R son mostrados. No obstante, se requieren nodos auxiliares que conectan las relaciones espaciales.

probabilistas tratan de evaluar todas las variables posibles como argumentos en cada predicado, hay una mayor cantidad de redes generadas. Sin embargo, la mayoría de estas redes generadas son de un solo nodo, de modo que solamente significa que están tratando de instanciarse en regiones de la imagen que *no tienen* el ejemplo a reconocer.

Redes lógicas de Markov. Siguiendo el método descrito en la sección 6.3.2 para escribir RLMs a partir de una gramática visual, las redes lógicas de Markov para el ejemplo de rostros tiene la siguiente estructura:

```
1.58 faceENMH(E,N,M,H) \/ aboveEM(E,M)

1.67 faceENMH(E,N,M,H) \/ aboveNM(N,M)

1.16 faceENMH(E,N,M,H) \/ withinEH(E,H)

1.25 faceENMH(E,N,M,H) \/ withinNH(N,H)
```

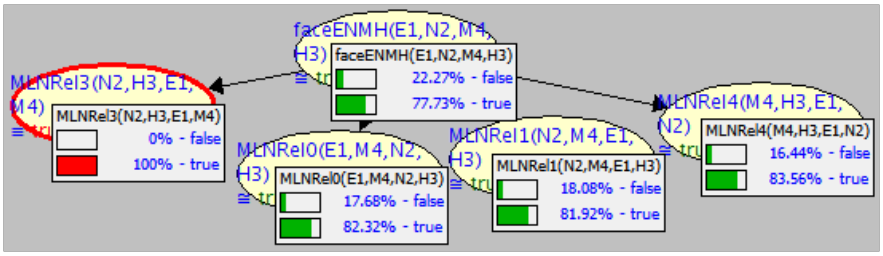


Figura 8.11: Red bayesiana obtenida de la gramática de rostros usando RLM. De los tres casos, siempre genera estructuras más compactas y corresponde en estructura a un clasificador bayesiano simple (o naïve bayes). Su principal ventaja es que tiene la representación más compacta y es relativamente sencillo pasar de la representación en la gramática hacia el modelo de red lógica de Markov. Propagación de evidencia es de abajo hacia arriba y sólo presenta los nodos de las relaciones espaciales V_R.

```
1.34 faceENMH(E,N,M,H) \/ withinMH(M,H)
```

y su transformación resultó en el modelo ilustrado como en la Fig. 8.11. En la construcción de la estructura de la red solo se crean nodos de las relaciones espaciales involucradas generando una estructura del tipo de naïve bayes. Esta representación al igual que en RBRs, compactó los nodos no terminales y terminales de la gramática. Solamente preservó los nodos asociados a las relaciones espaciales y el nodo inicial. Los nombres cambian de modo que se crea un nodo por cada enunciado que involucre los pares de predicados. Al igual que en las RBR, la instanciación produce desde un nodo (una región de la imagen donde no ha detectado el objeto este modelo) hasta seis en rostros-RLM (y nueve en ojo-RLM), que es en donde se cumplen todas las relaciones espaciales. El ejemplo de la Fig. 8.11 es un caso con cinco nodos. Este método muestra que, con respecto de los modelos de la red bayesiana y las RBR (Figs. 8.9 y 8.10), es posible compactar aún más la información provista por la gramática. No necesitar nodos auxiliares hace una representación más pequeña, tanto en estructura como en parámetros, puesto que las TPC de cada nodo solamente tienen un nodo como padre. No obstante, de los tres modelos, las RLM son las que generan en promedio, un mayor número de redes creadas cada vez que se evalúa un ejemplo de prueba. Esto se debe en buena medida a que en su definición, las RLM tienen menos instrumentos para restringir y podar la búsqueda de regiones candidatas que satisfacen la red. Aún así las RLM generan una similar cantidad de ejemplos que su contraparte con RBRs. En tiempo y en memoria se comportan de manera similar que las RBR.

Discusión de los modelos relacionales. Hay varios aspectos a considerar en este estudio comparativo. El aspecto más interesante que se encontró, es que gracias a la representación obtenida con las RLM, en general, es posible obtener una estructura más compacta de la red bayesiana (un naïve bayes) que con respecto del modelo propio con redes bayesianas (Fig. 8.9) y el modelo instanciado con redes bayesinas relacionales (Fig. 8.10). Otro aspecto observado es que queda pendiente la labor de "podar" las instancias que generan los modelos relacionales probabilistas, que principalmente es su punto más débil. Esta debilidad radica en que con facilidad puede crecer el número de instancias generadas haciendo poco viable la inferencia para modelos más grandes. En el modelo propuesto con RBs, debido a que se tiene conocimiento previo sobre los argumentos válidos para cada predicado, la inferencia ocurre con un número de instancias notablemente menor (aproximadamente la mitad para cada caso). Finalmente los modelos ocuparon espacios similares en memoria, aunque no debe perderse de vista que aún son ejemplos pequeños a medianos. Finalmente, en comparativa, se puede decir que la manipulación de una imagen con hasta 320 regiones teniendo diversas relaciones espaciales entre ellas aunque parece pequeño no es despreciable, si se ve desde el punto de vista que algunos trabajos [57], utilizan redes de tipo social en RBRs con alrededor de 40 nodos.

8.5. Reconocimiento de actividades de vestimenta usando Gramáticas Temporales-Relacionales

En esta sección se presentan los resultados del reconocimiento de actividades de vestimenta correcta utilizando las gramáticas que incluyen relaciones espaciales y temporales, vistas en el capítulo 7. El modelo parte de reconocer estas actividades utilizando la gramática TR sobre secuencias de imágenes. Los resultados son prometedores dado que el único conocimiento obtenido fue por parte del diccionario visual en adición a un procesamiento de eliminación de ruido de fondo. Los resultados se mostraron competitivos con un trabajo previo que sí utiliza información adicional, que son antenas RFID que fueron pegadas en las prendas de vestir de manera previa a la toma

de las imágenes.

A diferencia de los experimentos anteriores, en estas pruebas no se utilizó manejo de incertidumbre, en su lugar se utilizó un método de inferencia basado en reglas.

Hipótesis. Una gramática Temporal-Relacional previamente aprendida es capaz de reconocer secuencias visuales.

Objetivo. Mostrar la capacidad de representación y reconocimiento de secuencias visuales correctas e incorrectas por parte de una gramática Temporal-Relacional.

Condiciones del experimento. Se utilizó la misma base de datos que la utilizada en [74], sólo que la información sobre RFID no fue utilizada. Los autores de esta base de datos, buscan detectar las secuencias de vestimenta correcta e incorrecta con la intención de evaluar a pacientes con problemas de discapacidad física y/o mental. Para los propósitos del experimento prestado en esta sección, se evaluó la exactitud, la precisión y el recuerdo a fin de estudiar las capacidades de reconocimiento en secuencias de ejemplo. Esta base de datos contiene ejemplos de 11 personas que cometen 3 tipos de errores comunes al vestirse:

1. Error de orden: el sujeto se coloca al final una ropa que debió ponerse primero.
2. Error espacial: el sujeto no se coloca apropiadamente una prenda.
3. Error de uso de la prenda: el sujeto se coloca la prenda con la textura interior hacia el exterior (prenda al revés).

En total son 47 secuencias de imágenes. La BD tiene 25 ejemplos correctos y 22 tienen algún tipo de error: 10 fallas temporales, 5 fallas espaciales y 7 fallas relacionales. Los ejemplos se componen de sucesiones de imágenes en formato JPEG en alta resolución (2848 x 1602 píxeles). Estas pruebas parten de la información provista en las imágenes, únicamente. La gramática fue descrita manualmente y la inferencia, al estar basada en reglas, es menor a 100 milisegundos, para cada caso.

Figura 8.12: Se ilustran tres tipos de errores en la actividad diaria de vestirse: (a) vestimenta correcta b) falla temporal: orden de prendas es incorrecto, c) falla espacial: la prenda esta parcialmente puesta, d) falla relacional: ropa puesta al revés.

Dado que la tarea busca reconocer errores al vestirse, resulta natural buscar una manera de detectar las prendas de vestir, analizar si el usuario las viste adecuadamente, tanto en orden temporal como espacial, e incluso orden relacional (ponerse una playera al revés). En este sentido, los alfabetos visuales que se han propuesto en esta tesis tienen una intención más genérica, por ello para este caso se ha decidido construir un alfabeto visual que detecte prendas de vestir específicas, que puedan ser procesadas por la gramática visual a un nivel de abstracción más alto. Ejemplos de errores en el proceso de vestirse se ilustran en la Fig. 8.12.

Entrenamiento. Esta etapa consiste en reconocer prendas específicas, para después aprender secuencias de actividades de vestimenta. Para aprender las prendas de vestir, los puntos de interés local o bordes tienen menor poder discriminativo puesto que las prendas suelen ser parecidas en forma pero distintas en textura. De esta forma, se decidió utilizar información de textura, en particular histogramas de color, histogramas de gradiente [22] y patrones locales binarios [86]. El método seguido fue extraer recuadros (de manera parecida al método basado en recuadros) en conjunción con un método de aprendizaje supervisado, debido a que las prendas de vestir se pueden conocer *a priori*. En total se entrenaron 38 distintos tipos de detectores visuales de prendas. Las prendas se consideran diferentes a pesar de que visualmente sean parecidas (por ejemplo, dos pantalones azules de mezclilla). Para entrenar estos clasificadores se utilizaron máquinas de vectores de soporte

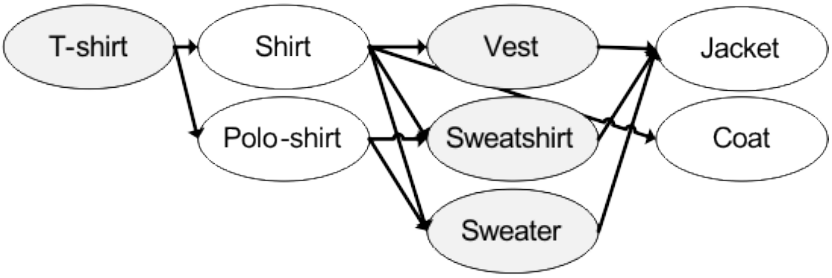


Figura 8.13: Ejemplo de la representación gráfica del ordenamiento correcto de prendas para el modelo propuesto. Después de aprender 38 clases de prendas (incluyendo pantalones de mezclilla, camisas, playeras, suéteres, etc.) se han compactado en el grafo el orden de las prendas de vestir superiores. Un ejemplo para el caso de las prendas inferiores es un grafo de dos nodos, puesto que dado que en la base de datos no hay cambios de tipo de prenda, Las opciones solamente son pantalón de mezclilla o pantalón de vestir.

[20] con el uso de un kernel lineal.

Los clasificadores se entrenan por separado de modo que cada clasificador reconozca sólo un tipo de ropa. El procesamiento realizado incluyó lo siguiente: i) resta de fondo, ii) supresión de no máximos, para remover falsos positivos cuando los clasificadores reportan más de una prenda como candidata en una misma región de la imagen, y iii) fusión de recuadros cuando comparten etiqueta de clasificación. Para definir las relaciones temporales, se utilizó un valor umbral de traslape entre objetos de $\varepsilon = 0.5$.

Para construir la gramática fue necesario representar de alguna forma la secuencia correcta de las prendas de vestir, en la Fig. 8.13 se ilustra con un grafo el orden correcto de prendas de vestir en la parte superior para la base de datos donde se evaluará este modelo. Con esta información se escribió la gramática TR que analiza los ejemplos en la fase de prueba.

Resultados. Los resultados obtenidos por nuestro modelo se pueden resumir en la Tabla 8.10.

Errores temporales. La confusión de errores temporales por errores espaciales se debieron a casos donde la segunda prenda no cubrió completamente la prenda anterior. Por ejemplo, una playera de manga corta puesta después de una sudadera es un error de orden, pero dado que se siguen viendo las mangas de la sudadera la gramática interpreta también un error espacial. En otros casos una persona se puso una playera encima de una abultada chamarra, el resultado es que se

Tabla 8.10: Matriz de confusión de los resultados obtenidos. Ropa puesta al revés fue el caso más difícil debido a que muchas prendas tienen la misma textura y color en el revés.

Tipo de Evento	Vestimenta correcta	Error temporal	Error espacial	Error de lexicón
Vestimenta correcta	80 %	4 %	16 %	0 %
Error temporal	0 %	80 %	20 %	0 %
Error espacial	0 %	40 %	40 %	20 %
Error de lexicón	28.5 %	14.3 %	14.3 %	42.9 %

sigue viendo la primera prenda.

Errores espaciales. En estos casos los errores en ocasiones se pueden interpretar no solamente como espaciales, sino en ocasiones también temporales. A veces la persona intenta vestir una prenda equivocada, pero no alcanza a ponérsela adecuadamente. El error es espacial porque no se la pudo poner. Si se la hubiera puesto, el error hubiera sido temporal.

Errores de lexicón. Dado que muchas prendas tienen la misma textura tanto en el derecho como en el revés (tal como playeras, camisetas e incluso suéteres) estos casos son particularmente difíciles de detectar. Sin embargo, en ocasiones los errores de lexicón se catalogaron como errores espaciales o temporales. El lexicón no descubrió la ropa al revés, pero en ocasiones el lexicón le asignó un tipo de ropa que generaba un error temporal. En otros casos la ropa al revés no pudo ponerse adecuadamente (por ejemplo, los botones no cierran por estar al revés) generando un error espacial.

Vestimenta correcta vs. vestimenta incorrecta. Si se considera que un sistema de vigilancia normalmente asiste a una persona para detectar e informar cuando ocurre una falla, es posible sugerir una versión más simple del modelo: aquella que solamente encuentra errores en los procesos de vestimenta sin explicar el tipo de error. La matriz de confusión en este caso es binaria (Tab. 8.11)

Tabla 8.11: Matriz de confusión considerando solamente dos clases. De 22 secuencias erróneas se descubren 20. La inferencia basada en reglas permite encontrar la mayoría de los errores a partir de la gramática TR.

	vestimenta correcta	vestimenta incorrecta
vestimenta correcta	19	6
vestimenta incorrecta	2	20

Tabla 8.12: Con dos clases, aunque la exactitud se mantiene, la precisión para la detección de fallas se incrementa a 91 %. El modelo tiene un buen compromiso en descubrir cuáles secuencias visuales no son aceptadas por la gramática TR.

Precisión y Recuerdo en fallas	
Precisión	91 %
Recuerdo	77 %
Exactitud	83 %

En este caso se obtiene una precisión del 91 % de detección de fallas, como se muestra en la Tabla 8.12. Lo anterior significa que el modelo detecta nueve fallas de diez de manera correcta.

Discusión y comparación con el trabajo anterior. El trabajo anterior que utiliza esta base de datos [74], utiliza antenas RFID pegadas en las prendas de vestir para hacer el reconocimiento de las secuencias de vestimenta correcta. Si se comparan los resultados con este trabajo previo se tiene un comportamiento significativo para la detección de secuencias de vestir correctas: 80 % vs. 83.9 %. Si se dividen los resultados por los tipos de fallas, las fallas temporales, presentan iguales resultados (80 %); las fallas espaciales y de lexicón tuvieron mejores resultados en el modelo basado en antenas RFID, aunque es de notarse que usando antenas para el caso de ropa al revés es más fácil puesto la detección de las etiquetas RFID izquierda y derecha se invierte. Los errores espaciales en los resultados presentados se confunden un poco con los errores temporales y los errores de prendas al revés, de acuerdo con la matriz de confusión de la Tabla 8.10. En otras palabras, la gramática está detectando de todas formas que hay un error, pero hay confusión en explicar el tipo de error. En resumen, las fallas en el sistema de visión se deben en su mayoría a los detectores de prendas basados en visión, es decir, el alfabeto visual. A pesar de esto, el modelo tiene una precisión que puede considerarse aceptable debido a que es capaz de discriminar bien vestimentas correctas con incorrectas. Algunas sugerencias para tratar de mejorar esto:

- Tratar de mejorar el diccionario visual, ya sea incorporando características de otros tipos (aparte de color y textura).

- Incorporar jerarquía en el lexicón que descubre prendas de vestir para aprender a diferenciar entre prendas muy parecidas. Muchas veces la ropa al revés tiene prácticamente la misma textura en un lado o en otro. También si esta aplicación se extiende a un mayor número de prendas de vestir, las ropas oscuras de distinto tipo podrían confundirse. Esta idea es extrapolable incluso para un lexicón en otros dominios.

- Tratar de detectar otros elementos en las imágenes para descubrir mejor los errores de ropa al revés, aunque esta cuestión se sale del objetivo planteado que es mostrar la capacidad de abstracción y representación del conocimiento que provee la gramática TR.

Algo que debe tenerse en cuenta es que para el caso de ampliar pruebas con gramáticas TR, pudieran considerarse relaciones temporales adicionales, como “ocurre antes/después de otro objeto”, “traslapado en tiempo”, “ocurre durante otro objeto”, entre otras.

Para dominios más amplios, se considera que este trabajo podría combinarse a futuro con modelos relacionales probabilistas, para poder representar de manera más “suave” las reglas espaciales y temporales. También el aprendizaje de la gramática a partir de ejemplos es una tarea aún abierta.

8.6. Discusión de los experimentos realizados

En este capítulo se mostraron los resultados del trabajo de tesis cubriendo tres aspectos distintos. En el primero el modelo construido con redes bayesianas se compara contra algunas bases de datos de imágenes de dominio público. Los resultados mostraron que la gramática recupera partes del objeto, de modo que la gramática abstrajo información visual invariante a lo largo del conjunto de entrenamiento. Al momento de compararse con otros trabajos relacionados, si bien hay avances en la medida de recuerdo, no se logró una mejora en términos de exactitud con respecto del resto. Esta limitación se debió a que alfabeto visual no fue lo suficientemente capaz de capturar de mejor

manera al objeto aprendido. De muchos elementos extraídos con el alfabeto visual se utilizaban menos del 10 % dentro de las gramáticas de acuerdo a los resultados del experimento de la sección 8.3.2. Aunque estos elementos eran los más invariantes, parece que aún se puede hacer más esfuerzo en esta tarea. En el mismo sentido, los trabajos recientes en aprendizaje profundo están empezando a mostrar una mayor capacidad de captura de partes que describen a un objeto. Esto motiva a realizar una mayor investigación del potencial de este tipo de características y de cómo podrían combinarse con las gramáticas visuales. No obstante lo anterior, la expresividad del modelo permite comprender mejor qué se está aprendiendo. El aprendizaje del modelo es particularmente rápido: un modelo se puede construir en menos de uno o dos minutos sin considerar optimización de ningún tipo. Parte de esta rapidez, es que logra capturar la información estructurada a partir de pocos ejemplos, en las pruebas realizadas aproximadamente 30.

En el segundo aspecto, el modelo se comparó contra dos modelos relacionales probabilistas [54, 99] a fin de observar las capacidades de transformar el conocimiento visual en este tipo de modelos, así como comparar si el modelo propio con redes bayesianas también lo hacía de la misma forma. Las pruebas se realizaron en experimentos pequeños, aunque realistas. Los resultados mostraron equivalencia entre los tres modelos, aunque una mayor compacidad de las redes lógicas de Markov, así como una representación paramétrica más amplia en las redes bayesianas relacionales.

En el tercer aspecto, una extensión de las gramáticas visuales (las gramáticas TR, propuestas en esta tesis) se probaron en un ambiente de evaluación de secuencias correctas o incorrectas de vestimenta de ropa. Los resultados indican factibilidad para representar los errores a partir de la gramática. Su comparación con un método que utilizó información adicional (etiquetas RFID pegadas en la ropa) mostró el potencial de las gramáticas para este tipo de aplicación a un costo relativamente bajo (una cámara web) y un uso que permite el manejo de la información con privacidad y/o rapidez: se puede procesar únicamente la gramática, pero no es necesario que las imágenes originales sean evaluadas por un experto. En el siguiente capítulo se muestran las conclusiones que se desprenden de estos experimentos y las contribuciones logradas bajo esta tesis.

Capítulo 9

Conclusiones y Trabajo Futuro

9.1. Resumen del trabajo realizado

El reconocimiento de objetos no es una tarea trivial. Es aún más complicada si se consideran situaciones de oclusión o ruido presente en las imágenes. Los modelos estructurados permiten descomponer el problema de detectar un objeto completo en buscar la detección de elementos más sencillos. De manera análoga, consideran a un objeto como una composición de elementos simples que organizados o relacionados de cierta manera logran describirlo. Al operar de esta manera, los modelos estructurados se comportan de una forma más robusta frente a la oclusión y el ruido. Una de las ventajas que proveen estos modelos es que permiten explicar qué se está aprendiendo ya que tienen un enfoque más visual, al buscar aprender la estructura que representa a un objeto. Asimismo, si se le integran métodos que traten con incertidumbre se convierten en modelos más robustos, pues permiten realizar reconocimiento en condiciones de ruido y oclusión parcial del objeto en las imágenes.

Por otra parte, en fechas recientes se ha abordado el reconocimiento de objetos con modelos que consideran esta tarea como una caja negra que se entrenan con estrategias de aprendizaje profundo. Estos métodos son los que, a pesar de su lentitud en entrenamiento y la necesidad de requerir grandes cantidades de ejemplos en dicha fase, han logrado las mejores tasas de exactitud en diver-

sas bases de datos. El enfoque seguido en esta tesis es sustancialmente diferente. En esta tesis se planteó un modelo que busca aprender a reconocer objetos a partir de una descripción composicional visual del mismo. Esta composición es una abstracción obtenida automáticamente a partir de ejemplos y explicada mediante una gramática de tipo simbólico-relacional. La gramática compone de manera estructurada lo que se puede ver en una imagen. Para poder automatizar el proceso, se creó un algoritmo que aprendió a escribir la gramática a partir de ejemplos. Dado que una gramática por sí sola no puede hacer tareas de inferencia en casos donde no todos los elementos de la misma se hallen en una imagen, la gramática se trasladó a un modelo que incorporó incertidumbre (una red bayesiana) de modo que la inferencia pueda ser realizada considerando casos como oclusión o imprecisión en la detección de los componentes del objeto. Los resultados mostraron que en algunos casos el modelo es competitivo en términos de recuerdo con respecto de otros modelos del trabajo relacionado. Este hallazgo es particularmente importante porque sugiere la capacidad de aprendizaje y abstracción de la gramática hecho de manera automática. En esta misma línea, el aprendizaje de reglas nuevas por sinonimia también mostró mejora de recuerdo en casos particulares. Por contraste, la precisión presentó menores resultados con respecto de otros trabajos que se compararon. Se encontró que esta precisión es más sensible al tipo de diccionario visual usado, mientras que el recuerdo es más sensible al uso de reglas *Or*. Posterior a ello, se decidió comparar este enfoque de transformación a red bayesiana con modelos relacionales probabilistas (dos de ellos), que concluyeron que el enfoque propuesto con redes bayesianas es equivalente, además de mostrar la factibilidad de la traslación de información visual a este tipo de enfoques. A manera de observar el potencial de las gramáticas visuales hacia una extensión en vídeos o secuencias de imágenes, en esta tesis se propusieron las gramáticas temporales-relacionales, que permiten tratar no solamente con relaciones de tipo espacial, sino también de tipo temporal y considerarlas de forma explícita. Los resultados aquí mostraron que fue posible utilizar una sola relación temporal con algunas relaciones espaciales descritas en la gramática para detectar errores en secuencias visuales. El enfoque de inferencia basado en reglas, aunque confundió en algunos casos los errores encontrados en las secuencias visuales, fue útil para discriminar secuencias correctas de las incorrectas.

9.2. Contribuciones

Durante el desarrollo de la tesis surgieron diversas contribuciones que se listan a continuación:

- Un nuevo modelo que representa la información visual mediante gramáticas visuales y además este modelo permite hacer inferencia para detectar el modelo aprendido en otras imágenes de ejemplo. El modelo propuesto es de tipo composicional de modo que construye a partir de elementos simples la estructura del objeto aprendido mediante composición utilizando relaciones espaciales. Esta representación visual con gramáticas se combinó con redes bayesianas para poder hacer inferencia y encontrar este objeto en otras imágenes. La capacidad de encontrar objetos en otras imágenes mostró un poder predictivo de la representación visual obtenida por la gramática.

- Un algoritmo que aprende un alfabeto visual usando recuadros como elementos base en combinación con estrategias de aprendizaje automático. Este algoritmo aprende de manera supervisada a seleccionar recuadros invariantes de una categoría visual. Haciendo una selección de este alfabeto se conformó el lexicón utilizado en la gramática visual.

- Una gramática visual que puede abstraer conocimiento visual. La gramática utilizada fue una gramática simbólico-relacional a la que se le incluyeron restricciones para poder utilizarse en combinación con modelos gráficos probabilistas.

- Una extensión de las gramáticas visuales para que pueda considerar relaciones temporales de forma explícita. Esta extensión se denominó como “gramáticas relacionales-temporales”. Esta extensión permite describir la posición de la secuencia en que están operando objetos que ocurren en una sucesión de imágenes. La gramática propuesta se puede utilizar para el caso donde los objetos ocurren entre dos cuadros sucesivos.

- Un algoritmo que aprende gramáticas visuales de tipo simbólico-relacional a partir de varios ejemplos de objetos visuales. La información solamente es mostrada a partir de predicados

que contienen información espacial acerca de regiones de las imágenes, obtenidas mediante un algoritmo de segmentación basado en color, o basado en recuadros.

- Un método que aprende reglas nuevas que se incorporan a la gramática a partir de una idea de "sinonimia visual". Este método ayudó a incrementar el recuerdo en una base de datos de poses puesto que las reglas incluidas no se habían aprendido inicialmente con el conjunto de entrenamiento.

- Un algoritmo que traslada la información provista en una gramática visual hacia una red bayesiana. De la gramática aprende una estructura, la cual aprende sus parámetros a partir de ejemplos usando un enfoque EM. Esta transformación es transparente y automática a partir de la gramática.

- Una transformación de una gramática visual hacia dos tipos de modelos relacionales probabilistas (redes bayesianas relacionales y redes lógicas de Markov) concluyendo que es factible la representación de conocimiento visual en este tipo de modelos y se compararon estos dos modelos con el trabajo realizado con redes bayesianas. Los resultados son equivalentes en exactitud, pero se observan algunas diferencias en cuanto a tamaño de los modelos en espacio de memoria y en claridad de la representación de los modelos instanciados para la inferencia.

9.3. Conclusiones

Se encontraron algunos hallazgos producto de la investigación desarrollada en esta tesis. Las conclusiones que se obtuvieron derivadas de esta investigación fueron:

- Sí es posible representar información visual a través de una gramática. Prueba de ello es que el modelo es capaz de reconocer el objeto aprendido en otras imágenes. Este aprendizaje se realiza sin considerar ninguna ayuda sobre *dónde* está ubicado el objeto en la imagen. El aprendizaje que realiza la gramática es automático, un costo computacional del orden

de pocos minutos y se obtiene a partir de pocos ejemplos (aproximadamente 30). El tipo de objeto a aprender debe tener una estructura subyacente. Otros enfoques no estructurados de la revisión bibliográfica [63, 52] usan miles de imágenes para entrenarse, con un costo computacional elevado.

- El alfabeto visual tiene la mayor importancia para mejorar las tasas de reconocimiento. La reducción de lexicón ayuda a la reducción de falsos positivos, mientras que ampliar los parámetros tiene un impacto menor. Un alfabeto visual adecuado para el conjunto de datos ayuda notablemente a incrementar los resultados en reconocimiento. Un alfabeto visual limitado (como el mostrado con segmentación y pocas relacionales espaciales) provoca una reducción notable en la capacidad de reconocimiento del modelo a aprender.

- Las relaciones espaciales juegan un papel importante para la realización de reconocimiento, aunque se encontró que incorporar muchas relaciones de tipo espacial hacen más lenta la inferencia. Las relaciones espaciales difusas como *Cerca* y *Le jos*, que pueden aparecer entre cada par de elementos de la imagen, producen una degradación de la inferencia puesto que la búsqueda se tiene que realizar exhaustivamente. Tener un compromiso adecuado en la cantidad de relaciones espaciales a describir permite tener una inferencia más rápida con una reducida afectación a una rica representación del conocimiento visual.

- El aprendizaje de gramáticas, en particular utilizando reglas de tipo *Or*, sí ayuda ampliar la cobertura (recuerdo) en las tareas de reconocimiento. Aprender *variantes* del objeto permite un mejor comportamiento de las tasas de reconocimiento, además de hacerlo de manera automática. Las reglas *Or* permiten descubrir variantes del objeto a lo largo de los ejemplos vistos durante el entrenamiento. Los ejemplos que no pudieron ser cubiertos con este enfoque tienen que ver con el hecho de que no se pudo aprender una configuración del conjunto de entrenamiento para esos ejemplos de prueba, lo cual dio pie al aprendizaje de reglas no vistas.

- El aprendizaje de reglas nunca vistas a través de la sinonimia visual, también ayuda a in-

crementar el recuerdo. La combinación de aprender reglas de tipo *Or* en conjunción con el aprendizaje de reglas nunca vistas mejoró el recuerdo para casos de objetos que presentan varias poses o tienen una estructura más flexible. Para evitar la pérdida de compromiso entre precisión y recuerdo, el podado del lexicón hace que se limite el aprendizaje de reglas nunca vistas y se restringe a utilizar las palabras visuales más cercanas a la categoría visual. En este sentido, las representaciones visuales obtenidas con la gramática que usan reglas nuevas, sí se corresponden con ejemplos positivos de la categoría. En otras palabras, el modelo aprendió a ver ejemplos que no conoció en el entrenamiento.

- Los modelos relacionales probabilistas al ser comparados con el modelo de visión propuesto con la red bayesiana muestran resultados equivalentes. Este resultado sugiere que ningún modelo está perdiendo información obtenida por la gramática y el aprendizaje a partir de ejemplos es equivalente en los tres casos. Las diferencias sugieren que la estructura de la red bayesiana y la red bayesiana relacional pueden compactarse hasta un clasificador bayesiano simple, puesto que ésta es la estructura obtenida por una red lógica de Markov. Una ventaja encontrada es que el traslado de la gramática a los modelos relacionales se da en un lenguaje muy similar: ambos casos usan lógica de predicados para establecer las relaciones entre los elementos terminales y no terminales. En el modelo de la red bayesiana la traslación es siempre hacia nodos y enlaces en la red.

- El enfoque propuesto de incorporar relaciones temporales a una gramática visual permitió representar el conocimiento de una forma compacta, abstracta y simple que facilitó una inferencia que es competitiva en una aplicación de evaluación de secuencias de vestimenta correcta. Se observó que el método propuesto tiene resultados similares a pesar de contar con menos información que el modelo con el cual se comparó.

9.3.1. Prueba de Hipótesis.

La hipótesis planteada se probó a través de los experimentos en esta tesis, de modo que sí fue posible abstraer una categoría visual utilizando gramáticas visuales, puesto que esta abstracción se pudo utilizar para reconocer la misma categoría visual en otras imágenes de prueba, sin otra información más que la definición del alfabeto visual, que es el espacio donde trabajó dicha gramática. La traslación hacia modelos gráficos probabilistas y relacionales probabilistas permitió manejar incertidumbre en la tarea de inferencia. En los tres casos, los resultados fueron equivalentes para un par de pruebas realizadas, pequeñas (en comparación con tareas tradicionales de reconocimiento de objetos) pero realistas (en comparación con trabajos que hacen aplicaciones con modelos relacionales probabilistas). Lo anterior concluye que la inferencia usando redes bayesianas es análoga a la realizada con modelos relacionales probabilistas y que sí reconoce al objeto aprendido. Las limitaciones en cuanto a precisión (la existencia de falsos positivos) están relacionadas con las capacidades y limitaciones del alfabeto visual. A mejor alfabeto visual, mayores capacidades de precisión en reconocimiento de objetos por parte de la gramática.

9.4. Trabajo Futuro

Durante el desarrollo de la investigación desarrollada en el trabajo de tesis, hay varias preguntas que surgieron y que se pueden considerar como trabajo futuro para esta tesis:

- Un estudio más profundo del alfabeto visual. En esta tesis se exploraron algunos tipos de alfabeto pero la conclusión observada fue que el alfabeto debe construirse de manera más cuidadosa con respecto del dominio de imágenes. En algunos casos características locales o de textura pueden funcionar bien, mientras en otros casos algoritmos basados en color o forma podrían ser más apropiados. No hay fórmula única. Aunque se exploró un algoritmo que obtiene recuadros a partir de técnicas de aprendizaje automático, aún hay diversas oportunidades de mejora. Una posibilidad es aprender alfabetos visuales bajo enfoques no supervisados, es decir, que sepan discriminar entre las imágenes pero sin estar atados a las

etiquetas de las categorías de los objetos a reconocer. Este problema se considera aún abierto y bajo investigación en los últimos años.

- Combinar ideas recientes de aprendizaje profundo con gramáticas visuales. Durante el desarrollo de esta tesis surgieron con fuerza enfoques que trabajan reconocimiento de objetos utilizando técnicas de aprendizaje profundo. Estas técnicas tienen el inconveniente de tener poco poder expresivo, aunque sus tasas de reconocimiento de objetos en bases de datos difíciles como Pascal VOC [28], son las más altas dentro de la literatura. Aunque en esta tesis se mostraron las potencialidades de representar el conocimiento con gramáticas visuales (tal como la posibilidad de representar con modelos relacionales la misma) pudiera explorarse a futuro mejorar la tasa de reconocimiento de objetos sin menoscabo del poder expresivo de la gramática. Esto pudiera hacerse combinando estrategias de aprendizaje profundo, quizás utilizando auto codificadores visuales [3, 5], para reducir la dimensionalidad de la información visual en combinación con una adaptación de algún tipo de gramática visual y/o relacional.

- Un estudio más profundo o búsqueda de aplicaciones donde el aprendizaje de reglas nunca vistas tenga mayor potencial. Los resultados sugieren que es útil, aunque aún es necesario explorar la importancia de este conocimiento nuevo a partir de lo encontrado en los ejemplos de entrenamiento.

- Investigar más aplicaciones realistas para los modelos relacionales probabilistas en combinación con visión. La comparación realizada sugiere que estos modelos son capaces de tratar con información visual estructurada, aunque hay limitaciones de eficiencia espacial y temporal en los mismos para tratar problemas más ambiciosos, tal como el procesamiento de grandes volúmenes de datos (*Big Data*). Esta área está en pleno crecimiento debido al incremento reciente de información visual ya no como imágenes, sino también como secuencias (vídeos). Los trabajos revisados en el estado del arte fueron igualmente pequeños, en cuanto a cantidad de datos, que los experimentos presentados.

- Explorar un mejor manejo de las restricciones de los datos en los modelos relacionales. El

objetivo es tratar de reducir el espacio donde operan estos modelos. Los modelos relacionales tienen el inconveniente de explorar todo el universo de posibilidades para las variables, aunque explorar maneras de podar esta búsqueda podría hacer los modelos más escalables. Un trabajo futuro es investigar cómo reducir esto para aplicaciones relacionadas con visión.

- El aprendizaje de las gramáticas de tipo relacional sigue siendo un área abierta. Aunque en esta tesis se propuso un algoritmo que aprende gramáticas automáticamente a partir de pocos ejemplos, el caso más general, (que involucra reglas de reescritura y que trate con producciones cíclicas) aún no es conocido dentro de los trabajos del área.

- El manejo de gramáticas de tipo relacional-temporal es aún poco conocido. Ahondar un poco más en las implicaciones de las relaciones temporales no es una tarea sencilla: para la representación pudieran incluirse todas las relaciones temporales propuestas por Allen a fin tener una gramática RT más flexible. En el caso de su aprendizaje, sólo se conoce su escritura a partir de un ejemplo; y en la inferencia sólo se conoce el caso a partir de reglas. Otro aspecto es incorporar incertidumbre a las gramáticas temporal-relacionales. En resumen, las gramáticas temporales-relacionales son un área bastante amplia para explorar a futuro.

- La creación de repositorios de imágenes donde la estructura de los objetos visuales sea el aspecto más importante para el reconocimiento. La mayoría de los repositorios tienen un enfoque más general, sin enfocarse tanto en si el objeto tiene una estructura visual que se pueda abstraer.

9.5. Publicaciones

Derivadas de esta tesis se desarrollaron las siguientes publicaciones:

1. Elias Ruiz, Augusto Meléndez, Luis Enrique Sucar, *Towards a General Vision System Based on Symbol-Relation Grammars and Bayesian Networks*, en Jürgen Schmidhuber, Kristinn R. Thórisson, Moshe Looks, ed., AGI vol. 6830, (Springer, 2011), pp. 291-296.

2. Elias Ruiz, Luis Enrique Sucar, *Object Recognition Based on Visual Grammars and Bayesian Networks*, en Francesca Rossi, ed., IJCAI 2013, Proceedings of the 23rd International Joint Conference on Artificial Intelligence, Beijing, China, August 3-9, 2013 (IJCAI/AAAI, 2013).

3. Elias Ruiz, Luis Enrique Sucar, *An Object Recognition Model Based on Visual Grammars and Bayesian Networks*, en Reinhard Klette, Mariano Rivera, Shiníchi Satoh, ed., Image and Video Technology - 6th Pacific-Rim Symposium, PSIVT 2013, Guanajuato, México, October 28-November 1, 2013. Proceedings vol. 8333, (Springer, 2013), pp. 349-359.

4. Elias Ruiz, Luis Enrique Sucar, *Recognizing Visual Categories with Symbol-Relational Grammars and Bayesian Networks*, en Eduardo Bayro-Corrochano, Edwin R. Hancock, ed., Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications - 19th Iberoamerican Congress, CIARP 2014, Puerto Vallarta, México, November 2-5, 2014. Proceedings vol. 8827, (Springer, 2014), pp. 540-547.

Bibliografía

[1] Narendra Ahuja and Sinisa Todorovic. Learning the taxonomy and models of categories present in arbitrary images. In *In ICCV*, 2007. [54]

[2] James F. Allen. Maintaining knowledge about temporal intervals. *Commun. ACM*, 26(11):832–843, 1983. [16, 113]

[3] Pierre Baldi. Autoencoders, unsupervised learning, and deep architectures. In Isabelle Guyon, Gideon Dror, Vincent Lemaire, Graham W. Taylor, and Daniel L. Silver, editors, *Unsupervised and Transfer Learning - Workshop held at ICML 2011, Bellevue, Washington, USA, July 2, 2011*, volume 27 of *JMLR Proceedings*, pages 37–50. JMLR.org, 2012. [170]

[4] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (surf). *Comput. Vis. Image Underst.*, 110(3):346–359, June 2008. [14]

[5] Yoshua Bengio. Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, 2(1):1–127, 2009. [170]

[6] B. Berlin. *Basic Color Terms: Their Universality and Evolution*. Berkeley/Los Angeles: University of California Press, 1969. [71, 72]

[7] Elie Bienenstock, Stuart Geman, and Daniel Potter. Compositionality, mdl priors, and object recognition. In *Neural Information Processing Systems*, pages 838–844. MIT Press, 1997. [57]

[8] Liefeng Bo, Kevin Lai, Xiaofeng Ren, and Dieter Fox. Object recognition with hierarchical kernel descriptors. In *The 24th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011, Colorado Springs, CO, USA, 20-25 June 2011*, pages 1729–1736. IEEE Computer Society, 2011. [14]

[9] Sean Borman. The expectation maximization algorithm – a short tutorial. Introduces the Expectation Maximization (EM) algorithm and fleshes out the basic mathematical results, inclusing a proof of convergence. The Generalized EM algorithm is also introduced., January 2009. [28]

[10] Lubomir Bourdev and Jitendra Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *International Conference on Computer Vision (ICCV)*, 2009. [129, 141, 147]

[11] Claus Brabrand, Robert Giegerich, and Anders Møller. Analyzing ambiguity of context-free grammars. *Sci. Comput. Program.*, 75(3):176–191, 2010. [18]

[12] Robert K. Bradley, Lior Pachter, and Ian Holmes. Specific alignment of structured RNA: stochastic grammars and sequence annealing. *Bioinformatics*, 24(23):2677–2683, 2008. [18]

[13] Jinhai Cai and Zhi-Qiang Liu. Integration of structural and statistical information for unconstrained handwritten numeral recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 21(3):263–270, 1999. [15]

[14] Lo-Bin Chang, Ya Jin, Wei Zhang, Eran Borenstein, and Stuart Geman. Context, computation, and optimal roc performance in hierarchical models. *International Journal of Computer Vision*, 93(2):117–140, 2011. [57]

[15] Mark Chavira, Adnan Darwiche, and Manfred Jaeger. Compiling relational bayesian networks for exact inference. *Int. J. Approx. Reasoning*, 42(1-2):4–20, 2006. [40, 109]

[16] N. Chomsky. *Knowledge of language: its nature, origin, and use*. Convergence Series. Praeger, 1986. [50]

[17] N. Chomsky. *Syntactic structures*. A Mouton classic. Mouton de Gruyter, 2002. [18, 50]

[18] C. Chow and C. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Trans. Inf. Theor.*, 14(3):462–467, September 2006. [30]

[19] Christophe Claramunt and Bin Jiang. An integrated representation of spatial and temporal relationships between evolving regions. *Journal of Geographical Systems*, 3(4):411–428, 2001. [17]

[20] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995. [137, 157]

[21] Gennaro Costagliola, Vincenzo Deufemia, Filomena Ferrucci, and Carmine Gravino. Using extended positional grammars to develop visual modeling languages. In *Proceedings of the 14th international conference on Software engineering and knowledge engineering, SEKE 2002, Ischia, Italy, July 15-19, 2002*, pages 201–208. ACM, 2002. [113]

[22] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In Cordelia Schmid, Stefano Soatto, and Carlo Tomasi, editors, *International Conference on Computer Vision & Pattern Recognition*, volume 2, pages 886–893, INRIA Rhône-Alpes, ZIRST-655, av. de lÉurope, Montbonnot-38334, June 2005. [73, 120, 128, 129, 141, 146, 147, 156]

[23] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977. [28, 95]

[24] S. Dickinson. The Evolution of Object Categorization and the Challenge of Image Abstraction. In S. Dickinson, A. Leonardis, B. Schiele, and M. Tarr, editors, *Object Categorization: Computer and Human Vision Perspectives*, pages 1–37. Cambridge University Press, 2009. [48]

[25] Pedro M. Domingos, Stanley Kok, Hoifung Poon, Matthew Richardson, and Parag Singla. Unifying logical and statistical AI. In *Proceedings, The Twenty-First National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference, July 16-20, 2006, Boston, Massachusetts, USA*, pages 2–9. AAAI Press, 2006. [43, 109]

[26] Christine du Toit and Andries van der Walt. Temporal grammars. In *Proceedings of the 2002 Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists on Enablement Through Technology*, SAICSIT '02, pages 205–211, Republic of South Africa, 2002. South African Institute for Computer Scientists and Information Technologists. [113]

[27] Max J. Egenhofer. A formal definition of binary topological relationships. In Witold Litwin and Hans-Jörg Schek, editors, *Foundations of Data Organization and Algorithms, 3rd International Conference, FODO 1989, Paris, France, June 21-23, 1989, Proceedings*, volume 367 of *Lecture Notes in Computer Science*, pages 457–472. Springer, 1989. [15]

[28] Mark Everingham, S. M. Ali Eslami, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, 2015. [170]

[29] L. Fei-Fei, R. Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *Proceedings of the 2004 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'04) Volume 12 - Volume 12*, CVPRW '04, pages 178–186, Washington, DC, USA, 2004. IEEE Computer Society. [6, 47, 51, 128, 148]

[30] Pedro F. Felzenszwalb. Object detection grammars. In *ICCV Workshops*, page 691. IEEE, 2011. [58, 73]

[31] Vittorio Ferrari, Frédéric Jurie, and Cordelia Schmid. Accurate object detection with deformable shape models learnt from images. In *2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2007), 18-23 June 2007, Minneapolis, Minnesota, USA*. IEEE Computer Society, 2007. [129, 141]

[32] Vittorio Ferrari, Frédéric Jurie, and Cordelia Schmid. From images to shape models for object detection. *International Journal of Computer Vision*, 87(3):284–303, 2010. [128, 144, 145]

[33] F. Ferrucci, G. Pacini, G. Satta, M. I. Sessa, G. Tortora, M. Tucci, and G. Vitiello. Symbol-relation grammars: a formalism for graphical languages. *Inf. Comput.*, 131(1):1–46, 1996. [21, 22, 63, 78, 111, 113]

[34] Sanja Fidler, Gregor Berginc, and Ales Leonardis. Hierarchical statistical learning of generic parts of object structure. In *CVPR (1)'06*, pages 182–189, 2006. [4, 51]

[35] Sanja Fidler, Marko Boben, and Ales Leonardis. Similarity-based cross-layered hierarchical representation for object categorization. In *CVPR'08*, pages –1–1, 2008. [51, 53]

[36] Sanja Fidler and Ales Leonardis. Towards scalable representations of object categories: Learning a hierarchy of parts. In *CVPR'07*, pages –1–1, 2007. [4, 51]

[37] Nir Friedman, Lise Getoor, Daphne Koller, and Avi Pfeffer. Learning probabilistic relational models. In *In IJCAI*, pages 1300–1309. Springer-Verlag, 1999. [32, 35, 36, 63]

[38] Kunihiko Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36(4):193–202, 1980. [70]

[39] D. Gabor. Theory of communication. *JIEE*, 93(3):429–459, 1946. [52, 70]

[40] Michael R. Garey and David S. Johnson. *Computers and Intractability; A Guide to the Theory of NP-Completeness*. W. H. Freeman & Co., New York, NY, USA, 1990. [49, 85]

[41] Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 6(6):721–741, 1984. [43]

[42] Stuart Geman, Daniel F. Potter, and Zhiyi Chi. Composition systems. *Quart. Appl. Math.*, 60(4):707–736, 2002. [57]

[43] Michael R. Genesereth and Nils J. Nilsson. *Logical foundations of artificial intelligence*. Morgan Kaufmann, 1988. [33]

[44] Lise Getoor and Ben Taskar, editors. *Introduction to Conditional Random Fields for Relational Learning*. MIT Press, 2006. [32, 36]

[45] Lise Getoor and Ben Taskar. *Introduction to Statistical Relational Learning*. The MIT Press, 2007. [4, 26, 32, 35, 62]

[46] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. *Markov Chain Monte Carlo in Practice (Chapman & Hall/CRC Interdisciplinary Statistics)*. Chapman and Hall/CRC, softcover reprint of the original 1st ed. 1996 edition, dec 1995. [31, 43]

[47] R. Girshick, P. Felzenszwalb, and D. McAllester. Object detection with grammar models. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 2011. [4]

[48] Ross B. Girshick, Pedro F. Felzenszwalb, and David A. McAllester. Object detection with grammar models. In John Shawe-Taylor, Richard S. Zemel, Peter L. Bartlett, Fernando C. N. Pereira, and Kilian Q. Weinberger, editors, *NIPS*, pages 442–450, 2011. [58, 59]

[49] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. Technical Report 7694, California Institute of Technology, 2007. [128, 135]

[50] Shih Hsiu Huang. *Compositional Approach To Recognition Using Multi-Scale Computations*. PhD thesis, Brown University, 2001. [57]

[51] D. H. Hubel and T. N. Wiesel. Receptive fields and functional architecture of monkey striate cortex. *J Physiol*, 195(1):215–243, March 1968. [52, 61]

[52] Brody Huval, Adam Coates, and Andrew Y. Ng. Deep learning for class-generic object detection. *CoRR*, abs/1312.6885, 2013. [3, 14, 167]

[53] IEEE. *2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2007), 18-23 June 2007, Minneapolis, Minnesota, USA*. IEEE Computer Society, 2007. [183, 187]

[54] Manfred Jaeger. Relational bayesian networks. In Dan Geiger and Prakash P. Shenoy, editors, *UAI*, pages 266–273. Morgan Kaufmann, 1997. [32, 35, 36, 38, 62, 64, 98, 161]

[55] Manfred Jaeger. Parameter learning for relational bayesian networks. In *Proceedings of the 24th international conference on Machine learning*, ICML '07, pages 369–376, New York, NY, USA, 2007. ACM. [62]

[56] Dominik Jain, Klaus von Gleissenthall, and Michael Beetz. Bayesian logic networks and the search for samples with backward simulation and abstract constraint learning. In Joscha Bach and Stefan Edelkamp, editors, *KI*, volume 7006 of *Lecture Notes in Computer Science*, pages 144–156. Springer, 2011. [35]

[57] Jiuchuan Jiang and Manfred Jaeger. Community detection for multiplex social networks based on relational bayesian networks. In Troels Andreasen, Henning Christiansen, Juan-Carlos Cubero, and ZbigniewW. Ras, editors, *Foundations of Intelligent Systems*, volume 8502 of *Lecture Notes in Computer Science*, pages 30–39. Springer International Publishing, 2014. [44, 154]

[58] Ya Jin. *Probabilistic Hierarchical Image Models*. PhD thesis, Brown University, 2006. [57]

[59] Ya Jin and Stuart Geman. Context and hierarchy in a probabilistic image model. In *CVPR (2)*, pages 2145–2152. IEEE Computer Society, 2006. [57, 58]

[60] Mayank Juneja, Andrea Vedaldi, C. V. Jawahar, and Andrew Zisserman. Blocks that shout: Distinctive parts for scene classification. In *2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013*, pages 923–930. IEEE, 2013. [75]

[61] Kristian Kersting and Luc De Raedt. Basic principles of learning bayesian logic programs. In Raedt et al. [96], pages 189–221. [35]

[62] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009. [25, 30, 31, 62]

[63] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In Peter L. Bartlett, Fernando C. N. Pereira, Christopher J. C. Burges, Léon Bottou, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States.*, pages 1106–1114, 2012. [167]

[64] Fred Lakin. Visual grammars for visual languages. In Kenneth D. Forbus and Howard E. Shrobe, editors, *Proceedings of the 6th National Conference on Artificial Intelligence. Seattle, WA, July 1987.*, pages 683–688. Morgan Kaufmann, 1987. [113]

[65] Bastian Leibe and Bernt Schiele. Analyzing appearance and contour based methods for object categorization. In *CVPR (2)*, pages 409–415. IEEE Computer Society, 2003. [128, 141, 142, 143]

[66] Rainer Lienhart and Jochen Maydt. An extended set of haar-like features for rapid object detection. In *ICIP (1)*, pages 900–903, 2002. [71]

[67] Liang Lin, Tianfu Wu, Jake Porway, and Zijian Xu. A stochastic graph grammar for compositional object representation and recognition. *Pattern Recognition*, 42(7):1297–1307, 2009. [56]

[68] D. G. Lowe. *Perceptual Organization and Visual Recognition*. Kluwer, Boston, 1984. [7, 56, 61, 72, 73]

[69] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004. [12, 14, 120]

[70] Jiwen Lu, Gang Wang, and Pierre Moulin. Image set classification using holistic multiple order statistics features and localized multi-kernel metric learning. In *IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013*, pages 329–336. IEEE Computer Society, 2013. [142]

[71] Tomasz Malisiewicz, Abhinav Gupta, and Alexei A. Efros. Ensemble of exemplar-svms for object detection and beyond. In Dimitris N. Metaxas, Long Quan, Alberto Sanfeliu, and Luc J. Van Gool, editors, *ICCV*, pages 89–96. IEEE, 2011. [75]

[72] Kim Marriott and Bernd Meyer. Towards a hierarchy of visual languages. In *Proceedings of the 1996 IEEE Symposium on Visual Languages, Boulder, Colorado, USA, September 3-6, 1996*, pages 196–203. IEEE Computer Society, 1996. [113]

[73] Kim Marriott, Bernd Meyer, and Kent B. Wittenburg. Visual language theory. In Kim Marriott and Bernd Meyer, editors, *Visual language theory*, chapter A survey of visual language specification and recognition, pages 5–85. Springer-Verlag New York, Inc., New York, NY, USA, 1998. [18]

[74] A. Matic, P. Mehta, J. M. Rehg, V. Osmani, and O. Mayora. Monitoring dressing activity failures through rfid and video. *Journal of Methods of Information in Medicine*, 51:45–54, Jan 2012. [123, 128, 129, 155, 159]

[75] Augusto Meléndez. Una gramática visual para detección de rostros. Master's thesis, Instituto Nacional de Astrofísica, Óptica y Electrónica, Calle Luis Enrique Erro No. 1, 2011. [4, 59, 101]

[76] Augusto Melendez, Luis Sucar, and Eduardo Morales. A visual grammar for face detection. In Angel Kuri-Morales and Guillermo Simari, editors, *Advances in Artificial Intelligence - IBERAMIA 2010*, volume 6433 of *Lecture Notes in Computer Science*, pages 493–502. Springer Berlin / Heidelberg, 2010. 10.1007/978-3-642-16952-6-50. [59, 60, 148]

[77] Krystian Mikolajczyk and Cordelia Schmid. A performance evaluation of local descriptors. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(10):1615–1630, 2005. [120]

[78] Brian Milch, Bhaskara Marthi, Stuart J. Russell, David Sontag, Daniel L. Ong, and Andrey Kolobov. BLOG: probabilistic models with unknown objects. In Leslie Pack Kaelbling and Alessandro Saffiotti, editors, *IJCAI-05, Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence, Edinburgh, Scotland, UK, July 30-August 5, 2005*, pages 1352–1359. Professional Book Center, 2005. [36]

[79] Eric Mjolsness. Visual grammars and their neural nets. In John E. Moody, Stephen Jose Hanson, and Richard Lippmann, editors, *Advances in Neural Information Processing Systems 4, [NIPS Conference, Denver, Colorado, USA, December 2-5, 1991]*, pages 428–435. Morgan Kaufmann, 1991. [113]

[80] Annette Morales-González and Edel B. García Reyes. Simple object recognition based on spatial relations and visual features represented using irregular pyramids. *Multimedia Tools Appl.*, 63(3):875–897, 2013. [142, 143]

[81] Stephen Muggleton. Inductive logic programming. *New Generation Comput.*, 8(4):295–318, 1991. [43]

[82] Stephen Muggleton. Learning structure and parameters of stochastic logic programs. In Stan Matwin and Claude Sammut, editors, *ILP*, volume 2583 of *Lecture Notes in Computer Science*, pages 198–206. Springer, 2002. [36]

[83] R. E. Neapolitan. *Learning Bayesian Networks*. Prentice Hall, 2003. [27]

[84] Jennifer Neville and David Jensen. Relational dependency networks. *Journal of Machine Learning Research*, 8:653–692, 2007. [35]

[85] Timo Ojala, Matti Pietikäinen, and David Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition*, 29(1):51–59, 1996. [73]

[86] Timo Ojala, Matti Pietikäinen, and David Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition*, 29(1):51–59, 1996. [120, 156]

[87] Bjorn Ommer and Joachim Buhmann. Object categorization by compositional graphical models. In *EMMCVPR*, pages 235–250. Springer, 2005. [52]

[88] Bjorn Ommer and Joachim M. Buhmann. Learning compositional categorization models. In *In ECCV*, pages 316–329. In: ECCV, Springer, 2006. [4, 52]

[89] Bjorn Ommer and Joachim M. Buhmann. Learning the compositional nature of visual objects. In *CVPR* [53]. [4, 52]

[90] Bjorn Ommer and Joachim M. Buhmann. Learning the compositional nature of visual object categories for recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32:501–516, 2010. [52, 54]

[91] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, 1988. [7, 25, 26, 29, 63]

[92] Avi Pfeffer. Ibal: A probabilistic rational programming language. In Bernhard Nebel, editor, *IJCAI*, pages 733–740. Morgan Kaufmann, 2001. [36]

[93] Hamed Pirsiavash and Deva Ramanan. Parsing videos of actions with segmental grammars. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 612–619. IEEE, 2014. [61, 64, 113]

[94] John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *ADVANCES IN LARGE MARGIN CLASSIFIERS*, pages 61–74. MIT Press, 1999. [76]

[95] Jake Porway, Qiongchen Wang, and Song Chun Zhu. A hierarchical and contextual model for aerial image parsing. *Int. J. Comput. Vision*, 88:254–283, June 2010. [56]

[96] Luc De Raedt, Paolo Frasconi, Kristian Kersting, and Stephen Muggleton, editors. *Probabilistic Inductive Logic Programming - Theory and Applications*, volume 4911 of *Lecture Notes in Computer Science*. Springer, 2008. [180, 184]

[97] Luc De Raedt and Kristian Kersting. Probabilistic inductive logic programming. In Raedt et al. [96], pages 1–27. [36]

[98] George Rebane and Judea Pearl. The recovery of causal poly-trees from statistical data. In Laveen N. Kanal, Tod S. Levitt, and John F. Lemmer, editors, *UAI '87: Proceedings of the Third Annual Conference on Uncertainty in Artificial Intelligence, Seattle, WA, USA, July 10-12, 1987*, pages 175–182. Elsevier, 1987. [30]

[99] Matthew Richardson and Pedro Domingos. Markov logic networks. *Machine Learning*, 62(1-2):107–136, 2006. [32, 35, 41, 43, 62, 64, 98, 161]

[100] Sebastian Riedel and Ewan Klein. Genic interaction extraction with semantic and syntactic chains. In *In Proceedings of the Fourth Workshop on Learning Language in Logic*, pages 69–74, 2005. [44]

[101] Maximilian Riesenhuber and Tomaso Poggio. Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2:1019–1025, 1999. [52, 55, 70, 71]

[102] Ivan A. Sag, Thomas Wasow, and Emily M. Bender. *Syntactic Theory: A Formal Introduction*. CSLI, 2nd edition, 2003. [81]

[103] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117, 2015. [2, 3, 14]

[104] Georg Schneider, Heiko Wersing, Bernhard Sendhoff, and Edgar Körner. Evolutionary optimization of a hierarchical object recognition model. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 35(3):426–437, 2005. [60]

[105] Ehud Y. Shapiro. The model inference system. In Patrick J. Hayes, editor, *Proceedings of the 7th International Joint Conference on Artificial Intelligence (IJCAI '81), Vancouver, BC, Canada, August 1981*, page 1064. William Kaufmann, 1981. [43]

[106] A. Rashid B. M. Shariff, Max J. Egenhofer, and David M. Mark. Natural-language spatial relations between linear and areal objects: The topology and metric of english-language terms. *International Journal of Geographical Information Science*, 12(3):215–245, 1998. [15]

[107] R. Shneider and M. Riesenhuber. A detailed Look at Scale and Translation Invariance in a HIerarchical Neural Model of Visual Object Recognition. *CBCL Paper / AI Memo*, 218(Memo 2002-011), 2002. [52]

[108] Saurabh Singh, Abhinav Gupta, and Alexei A. Efros. Unsupervised discovery of mid-level discriminative patches. In *European Conference on Computer Vision*, 2012. [73, 75]

[109] Lauro Snidaro, Ingrid Visentini, and Karna Bryan. Fusing uncertain knowledge and evidence for maritime situational awareness via markov logic networks. *Information Fusion*, 21:159–172, 2015. [44, 45]

[110] Lauro Snidaro, Ingrid Visentini, Karna Bryan, and Gian Luca Foresti. Markov logic networks for context integration and situation assessment in maritime domain. In *15th International Conference on Information Fusion, FUSION 2012, Singapore, July 9-12, 2012*, pages 1534–1539. IEEE, 2012. [44, 45]

[111] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT press, 2nd edition, 2000. [29]

[112] Bartłomiej Stasiak and Mykhaylo Yatsymirskyy. *Methods and Supporting Technologies for Data Analysis*, chapter Frequency Domain Methods for Content-Based Image Retrieval in Multimedia Databases, pages 137–166. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009. [143]

[113] T. Poggio T. Serre, M. Kouh, C. Cadieu, U. Knoblich, G. Kreiman. A theory of object recognition: computations and circuits in the feedforward path of the ventral stream in primate visual cortex. Technical Report CBCL-259, MIT Artificial Intelligence Laboratory, dec 2005. [70]

[114] Benjamin Taskar, Pieter Abbeel, and Daphne Koller. Discriminative probabilistic models for relational data. In Adnan Darwiche and Nir Friedman, editors, *UAI*, pages 485–492. Morgan Kaufmann, 2002. [35]

[115] Sinisa Todorovic and Narendra Ahuja. Extracting subimages of an unknown category from a set of images. In *in CVPR*, pages 927–934, 2006. [54, 55]

[116] Sinisa Todorovic and Narendra Ahuja. Unsupervised category modeling, recognition, and segmentation in images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(12):2158–2174, 2008. [54, 128, 137, 140]

[117] Sinisa Todorovic and Michael C. Nechyba. Interpretation of complex scenes using dynamic tree-structure bayesian networks. *Comput. Vis. Image Underst.*, 106:71–84, April 2007. [54]

[118] Son Dinh Tran and Larry S. Davis. Event modeling and recognition using markov logic networks. In David A. Forsyth, Philip H. S. Torr, and Andrew Zisserman, editors, *Computer Vision - ECCV 2008, 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part II*, volume 5303 of *Lecture Notes in Computer Science*, pages 610–623. Springer, 2008. [45]

[119] Shimon Ullman. *High-Level Vision: Object Recognition and Visual Cognition*. The MIT Press, 1 edition, July 2000. [13]

[120] Paul A. Viola and Michael J. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR (1)*, pages 511–518. IEEE Computer Society, 2001. [71, 101]

[121] Kent Wittenburg and Louis Weitzman. Relational grammars: Theory and practice in a visual language interface for process modeling. In *In Workshop on Theory of Visual Languages*, pages 27–29. Springer Verlag, 1996. [19]

[122] Tianfu Wu, Gui-Song Xia, and Song Chun Zhu. Compositional boosting for computing hierarchical image structures. In *CVPR* [53]. [56]

[123] Tianfu Wu and Song Chun Zhu. A numerical study of the bottom-up and top-down inference processes in and-or graphs. *International Journal of Computer Vision*, 93(2):226–252, 2011. [56]

[124] Ying Nian Wu, Zhangzhang Si, Haifeng Gong, and Song Chun Zhu. Learning active basis model for object detection and recognition. *International Journal of Computer Vision*, 90(2):198–235, 2010. [60]

[125] Wei Zhang. *Statistical Inference and Probabilistic Modeling in Compositional Vision*. PhD thesis, Brown University, 2009. [57]

[126] Long Zhu, Yuanhao Chen, Alan L. Yuille, and William T. Freeman. Latent hierarchical structural learning for object detection. In *CVPR*, pages 1062–1069. IEEE, 2010. [60]

[127] Long Leo Zhu, Yuanhao Chen, and Alan Yuille. Unsupervised learning of a probabilistic grammar for object detection and parsing. In *in Advances in Neural Information Processing Systems 19*. MIT Press, 2007. [60]

[128] Song Chun Zhu and David Mumford. A stochastic grammar of images. *Foundations and Trends in Computer Graphics and Vision*, 2(4):259–362, 2006. [56, 57, 61]

[129] X. Zhu, D. Anguelov, and D. Ramanan. Capturing long-tail distribution of object subcategories. In *CVPR*, pages 1600–1607. IEEE, 2014. [89]

Glosario

Concatenación conjuntiva: se trata de enlazar de manera anidada funciones o predicados. Un ejemplo es: $A(B(C(arg)))$, donde A,B y C son funciones donde lo que devuelve una es argumento de otra.

Cuantificadores: permiten establecer al rango de valores que admite una variable. Los cuantificadores más conocidos son dos: cuantificador universal ($\forall$) y cuantificador existencial ($\exists$). El primero establece que deben aplicarse todos los valores posibles que admite la variable mientras que el cuantificador existencial establece que al menos uno de ellos debe admitirlo.

Lógica de primer orden: también se le llama cálculo de predicados, es una extensión de la lógica proposicional que incluye el uso de cuantificadores, que se aplican sobre los argumentos de los predicados. Un ejemplo de lógica de primer orden es $\forall x, f(x) \geqslant 0, x \in \Re$ Si la la expresión anterior es verdadera, sugiere que f probablemente es una función que dibuja una parábola. en este caso f es el predicado, y x es el argumento. Notar que el cuantificador universal $\forall$ se aplica sobre el argumento x.

Lógica de segundo orden: es una extensión de la lógica de primer orden, en donde los cuantificadores se aplican también sobre los predicados o funciones. Un ejemplo de lo anterior es: $\forall x, \exists f, f(x) \geqslant 0, x \in \Re$, la expresión anterior sugiere que es verdadera, porque en efecto, existen algunas funciones que pueden ser capaces de satisfacer la expresión (la familia de parábolas, por ejemplo) si el cuantificador fuera el universal, el enunciado sería falso, considerando que el universo de f son todas las funciones polinómicas posibles, y también sería falso si se incluyera la familia de funciones trascendentes.

Radiofrecuencia, identificación por: abreviado como RFID, son dispositivos pequeños en forma de etiquetas que tienen una antena en su interior con la intención de identificar de manera única al objeto en donde se encuentra pegada dicha etiqueta.

Sobrecarga de argumentos: en lenguajes de programación como C/C++ una función puede tener el mismo nombre pero argumentos diferentes, por ejemplo: $suma(int, int)$ y $suma(float, float)$. En algunos casos este tipo de sobrecarga de funciones puede estar prohibido. Cuando la prohibición es explícita, se debe renombrar la función de tal forma que quede: $sumaI(int, int)$, $sumaF(float, float)$.