Una gramática visual para la detección de rostros

por

Augusto Meléndez Teodoro

Tesis sometida como requisito parcial para obtener el grado de

Maestro 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

Una gramática visual para la detección de rostros

Tesis de Maestría

POR:

Augusto Meléndez Teodoro

ASESOR:

DR. LUIS ENRIQUE SUCAR SUCCAR

Instituto Nacional de Astrofísica Óptica y Electrónica
Coordinación de Ciencias Computacionales

TONANTZINTLA, PUEBLA. ENERO 2011

...A esa fuerza enorme y poderosa que hace que se mueva el Universo y puso cada cosa en su lugar...

...A mi padre que con su esfuerzo y ejemplo hoy termino una aventura más...

...A mi familia, que con su apoyo incondicional me motivaron a seguir adelante...

...A esas personas que me permitieron compartir o que compartieron conmigo segundos, minutos, horas o días de su tiempo...

...A esas personas que dentro del aula y fuera de ellas, compartieron su conocimiento y su experiencia conmigo, y así formar el ser humano que soy...

...Gracias B–Beto, porque sin palabras en un gimnasio me enseñaste a reír y a disfrutar cada momento que la vida te va regalando. Gracias por enseñarme a capturar esas pequeñas cosas que el mundo te regala. Gracias por mostrarme que quizá haya cosas que no se pueden cambiar, pero que de nosotros depende seguir viviendo. Gracias por mantenerme con los pies en la Tierra...

Resumen

Existen varios métodos de detección de rostros que se han desarrollado con cierto éxito, estos métodos suelen incluir características como textura, color de la piel, plantillas predefinidas o deformables, etc.; sin embargo, estos tienden a fallar en condiciones difíciles, tales como oclusiones parciales y cambios en la orientación y la iluminación en los rostros de las imágenes.

En esta tesis proponemos una nueva técnica para la detección de rostros basada en una gramática visual. Primero, definimos una *gramática simbólica–relacional* para rostros, para poder representar los elementos visuales de un rostro y sus relaciones espaciales. Después, esta gramática se transforma en una representación con base en una red bayesiana. La estructura de la red bayesiana es derivada a partir de la gramática y sus parámetros se obtienen a partir de datos; es decir, a partir de ejemplos positivos y negativos de rostros. Luego la red bayesiana es utilizada para la detección de rostros a través de la inferencia probabilística, utilizando un conjunto de detectores débiles para los diferentes componentes del rostro.

Evaluamos nuestro método utilizando un conjunto de imágenes de rostros en condiciones difíciles, y lo comparamos con un modelo simplificado sin relaciones espaciales, y el detector de rostros AdaBoost.

Los resultados muestran una mejora significativa en la detección de rostros al usar nuestro método basado en una gramática visual. Aunque la gramática está restringida a la representación de rostros, es posible extenderla para representar a una persona completa u otro tipo de objetos.

Abstract

Several methods for face detection have been developed with certain success, these methods typically include features like texture, skin color, some predefined templates or deformable templates, etc., however these tend to fail under “difficult” conditions such as partial occlusions and changes in orientation and illumination.

We propose a novel technique for face detection based on a visual grammar. We first define a symbol relational grammar for faces, representing the visual elements of a face and their spatial relations. This grammar is then transformed to a Bayesian network representation. The structure of the Bayesian network is derived form the grammar, and its parameters are obtained from data, i.e., from positive and negative examples of faces. Then the Bayesian network is used for face detection via probabilistic inference, using as evidence a set of weak detectors for different face components.

We evaluated our method on a set of sample images of faces under “difficult” conditions, and contrasted it with a simplified model without spatial relationships, and the AdaBoost face detector.

The results show a significant improvement when using our method based on a visual grammar. Although the grammar is restricted to the representation of faces, it is possible to extend it to represent a complete person or other object.

Tabla de Contenido

Lista de Figuras

Lista de Tablas

Capítulo 1

Introducción

1.1. Antecedentes

La detección y reconocimiento de rostros de forma automática es una tarea difícil e importante dentro del ámbito de la computación. Esta tarea es complicada debido a las condiciones difíciles en las que son tomadas las imágenes que se van a utilizar a la hora de la detección. Las formas en que se representa un rostro para ser procesados por una computadora son variadas, ahora ¿qué pasaría si pudiéramos representar un rostro con base en los elementos que lo componen y pudiéramos así representar las relaciones existentes entre dichos elementos? El reconocimiento de rostros a partir de reconocer por separado cada uno de sus elementos y después combinar estos resultados puede mejorar el reconocimiento del rostro completo.

Las diferentes condiciones en las que son tomadas las imágenes que se utilizan en la tarea de reconocimiento de rostros pueden presentar problemas como cambios en la iluminación, oclusión de objetos, ruido, diferentes posiciones del objeto de interés dentro de la imagen, etc. A estas condiciones las llamaremos de aquí en adelante *condiciones difíciles*. Cuando las imágenes son adquiridas en condiciones difíciles el proceso de reconocimiento resulta ser una tarea complicada y no siempre se obtiene un buen resultado, es decir, no siempre se logra detectar al rostro en la imágenes.

El representar al rostro a partir de sus elementos y de las relaciones que pueden existir entre los elementos nos puede ayudar a reconocer el rostro completo. Los elementos del rostro y sus relaciones se pueden representar con una gramática, en este caso la llamaremos *Gramática Visual*.

Esta gramática nos va a ayudar a representar un rostro para poder después hacer el proceso de detección en una imagen. Se propone definir una grámatica visual, uti-

lizando el formalismo para definir una *Gramática Simbólica–Relacional*, dado que existe *incertidumbre* en el reconocimiento y detección del rostro, representar esta incertidumbre es importante y un *Modelo gráfico probabilista*, en específico, representar a la gramática mediante una *red bayesiana*, nos permite hacer inferencia sobre las variables (elementos del rostro) de la red y con esto reconocer y detectar al rostro.

Se espera que con el uso de la gramática visual se mejore la *precisión*, considerando la reducción de los falsos positivos, en el proceso de detección del rostro en las condiciones difíciles antes mencionadas.

1.2. Definición del problema

La detección de rostros es el primer paso para el reconocimiento de rostros de forma automática. Por una parte en el *reconocimiento de rostros* se compara una imagen de entrada que contiene un rostro contra una base de datos de rostros y si existe alguna coincidencia se informa. Mientras que en la *detección de rostros* a partir de una imagen arbitraria se debe determinar si hay ó no hay rostros en la imagen, en el caso de que exista algún rostro en la imagen, se informa de la localización y tamaño de cada rostro (Yang, Kriegman, y Ahuja, 2002).

La confiabilidad de un sistema de detección de rostros tiene una influencia importante en el funcionamiento y utilidad del sistema completo de reconocimiento de rostro. Dado una imagen o un vídeo, un detector ideal de rostros debe poder identificar y situar todos los rostros presentes, sin importar su posición, escala, orientación, edad y expresión. Además, la detección se debe llevar a cabo independendientemente de condiciones de la iluminación y del contenido de la imagen o del vídeo (Li, Jain, y Li, 2005).

Los reconocedores existentes trabajan con imágenes en condiciones muy específicas. Por ejemplo, para el caso de la detección de rostros, las personas en las imágenes deben estar de frente para lograr su detección ya que si están en alguna otra posición no se logra un buen resultado. Cuando a una persona no se le distingue el rostro completo es difícil reconocerla en la imagen.

El objetivo de la detección de rostros es identificar toda la región de la imagen la cual contenga un rostro sin importar su posición tridimensional, orientación y condiciones de luz. El rostro es un objeto dinámico y tiene un alto grado de variabilidad en su apariencia, tamaño, forma, color y textura, lo cual hace que la detección de rostros sea un problema difícil en tareas de visión por computadora (Hjelmas y Low, 2001),

(Yang, Kriegman, y Ahuja, 2002).

De manera general podemos decir que los trabajos que tratan de solucionar la detección de objetos tienen problemas bajo condiciones difíciles en las imágenes.

El método de detección de rostros que se propone en esta tesis se basa en definir una gramática que sea capaz de representar un rostro. La gramática que se defina debe integrar tanto a los elementos que definen a un rostro como a las relaciones espaciales que existen entre dichos elementos. Por otra parte, se va a reconocer por separado los elementos que componen a un rostro completo, es decir, tendremos un reconocedor para cada elemento, después se integrarán los resultados de las detecciones individuales para obtener el resultado final.

1.3. Objetivos de la tesis

El objetivo general de esta tesis es desarrollar un modelo para representar rostros basado en una gramática visual y utilizarlo para la detección de rostros en condiciones difíciles.

Como objetivos específicos se plantean los siguientes:

1.- Definir de manera conceptual la gramática que nos permita definir un rostro.

2.- Desarrollar una representación de la gramática utilizando un modelo gráfico probabilista.

3.- Desarrollar un método para aprender los parámetros del modelo que representa la gramática.

4.- Desarrollar un método para reconocer rostros utilizando el modelo gráfico probabilista.

5.- Probar el modelo propuesto para detectar rostros en condiciones difíciles.

1.4. Solución propuesta

Este trabajo parte del hecho de que los métodos actuales para la detección de rostros no funcionan de manera eficiente cuando las imágenes de los rostro son adquiridas en algunas de las condiciones difíciles antes mencionadas. Por tal motivo, proponemos definir una gramática visual que nos permita representar los elementos que tiene un

rostro y algunas de las relaciones que se puedan establecer entre los elementos que se hayan definido.

Primero, analizamos los elementos en los que podemos dividir a un rostro; teniendo estos elementos definidos, establecimos una serie de reglas que incluyeran tanto a los elementos como a las relaciones espaciales entre dichos elementos. De manera manual se definió una gramática visual relacional para rostros, tomando en cuenta los elementos que se definieron y las relaciones espaciales entre dichos elementos.

En esta tesis se presenta la definición de una gramática visual, basada en el formalismo de una *gramática simbólica relacional*, para representar un rostro y después reconocerlo en imágenes que fueron adquiridas en condiciones difíciles. La gramática propuesta es transformada en una red bayesiana, la cual intengra tanto los elementos que componen a un rostro, así como las relaciones entre estos elementos. Después la red bayesiana es utilizada para la detección de rostros en condiciones difíciles haciendo inferencia sobre el modelo propuesto.

Se describen también los experimentos realizados al utilizar la gramática propuesta. En los experimentos realizados nos comparamos contra detectores basados en AdaBoost especializados en la detección de rostros. En comparación con los detectores de AdaBoost, nuestro método mostró una mejora significativa en la detección de rostros en condiciones difíciles.

1.5. Organización de la tesis

En el capítulo 2 se realiza un estudio de algunos formalismos que sirven para definir gramáticas visuales. En el capítulo 3 se presenta la teoría básica de las redes bayesianas. En el capítulo 4 se presentan y analizan trabajos relacionados en la detección de rostros. En el capítulo 5 se describe de manera detallada el método propuesto para solucionar la detección de rostros en condiciones difíciles. El capítulo 6 presenta los experimentos realizados y los resultados obtenidos y, finalmente, en el capítulo 7 se dan las conclusiones del trabajo realizado y se proponen algunos trabajos que en un futuro se pueden realizar.

Capítulo 2
Gramáticas visuales

En este capítulo se analizan algunos formalismos para definir Gramáticas visuales. Posteriormente se mencionan los trabajos más representativos en cuanto al uso de las gramáticas visuales para diferentes tareas.

2.1. Introducción

Los propósitos en la investigación de los lenguajes visuales son variados. Uno de los objetivos más importantes es lograr entender cómo los lenguajes visuales pueden ser clasificados de manera natural y cómo pueden ser especificados de manera concisa también de manera natural. Una de las principales motivaciones en la investigación de los lenguajes visuales ha sido facilitar la comunicación multimodal y facilitar también la interacción entre humanos y computadoras. Podemos decir que las gramáticas visuales surgen como uno de los enfoques para representar un lenguaje visual. El enfoque gramatical para representar un lenguaje visual ha sido fuertemente influenciado por el trabajo en teoría lingüística y teoría de lenguajes formales en las ciencias computacionales. Especificar un lenguaje visual al modificar una gramática de cadenas (lingüística) tiene la desventaja de que sólo se pueden especificar clases muy restringidas para lenguajes visuales (Marriot y Meyer, 1998).

Existen varios formalismos que nos permiten definir gramáticas visuales y representar lenguajes visuales. Todos los formalismos que se abordan en este capítulo están constituidos por una serie de reglas que definen las posibles configuraciones en el lenguaje visual. Una *gramática estocástica* (Zhu y Mumford, 2006) define un marco de trabajo que incluye tanto la representación de un objeto empleando un modelo probabilista, así como un proceso de aprendizaje y categorización de objetos. Otro formalismo

que se abordará es el que define una *gramática posicional* (Marriot y Meyer, 1998), que se basa en saber dónde está el siguiente símbolo de una gramática en relación con el símbolo actual que se esté tratando. En este tipo de gramáticas se incluyen ciertas relaciones arbitrarias entre los símbolos, estas relaciones dan información de la posición relativa de los símbolos que se estén analizando. Por otra parte, una *gramática relacional* (Marriot y Meyer, 1998) es un formalismo de alta dimensionalidad. Este enfoque hereda ideas de la lingüística computacional. Define una serie de limitaciones basadas en unificación para expresar las limitaciones espaciales y permitir el cálculo de estructuras de características en los atributos de alguna gramática pasiva. En una *gramática de relación* (Marriot y Meyer, 1998) se representan objetos visuales elementales partiendo de ciertas ocurrencias de símbolos, estos símbolos se relacionan a partir de ciertos elementos binarios. Con este tipo de gramáticas se logra un alto nivel de descripción en los lenguajes visuales.

2.2. Gramáticas estocásticas

A diferencia de una gramática tradicional que contiene cuatro elementos (conjunto de símbolos iniciales, conjunto de símbolos terminales, conjunto de reglas de producción y un símbolo inicial); una *gramática estocástica* contiene además un conjunto de probabilidades *P*, que reflejan un número alternativo de formas de reescribir una regla de producción en la gramática. En este tipo de gramáticas, cada regla de producción tiene asociado un valor de probabilidad. Una gramática debe de cubrir los siguientes puntos (Zhu y Mumford, 2006):

1.- *Un marco de trabajo común para la representación de conocimiento visual y la categorización de objetos.* La composición jerárquica y estructural es el concepto principal detrás de las gramáticas. La gramática representa tanto la descomposición jerárquica de escenas a objetos, sus partes, primitivas y píxeles para nodos terminales y no terminales, así como para representar el contexto para relaciones espaciales y funcionales por medio de ligas horizontales entre los nodos. Esto define cada categoría de objeto como el conjunto de todas las configuraciones válidas posibles producidas por la gramática.

2.- *Entrenamiento y generalización a partir de un conjunto pequeño de ejemplos.* Se define un modelo probabilista para representar la ocurrencia de frecuencia de los

objetos y sus partes, así como sus relaciones. Este modelo aprende de un conjunto de entrenamiento relativamente pequeño, y después se prueba para sintetizar un gran número de configuraciones para cubrir nuevas instancias de objetos en el conjunto de prueba.

3.- *Mapeo del vocabulario visual para llenar el vacío semántico.* Para llenar el vacío que existe entre los símbolos y las imágenes originales, la gramática incluye una serie de diccionarios visuales que organiza a través de una composición de grafos. El nivel inferior del diccionario es un conjunto de imágenes primitivas las cuales tienen un punto de referencia con lazos abiertos para ligarse con otras primitivas.

A continuación describiremos de manera breve el grafo *And-Or*, como ejemplo de una gramática estocástica que se utiliza para representar objetos en las imágenes (Zhu y Mumford, 2006).

Grafo *And-Or*

Un grafo *And-Or* representa la gramática de la imagen completa y contiene todos los *parse graphs* válidos. Un *parse graph* es una interpretación de una imagen específica. En el grafo *And-Or*, cada nodo no terminal *A* puede ser representado por un nodo-*Or* con *n(A)* estructuras alternativas, las cuales conforman un nodo-*And* compuesto por un número de sub-estructuras. Por tanto el grafo *And-Or* nos permite representar una gramática de imagen sensible al contexto.

De manera formal un grafo *And-Or* es una 6-tupla para representar una gramática de imagen *G*.

$$G_{and-or} =< S, V_N, V_T, R, \Sigma, P >$$

donde:

- *S* es el nodo raíz para una escena o algún objeto.
- $V_N = V^{and} \cup V^{or}$ es un conjunto de nodos no terminales que incluye un conjunto de nodos-And, V^{and}, y un conjunto de nodos-Or, V^{or}. Los nodos-And representan las producciones y los nodos-Or el vocabulario de los elementos.

- V_T es un conjunto de nodos terminales para primitivas, partes de objetos y objetos (note que un objeto en baja resolución podría terminar sin descomposición directamente).
- *R* es el número de relaciones entre los nodos.
- Σ es el conjunto de todas las configuraciones derivables válidas de la gramática, es decir, su lenguaje.
- *P* es el modelo de probabilidad definido para el grafo *And-Or*.

El la Figura 2.1 podemos observar un ejemplo del empleo del grafo *And-Or* para representar una imagen.

2.3. Gramática Posicional (*Positional Grammars*)

Una Gramática posicional (Marriot y Meyer, 1998) se basa en el hecho de saber dónde está el siguiente símbolo en relación con el símbolo actual. Este proceso se realiza al permitir una relación espacial arbitraria REL_i, la cual da información acerca de la posición relativa del siguiente símbolo $\alpha_{(i+1)}$ con respecto al símbolo actual α_i. Las producciones en una gramática posicional tienen la forma:

$$A \rightarrow \alpha_1 REL_1 \alpha_2 REL_2 ... REL_{n-1} \alpha_n$$

donde *A* es un símbolo no terminal , los α_i pueden ser símbolos terminales o no terminales y REL_i especifica la posición de α_{i+1} en relación con α_i.

Las gramáticas posicionales están bien adaptadas para describir la composición bidimensional del texto y describir algunas formas de lenguajes icónicos. Sin embargo, una gramática posicional en su forma básica tiene muchas restricciones en lo que puede expresar. La principal dificultad se deriva de la necesidad de que cada símbolo debe tener una coordenada simple y con esto debe ser posible determinar las coordenadas del sucesor único del siguiente token en la entrada dada la coordenada del símbolo actual.

De manera formal podemos definir una *Gramática Posicional Libre de Contexto, PG,* como una 6-tupla

$$PG = (S, N, T, POS, P, PE)$$

donde:

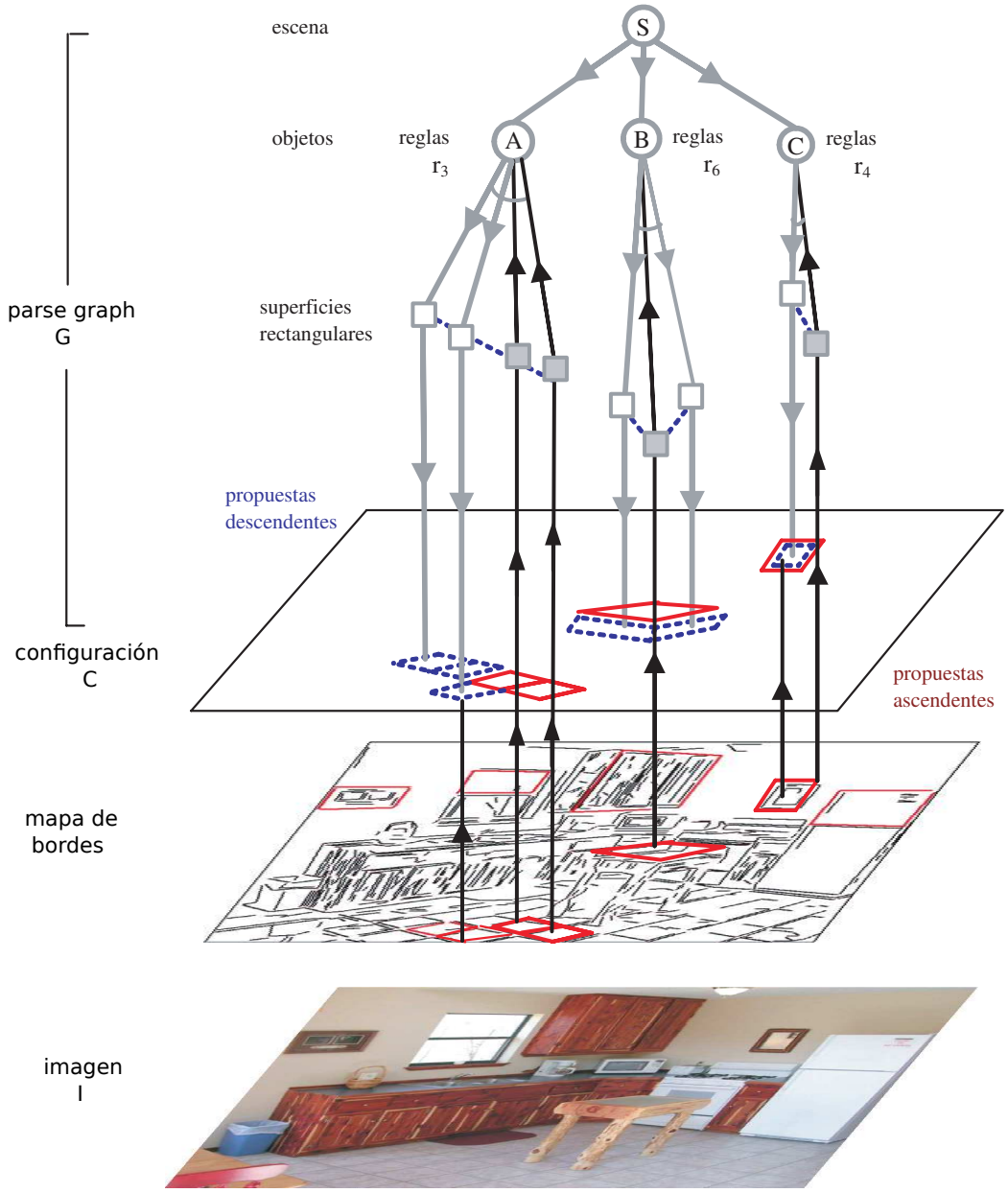


Figura 2.1: El grafo And-Or refleja la estructura de la imagen. Cada nodo cuadrado representa una primitiva del objeto, mientras que nodos circulares representan la combinación de estas primitivas para formar un objeto completo. El nodo *S* representa el símbolo inicial de la gramática. Tomado de (Zhu y Mumford, 2006).

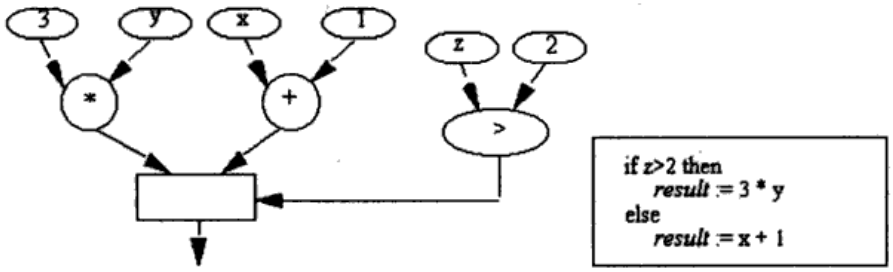


Figura 2.2: Forma de representar un algoritmo utilizando una Gramática Posicional. Para esta gramática los nodos circulares representan operadores como: +, -, *, /. Los nodos ovalados grandes representan símbolos de comparación como: $<$, $>$, $=$, $\neq$. Los óvalos pequeños representan variables en el algoritmo. El nodo rectangular recibe varias entradas de los nodos anteriores y ofrece una salida. Tomado de (Marriot y Meyer, 1998).

- $S \in N$ es el elemento inicial y es un elemento no terminal.
- N es un conjunto finito no vacío de elementos no terminales.
- T es un conjunto finito no vacío de elementos terminales, con $N \cap T = \phi$.
- POS es un conjunto finito de identificadores de las relaciones binarias.
- P es un conjunto finito de producciones.
- PE es un evaluador pictórico. Es una función la cual transforma un enunciado derivado a partir de la gramática en su correspondiente forma pictórica.

En la Figura 2.2 podemos observar cómo se representa una parte de un algoritmo. El algoritmo es un ejemplo de la condicional *IF*.

2.4. Gramáticas de Relación (*Relation Grammars*)

Una Gramática simbólica relacional, *Symbol Relation Grammars* (Marriot y Meyer, 1998), es un formalismo en el que cada frase está compuesta por un conjunto de ocurrencias de símbolos *(symbol occurrences)*, que representan objetos visuales elementales, los cuales están relacionados por medio de un conjunto de elementos relacionales *(relational elementos)* binarios. La principal característica de una *Gramática simbólica relacional* es que utiliza las producciones libres de contexto para volver a escribir la ocurrencia de los símbolos así como los elementos relacionales.

Un enfoque común para la descripción formal de lenguajes visuales hace uso de gramáticas formales y reescribe producciones capaces de generar enunciados visuales.

Un enunciado visual es una disposición de elementos pictóricos (imágenes) relacionados en dos o más dimensiones.

El formalismo para definir una *Gramática simbólica relacional* proporciona un nivel de descripción alto de los lenguajes visuales. En una *Gramática simbólica relacional* un enunciado es visto como un conjunto de ocurrencias de símbolos (*s-elementos*) y un conjunto de elementos relacionales (*r-elementos*) sobre las ocurrencias de los símbolos. Se define un mecanismo para volver a escribir tanto los *s-elementos* como los *r-elementos* por medio de reglas de producción libres de contexto. Como resultado, las relaciones sobre los símbolos terminales son especificados de una manera muy natural independientemente de la implementación de alguna interfaz gráfica.

De manera formal una *Gramática simbólica relacional* es una 6-tupla

$$G = (S, V_N, V_T, V_R, P, R)$$

donde:

- $S \in V_N$ es el símbolo inicial.
- V_N es un conjunto no vacío de símbolos no terminales.
- V_T es un conjunto no vacío de símbolos terminales.
- V_R es un conjunto finito de símbolos relacionales (*relation symbols*).
- P es un conjunto de reglas de re-escritura llamadas *producciones s-elementos* de la forma:

  $$l : Y^0 \rightarrow < M, R >$$

  donde:

  - *l* es un identificador único de la *producción s-elemento*.
  - $< M, R >$ es un enunciado sobre V_R y $V_N \cup V_T$
    - *M* es el conjunto de s-elementos (v, i), $v \in V_T$ e *i* es un número natural. Por simplicidad cada s-elemento se escribe v_i.
    - *R* es el conjunto de r-elementos de la forma $r(X_i, Y_j), X_i, Y_j \in M, r \in V_R$ indicando que una relación *r* se mantiene entre X_i e Y_j.
  - $Y \in V_N, Y^0 \notin M$

- *R* es un conjunto finito de reglas re-escritas llamadas *producciones r-elementos*, de la forma

$$s(Y^0, X^1) \rightarrow [l]Q \; o \; s(X^1, Y^0) \rightarrow [l]Q$$

donde:

  - $s \in V_R$.
  - *l* es la etiqueta de una *producción r-elemento* $Y^0 \rightarrow < M, R >$
  - $X \in V_N \cup V_T$ y $X^1 \notin M$
  - $Q \neq \phi$ es un conjunto finito de r-elementos de la forma $r(Z^i, X^1)$ o $r(X^1, Z^i)$, $Z^i \in M$

2.5. Trabajos relacionados con gramáticas visuales

La utilización de las gramáticas visuales para reconocer objetos entró en una etapa de "hibernación" en los 1970s debido a diversas dificultades que aún hoy en día siguen siendo un problema (Zhu y Mumford, 2006). Algunas de estas dificultades son:

- Hay una enorme cantidad de conocimiento visual en escenas del mundo real que tienen que ser representadas en la computadora para hacer inferencia sobre ellas.
- La complejidad computacional es enorme. En una imagen podemos encontrar una gran cantidad de objetos y tanto los algoritmos de aprendizaje como los sistemas de visión por computadora rara vez poseen un conocimiento visual suficiente para realizar tareas de razonamiento sobre dichas imágenes.
- Uno no puede calcular de manera fiable los símbolos de las imágenes originales. Los sistemas confunden con regularidad los objetos que aparecen en las imágenes.

A pesar de que aún hay dificultades, con los progresos recientes en el aprendizaje automático y en los modelos estadísticos, el importante y constante aumento en la potencia de cálculo, además del desarrollo de detectores más robustos, han hecho posible el renacimiento de las gramáticas visuales con enfoques interesantes en estos últimos años. Lo que han hecho que las limitaciones antes mencionadas puedan ser atacadas.

Song-Chun y Munford (Zhu y Mumford, 2006) presentan uno de los trabajos más importantes realcionado con las gramáticas visuales. En este trabajo se define una

Gramática de Imagen utilizando un *Grafo And-Or*. Song-Chun y Munford proponen desarrollar un marco de trabajo consistente de representación para la gran cantidad de conocimiento visual en todos los niveles de abstracción. Dentro del grafo *And-Or*, se define un *Parse Graph* que es una representación de una imagen específica, mientras que el grafo *And-Or* incluye la gramática de la imagen completa y contiene todos los *parse graph* válidos. En este trabajo también se menciona que una gramática debe de ser capaz de representar la descomposición jerárquica de alguna imagen, así como las relaciones espaciales y funcionales de esta imagen con las partes que la componen. La imagen debe ser representada utilizando alguna estructura de datos, por ejemplo utilizando grafos. La gramática que se está definiendo debe ser representada utilizando algún modelo probabilista con el fin de relacionar tanto a la imagen como a sus partes. Song-Chun y Munford utilizan diferentes tipos de imágenes como animales, rostros, automóviles, cocinas, etc. Para representar la imagen dividida en sus elementos se utiliza un *Grafo de Regiones Adyacentes (RAG)* donde se agregan arcos horizontales que representen la adyacencia de las diferentes regiones o elementos de la imagen. Otra manera presentada en (Zhu y Mumford, 2006) para representar a la imagen dividida es utilizando diferentes tipos de imágenes básicas o primitivas dependiendo del tamaño de la imagen a dividir. Para poder llevar a cabo el reconocimiento de los objetos, realizan inferencia *descendente (top-down)* y *ascendente (bottom-up)* en el RAG de cada imagen.

Tian-Fu, Gui-Song y Song-Chun (Wu, Xia, y Zhu, 2007) proponen un algoritmo de boosting composicional, el cual está formado de dos etapas: *propuesta ascendente (bottom-up)* en la cual se generan candidatos para alguna parte de una imagen y *validación descendente (top-down)* que se encarga de verificar si los candidatos generados en la etapa anterior son válidos. Toman en cuenta la forma en que están conectadas las estructuras a detectar en las imágenes. Este algoritmo es utilizado para detectar y reconocer 17 estructuras de imágenes comunes en tareas de visión de nivel bajo-medio. Estas estructuras son utilizadas como primitivas para representar diferentes imágenes en este trabajo. Tian-Fu, Gui-Song y Song-Chun toman en cuenta la forma en que están conectadas las estructuras a detectar en las imágenes. Se utiliza un Grafo *And-Or* para representar a la imagen y los arcos tienen asociados un valor de peso el cual es obtenido de manera probabilista dependiendo de un segmento de imagen dada la imagen completa.

Feng Han y Song-Chun (Han y Zhu, 2005) proponen una *Gramática para un Grafo de Atributos* para analizar escenas con objetos hechos por el hombre, como edificios, pasillos, cocinas y habitaciones. En este trabajo sólo se utiliza una primitiva 3D, un rectángulo, y se trabaja con seis reglas de producción. Las reglas de producción, además de expandir un nodo o segmento de la imagen del grafo en sus componentes, también incluyen un número de ecuaciones que limitan los atributos de un nodo padre y los de sus hijos. Feng Han y Song-Chun utilizan vectores de atributos para los nodos terminales y los no terminales, estos vectores después son utilizados para definir ecuaciones que limitan los atributos de un nodo padre a sus hijos aplicando una serie de reglas de producción.

Zhu, Chen y Yuille (Zhu, Chen, y Yuille, 2009) proponen una Gramática probabilística utilizando Modelos de Markov. Esta gramática está definida por un *campo aleatorio de Markov* el cual es importante porque con este modelo se pueden integrar relaciones espaciales a la gramática. La gramática propuesta es probada en varios tipos de objetos, incluyendo imágenes de rostros, aviones, bicicletas y algunos animales.

Varios ejemplos de aplicaciones del uso de gramáticas son presentados en (Marriot y Meyer, 1998). En uno de los ejemplos, se muestra la traducción de un diagrama de flujo semi-estructurado de su representación gráfica al código fuente en Pascal. Para esta grámatica se definen 8 puntos terminales, 13 puntos de unión y 14 reglas de producción. Algunos puntos terminales representan condiciones Booleanas e instrucciones, también hay puntos que sirven para representar enunciados compuestos y puntos que sirven para definir ciclos. Los otros ejemplos son del mismo estilo, es decir, se hace la traducción de diagramas de flujo a su equivalente en algún código fuente de algún lenguaje. En estos ejemplos no se hace la representación de imágenes o de objetos, sino que se trabaja con diagramas de flujo.

2.6. Discusión sobre las gramáticas

Los formalismos para definir una gramática visual que se mencionaron en este capítulo nos ofrecen ideas para definir la gramática para representar el rostro.

Por una parte, las *gramáticas estocásticas* nos permiten integrar a la gramática del rostro, ciertos valores de probabilidad para cada una de las relaciones espaciales que se establecieron entre los elementos del rostro. Este tipo de gramáticas nos permite utilizar un modelo gráfico probabilista para representar la gramática del rostro. Con este tipo de

gramáticas se puede establecer un proceso de detección del rostro, a partir del modelo gráfico que se haya definido para la representación de la gramática.

Por otro lado, la *gramática símbolica relacional* nos permite definir de manera formal, tanto las reglas que nos definen un rostro, como las relaciones espaciales que existen entre los elementos del rostro que se tomaron en cuenta. Vale la pena señalar que un *enunciado simbólico–relacional*, el cual es concebido como un conjunto de ocurrencias de símbolos junto con un conjunto de relaciones entre dichos símbolos, puede ser facilmente interpretado como un grafo dirigido etiquetado, donde los s-elementos corresponden a los nodos etiquetados y los r-elementos corresponden a los arcos etiquetados.

Un grafo dirigido etiquetado nos ofrece la ventaja de poder incluir las realciones espaciales, pero tenemos la desventaja de no poder representar la incertidumbre. Por otra parte, una red bayesiana nos permite representar la incertidumbre y también nos permite representar las relaciones espaciales.

Tomando en cuenta lo anterior, representamos la gramática para un rostro utilizando una red bayesiana, en donde los elementos y las relaciones espaciales entre los elementos son representados como nodos de la red.

2.7. Conclusiones

En este capítulo se abordó la teoría básica para definir una gramática y así, definir una gramática para representar un *rostro*.

La gramática que se define integra tanto los elementos del rostro, como algunas *relaciones espaciales* entre éstos elementos. Los elementos terminales y no terminales de la grámatica se tomaron en cuenta dado el conocimiento previo que tenemos de los posibles elementos en los cuales podemos descomponer el rostro y también considerando los detectores que tenemos disponibles.

A diferencia de los trabajos antes descritos, nosotros no utilizamos primitivas de bajo nivel (píxeles, líneas, etc.), nosotros trabajamos con objetos más complejos los cuales son detectados por los reconocedores basados en AdaBoost.

Hasta este momento tenemos definida de manera conceptual la gramática visual, ahora tenemos que encontrar un modelo que nos permita representar la gramática del rostro. Este modelo nos debe permitir integrar tanto los símbolos de la gramática como las relaciones entre dichos símbolos.

En el siguiente capítulo se aborda la teoría de las redes bayesianas, que es el modelo que utilizamos para representar la gramática.

Capítulo 3

Redes bayesianas

En este capítulo se presenta de manera breve la teoría sobre las *redes bayesianas*, ya que es el modelo utilizado para representar la gramática visual. Comenzamos dando una definición de red bayesiana, posteriormente se explica la manera en la que se hace inferencia y propagación en una red bayesiana. Posteriormente, se exponen algunos trabajos donde se aplican redes bayesianas a imágenes.

3.1. Introducción

Las redes bayesianas (Sierra, 2006) modelan un fenómeno mediante un conjunto de variables y las relaciones de dependencia entre ellas. Dado este modelo, se puede hacer inferencia bayesiana; es decir, estimar la probabilidad posterior de las variables no conocidas, con base en las variables conocidas. Además, las redes bayesianas pueden dar información interesante en cuanto a cómo se relacionan las variables del dominio, las cuales pueden ser interpretadas como relaciones de causa-efecto.

Inicialmente, estos modelos eran construidos manualmente basados en un conocimiento experto, pero en los últimos años se han desarrollado diversas técnicas para aprender a partir de datos, tanto la estructura como los parámetros asociados al modelo, aunque también es posible combinar conocimiento experto con los datos para aprender el modelo (Sierra, 2006).

3.2. Redes bayesianas

Las *redes bayesianas* son estructuras gráficas para representar las relaciones de probabilidad sobre un gran número de variables, y para hacer inferencia probabilista con

estas variables. La naturaleza gráfica de las redes bayesianas nos da una comprensión intuitiva de las relaciones entre las variables. Una forma de construir redes bayesianas es crear arcos que representen influencias directas entre las variables. Las probabilidades en la red son las probabilidades condicionales de los valores de cada variable, dada cada combinación de valores de los padres de las variables en la red, excepto en el caso del nodo raíz, donde se requieren las probabilidades *a priori*. La inferencia probabilista sobre las variables puede ser realizada usando la red bayesiana (Neapolitan, 2004).

En una red bayesiana todas las relaciones de independencia condicional representadas en el grafo corresponden a relaciones de independencia en la distribución de probabilidad. Dichas independencias simplifican la representación del conocimiento y el razonamiento. Una red bayesiana representa en forma gráfica las dependencias e independencias entre variables aleatorias, en particular las independencias condicionales. Lo anterior se representa con la siguiente notación, para el caso de X independiente de Y dado Z:

- Independencia en la distribución: $P(X|Y,Z) = P(X|Z)$.
- Independencia en el grafo: $I < X|Z|Y >$

Por ejemplo, en la Figura 3.1 podemos ver un ejemplo de una red bayesiana en donde tenemos cinco variables aleatorias: *Humedad, Temperatura, Nubosidad, Lluvia, Organización de Fiesta*. En esta red, "Organización de Fiesta" es *independiente* de "Temperatura" dado "Lluvia".

3.3. Inferencia

El razonamiento probabilístico o propagación de probabilidades consiste en propagar los efectos de la evidencia a través de la red para conocer la probabilidad *a posteriori* de las variables. Es decir, se le dan valores a ciertas variables (evidencia), y se obtiene la probabilidad posterior de las demás variables dadas las variables conocidas (si el conjunto de variables conocidas es vacío, entonces se obtienen las probabilidades *a priori*). Existen diferentes tipos de algoritmos para calcular las probabilidades posteriores, que dependen del tipo de grafo y de si obtienen la probabilidad de una variable a la vez o de todas. Los principales tipos de algoritmos de inferencia son (Sierra, 2006):

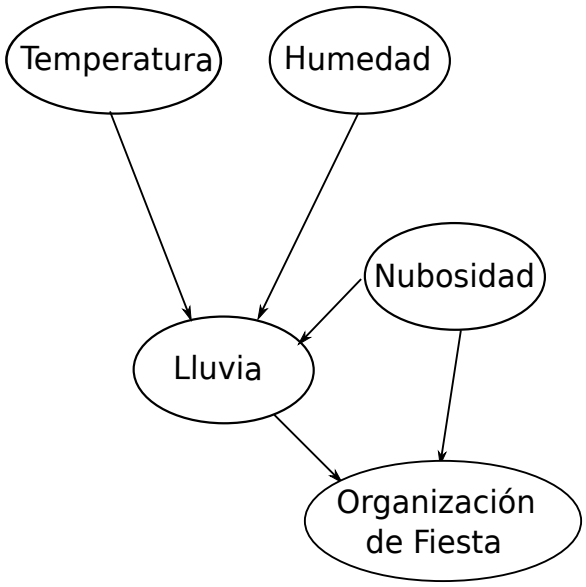


Figura 3.1: Ejemplo de una red bayesiana. Los nodos representan variables aleatorias y los arcos relaciones de dependencia.

1.- Una variable, cualquier estructura: algoritmo de eliminación (*variable elimination*). El método de eliminación de variables consiste en distribuir sumas dentro de productos, esto para evitar cálculos innecesarios. Este algoritmo requiere como entrada un orden de eliminación, el cual es un problema NP-completo.

2.- Cualquier variable, estructuras sencillamente conectada: algoritmo de propagación de Pearl (Pearl y Dechter, 1989). Este algoritmo determina $P(X|e)$ para todos los valores e de la variable E de la red. El paso de mensajes es realizado iniciando mensajes desde cada variable instanciada hacia sus vecinos, a su vez éstos pasan mensajes a sus vecinos. La evidencia puede llegar en cualquier orden. Este método es aplicable a redes bayesianas con estructura de árbol y poliárbol.

3.- Cualquier variable, cualquier estructura: (i) Agrupamiento (*junction tree*). El método de agrupamiento consiste en transformar la estructura de la red para obtener un árbol, mediante agrupación de nodos usando la teoría de grafos. (ii) Simulación estocástica. Se asignan valores aleatorios a las variables no instanciadas, se calcula la distribución de probabilidad y se obtienen valores de cada variable dando una muestra. Se repite el procedimiento para obtener un número apreciable de muestras y en base al número de ocurrencias de cada valor se determina la probabilidad de dicha variable. (iii) Condicionamiento. Si instanciamos una

variable, ésta bloquea las trayectorias de propagación. Entonces asumiendo valores para un grupo seleccionado de variables podemos descomponer la gráfica en un conjunto de grafos conectados de forma sencilla. Propagamos para cada valor posible de dichas variables y luego promediamos las probabilidades ponderadas.

Propagación en redes multiconectadas

El algoritmo general más común (Sierra, 2006) en redes bayesianas es el de agrupamiento o "árbol de uniones" (*junction tree*). Este método consiste en transformar la estructura de la red para obtener un árbol, mediante agrupación de nodos usando la teoría de grafos. Para ello, se hace una transformación de la red a un árbol de uniones (grupos de nodos) mediante el siguiente procedimiento:

1.- Eliminar la dirección en los arcos.

2.- Ordenar los nodos por máxima cardinalidad.

3.- Moralizar el grafo (añadir arcos entre nodos con hijos comunes, es decir, todos los nodos padres están relacionados).

4.- Triangular el grafo.

5.- Obtener los cliques y ordenarlos.

6.- Construir el árbol de cliques.

Un *clique* es un subconjunto de nodos completamente conectados, de forma que hay un arco entre cada par de nodos, y no existe un conjunto completamente conectado del que éste sea subconjunto.

Una vez transformado el grafo, la propagación se hace mediante el envío de mensajes en el árbol de uniones o cliques. Inicialmente se calcula la probabilidad conjunta (potencial) de cada clique, y la condicional dado el padre. Dada cierta evidencia se recalculan las probabilidades de cada clique. La probabilidad individual de cada variable se obtiene de la del clique por marginalización.

En el peor de los casos, la propagación en redes bayesianas es un problema NP-duro. En la práctica se tienen redes no muy densamente conectadas y la propagación es eficiente aún para redes muy grandes (función del clique mayor). Para redes muy complejas, es decir, que tienen muchas conexiones, la mejor alternativa son técnicas de simulación estocástica o técnicas aproximadas.

3.4. Aprendizaje de redes bayesianas

El aprendizaje de redes bayesianas consiste en inducir un modelo, estructura y parámetros asociados, a partir de datos (Sierra, 2006). Este puede dividirse en dos partes:

1.- Aprendizaje estructural, en el cual se obtiene la estructura o topología de la red.

2.- Aprendizaje paramétrico, en el cual dada la estructura, se obtienen las probabilidades asociadas.

3.4.1. Aprendizaje paramétrico

El método más común para obtener los parámetros de una red bayesiana suponiendo que la estructura ya está dada, es el llamado *estimador de máxima verosimilitud* bajo el cual se estiman las probabilidades con base en las frecuencias de los datos. Para una red bayesiana se tienen dos casos:

- Nodos raíz. Se estima la probabilidad marginal. Por ejemplo: $P(A_i) \sim NA_i/N$, donde NA_i es el número de ocurrencias del valor i de la variable A, y N es el número total de casos o registros.

- Nodos hojas. Se estima la probabilidad condicional de la variable dados sus padres. Por ejemplo: $P(B_i|A_j, C_k) \sim NB_iA_jC_k/NA_jC_k$, donde $NB_iA_jC_k$ es el número de casos en que $B = B_i$, $A = A_j$ y $C = C_k$, y NA_jC_k es el número de casos en que $A = A_j$ y $C = C_k$.

Discretización de variables continuas

En general las redes bayesianas se aplican a variables discretas, por lo que si se tienen variables continuas hay que discretizarlas (existen métodos de inferencia para variables continuas pero están limitados a distribuciones gaussianas).

Uno de los métodos no supervisados más simples para la discretización de variables (Sierra, 2006), consiste en dividir el rango de valores de cada atributo, $[Xmin, Xmax]$, en k intervalos, donde k es dado por el usuario u obtenido usando una cierta medida de información sobre los valores de los atributos.

En cuanto a los métodos supervisados, que consideran la variable clase, los puntos de división para formar rangos en cada atributo son seleccionados en función del valor de la clase. El problema de encontrar el número óptimo de intervalos y de los límites

correspondientes se puede considerar como un problema de búsqueda. Es decir, podemos generar todos los puntos posibles de división para formar intervalos sobre la gama de valores de cada atributo (donde hay un cambio en la clase), y estimamos el error de clasificación para cada partición posible.

Para el caso general de redes bayesianas, existe un método para la discretización de atributos continuos, mientras se aprende la estructura de la red bayesiana (Sierra, 2006). La discretización está basada en el principio de MDL (por sus siglas en inglés de Longitud de Descripción Mínima), considerando el número de intervalos de una variable con respecto a sus vecinos en la red. Para una estructura dada, un procedimiento de búsqueda local encuentra la discretización de una variable que reduce al mínimo la longitud de la descripción referente a los nodos adyacentes en la red, y éste se repite en forma iterativa para cada una de las variables continuas.

3.4.2. Aprendizaje estructural

El aprendizaje estructural consiste en encontrar las relaciones de dependencia entre las variables, de forma que se pueda determinar la topología o estructura de la red bayesiana.

Existen dos clases de métodos para el aprendizaje genérico de redes bayesianas (Sierra, 2006). Estos son:

1.- Métodos basados en medidas de ajuste y búsqueda.

En esta clase de métodos se tiene una evaluación global de la estructura respecto a los datos. Es decir, se generan diferentes estructuras y se evalúan respecto a los datos utilizando alguna medida de ajuste. Las variantes en estos métodos dependen de la medida de ajuste de la estructura a los datos y el método de búsqueda de la mejor estructura.

En cuanto a las medidas de ajuste, las dos más comunes son la medida bayesiana y la medida basada en el principio de longitud de descripción mínima (MDL). La medida bayesiana busca maximizar la probabilidad de la estructura dados los datos, mientras que la medida MDL hace un compromiso entre la exactitud y la complejidad del modelo. La exactitud se estima midiendo la información mutua entre los atributos y las clases; y la complejidad contando el número de parámetros.

Una vez establecida una forma de medir la calidad de la estructura, se establece

un método para hacer una búsqueda de la mejor estructura entre todas las estructuras posibles. Dado que el número de posibles estructuras es exponencial en el número de variables, es imposible evaluar todas las estructuras, por lo que se hace una búsqueda heurística. Una estrategia común es utilizar búsqueda de ascenso de colina (*hill climbing*), en la cual se inicia con una estructura simple que se va mejorando hasta llegar a la mejor estructura.

2.- Métodos basados en pruebas de independencia.

Este enfoque se basa en medidas de dependencia local entre subconjuntos de variables. El caso más sencillo es el del algoritmo de Chow y Liu (Chow, Member, y Liu, 1968), en el cual se mide la información mutua entre pares de variables. A partir de estas medidas se genera una red bayesiana en forma de árbol.

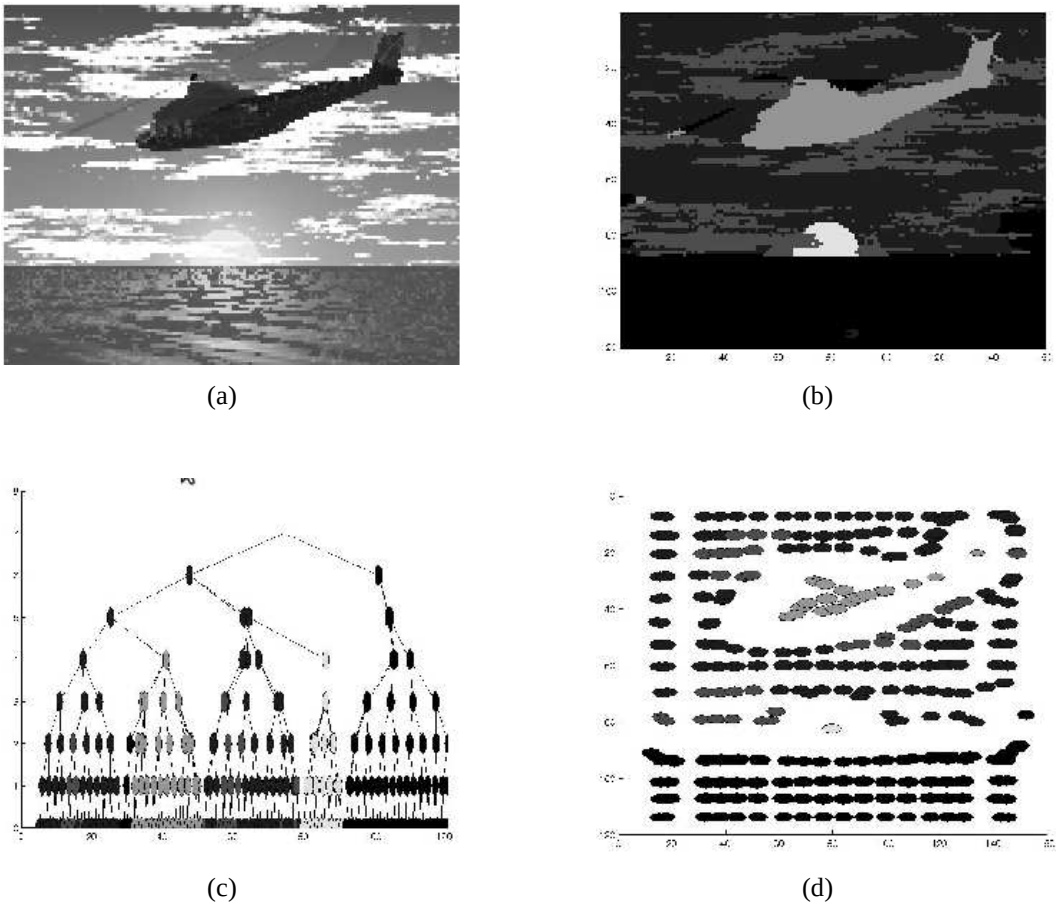
Este enfoque se puede generalizar para el aprendizaje de redes multiconectadas, haciendo pruebas de dependencia entre subconjuntos de variables, normalmente dos o tres variables. La desventaja es que se pueden generar muchos arcos innecesarios, por lo que se incorporan formas de eliminar estos arcos.

3.5. Redes bayesianas en la representación de imágenes

Una de las maneras de llevar el reconocimiento de objetos es descomponiendo la imagen en cierto número de elementos, cuando se lleva a cabo esta descomposición es necesario definir algún modelo que nos permita representar todos los elementos de la imagen. En este trabajo de tesis se propone utilizar una red bayesiana como modelo para representar la estructura del rostro. A continuación se presentan los trabajos más relevantes enfocados a describir la estructura de las imágenes basados en modelos probabilistas.

Storkey y Williams (Storkey y Williams, 2003) proponen utilizar un modelo probabilista basado en árboles dinámicos llamado *Position-Encoding Dynamic Tree (PEDT)* para representar una imagen. Un PEDT es un modelo probabilista que mejora un árbol dinámico normal al permitir que la posición de los objetos formen parte del modelo. Esto incrementa la flexibilidad del modelo sobre los árboles dinámicos y permite que las posiciones de los objetos puedan ser ubicados y manipulados. El PEDT usa estructura de árbol jerárquico de nodos. Cada nodo tiene una etiqueta que denota el tipo de objeto que se está reresentando, además cuenta con la posición que representa la ubicación

espacial del objeto que está siendo representado. Los nodos que están más arriba del árbol representan regiones grandes de la imagen, mientras que los nodos de más abajo representan la etiqueta de un píxel o de un conjunto de píxeles. Al utilizar esta estructura tenemos el problema de que, al tomar en cuenta píxeles o conjuntos de píxeles como nodos del árbol, la estructura crece en gran medida, por lo que a la hora de hacer el recorrido del árbol se consume mucho tiempo. Este modelo es utilizado para resolver el problema de etiquetado de imágenes. En la Figura 3.2 se puede observar un ejemplo de un PEDT para representar una imagen.



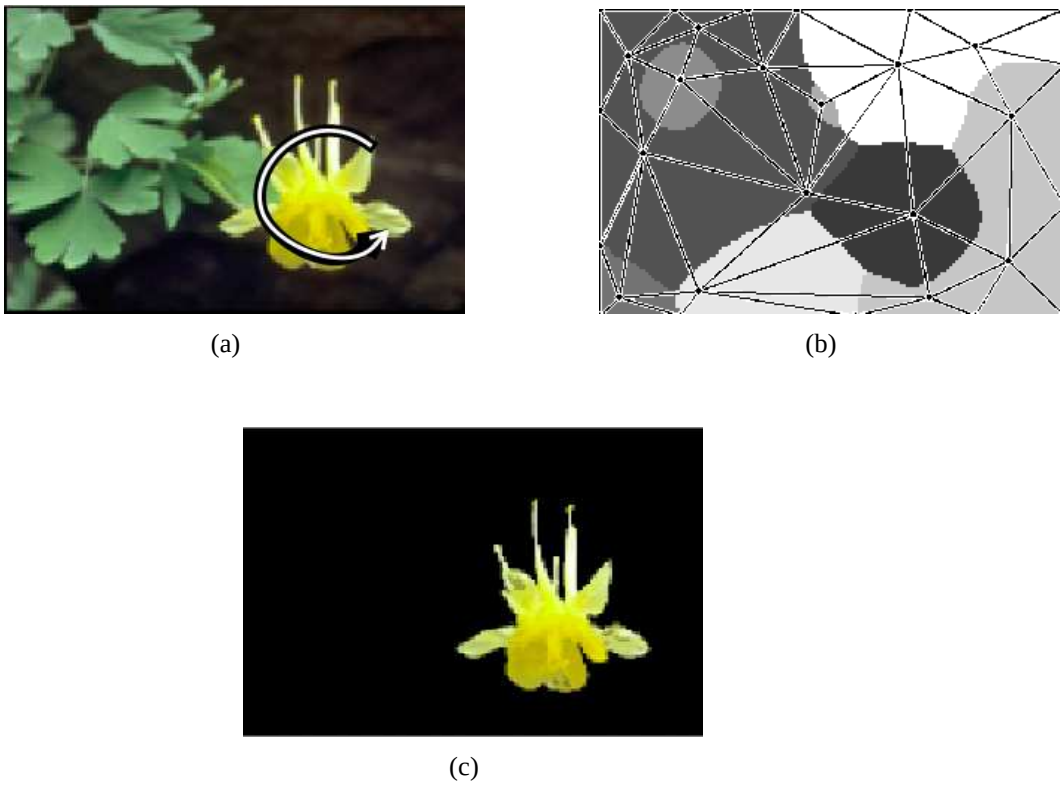
(a) (b)

(c) (d)

Figura 3.2: Ejemplo de un PEDT para representar una imagen. (a) Imagen original. (b) Árbol dinámico posicional etiquetado. (c) Una proyeccion de una parte del PEDT. (d) Representación de los nodos del PEDT en la imagen. Tomada de (Storkey y Williams, 2003).

Tan y Ahuja en (Tan y Ahuja, 2001) definen una estructura que está basada en utilizar triángulos para descomponer la imagen, cada triángulo es una región que representa a la imagen. Tomando en cuenta los vértices y bordes de los triángulos se busca encontrar regiones de la imagen que sean adyacentes y realizar la segmentación

de la imagen. Cuando se realiza esta segmentación se obtiene representación aproximada para la extensión espacial de los objetos en la imagen original. El problema de utilizar este tipo de representación es que al utilizar triángulos para representar la imagen, perdemos la forma del objeto que está en el triángulo; otro problema es que no tenemos información espacial importante de los objetos en la imagen, sólo sabemos si son adyacentes, pero no sabemos la posición de los objetos. Este trabajo se centra en resolver el problema de segmentación de imágenes. En la Figura 3.3 podemos observar un ejemplo de la representación de una imagen utilizando el método de Tan y Ahuja.



(a) (b)

(c)

Figura 3.3: Ejemplo de la representación de una imagen con base en un conjunto de triángulos que contienen a los diferentes objetos de una imagen. (a) Imagen original. (b) Representación de la imagen utilizando triángulos. (c) Imagen con el objeto de interés segmentada. Tomada de (Tan y Ahuja, 2001)

Park y Aggarwal en (Park y Aggarwal, 2004) presentan un método para reconocer la interacción de dos personas usando una red bayesiana jerárquica. El nodo raíz de la red bayesiana representa la región completa del cuerpo, y los nodos hijos representan la cabeza, parte superior del cuerpo, parte inferior del cuerpo, respectivamente. Cada parte del cuerpo es subdividida en partes que contienen piel y partes que no contienen piel. Por tanto, se propone un modelo gráfico estocástico para reconocer la interacción entre

dos personas. En este trabajo también se integran descripciones verbales semánticas del tipo *sujeto + verbo + (objeto)*. El sujeto corresponde a la persona de interés en la imagen, el verbo a la acción de movimiento del sujeto y el objeto un objetivo opcional de movimiento (es decir, usualmente otra parte del cuerpo de una persona). Un ejemplo de de esta descripción es: *estar quieto con las manos hacia abajo*.

Una de las principales limitaciones es que no se considera el hecho de que las imágenes presenten oclusiones. Otro aspecto es que no se incorporan relaciones espaciales entre los elementos de la persona. Para llevar a cabo la detección de la persona se necesita información adicional como la distancia entre las personas involucradas en la imagen. En la Figura 3.4 podemos observar la manera en la que se representa al cuerpo de una persona y llevarse a cabo el reconocimiento de la interacción entre dos personas.

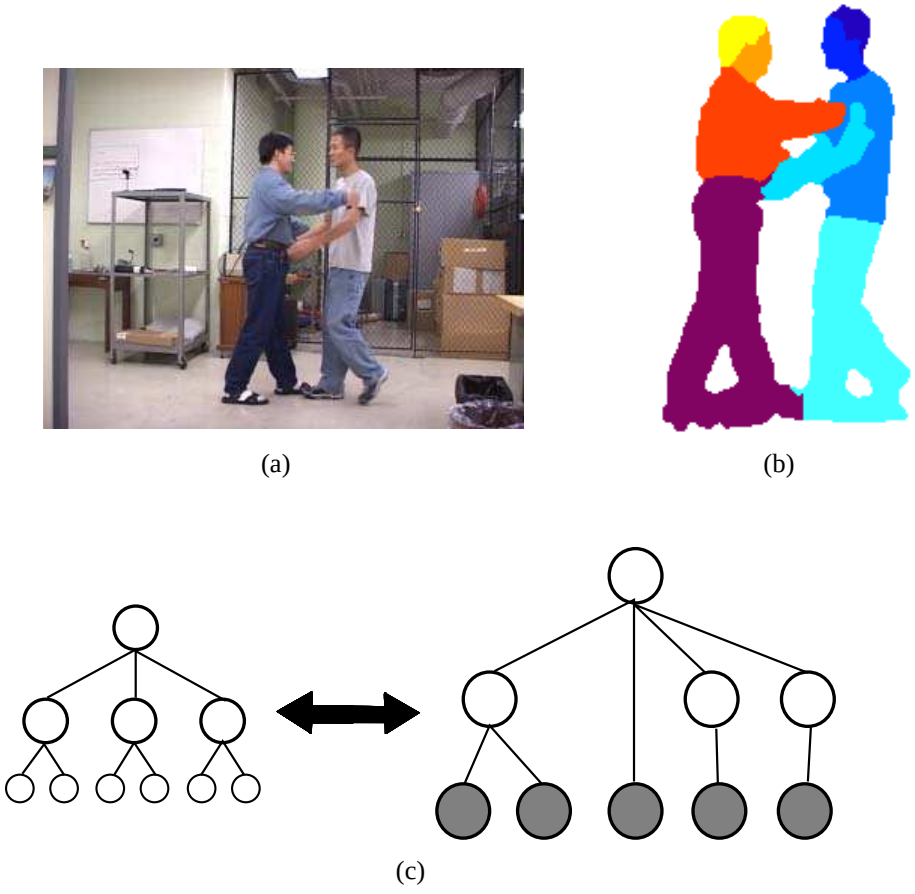


Figura 3.4: Ejemplo de una red bayesiana para representar el cuerpo de una persona. (a)Imagen original. (b)Cada parte del cuerpo es ubicada de diferentes colores. (c)Ejemplo de las redes bayesianas que utilizan para representar a una persona. La red del lado izquierdo, representa la persona completa, mientras que la red del lado derecho, es la red bayesiana usada para estimar la pose del cuerpo completo de la persona. Tomada de (Park y Aggarwal, 2004)

3.6. Conclusiones

En este capítulo se abordó de manera breve la teoría básica de las redes bayesianas. Posteriormente se describieron algunos trabajos donde se representan los objetos de interés en una imagen utilizando diferentes modelos gráficos y una representación a base de triángulos.

Para este trabajo en específico utilizamos una *red bayesiana* como modelo gráfico probabilista para representar la gramática visual que definimos. La red bayesiana nos permite hacer inferencia de una manera sencilla, además de que nos permite representar los elementos del rostro y sus relaciones de una manera natural.

La red bayesiana también nos permite incluir la incertidumbre en la detección de los elementos, es decir, nos permite considerar los problemas que pueden presentar los detectores ques se utilizaron para cada elemento de rostro.

En el siguiente capítulo se analizan algunos de los métodos utilizados en la detección de rostros.

Capítulo 4
Detección de rostros

4.1. Introducción

La detección de rostros es el primer paso en el reconocimiento automático de rostros. La buena detección tiene una gran influencia sobre el desempeño y usabilidad del sistema completo de reconocimiento de rostros. Dada una imagen o un video, un detector ideal de rostros debe ser capaz de identificar y localizar todos los rostros presentes sin tener en cuenta su posición, escala, orientación, edad y expresión. Además, la detección debe ser independiente de las condiciones de iluminación y del contenido de la imagen o del video.

La detección de rostros se puede realizar tomando en cuenta varias señales: color de piel (para rostros en imágenes o video a color), movimiento (para rostros en video), forma facial o de la cabeza, aspectos faciales que consisten en la presencia de lentes, bigote, barba en el rostro; o una combinación de estos parámetros. Los algoritmos de detección de rostros más exitosos son los basados en apariencia sin el uso de otros parámetros, en los cuales se utlizan conjuntos de entrenamiento de rostros y no-rostros para crear modelos del rostro. A continuación trataremos de manera breve algunos de los métodos más utilizados en la detección de rostros.

4.2. Métodos para la detección de rostros

Existen diversos métodos para la detección de rostros en imágenes, cada método tiene sus ventajas y desventajas (García, 2006). Se clasifican estos métodos en cuatro categorías; aunque algunos métodos pueden ser clasificados en más de una:

1.- Métodos basados en el conocimiento. Estos métodos se basan en el conocimiento

humano de las características que definen al rostro. Este conocimiento se usa para formar reglas que permitan distinguir entre rostros y no-rostros

2.- Métodos de características invariantes. Estos métodos buscan las características del rostro que son persistentes a diferentes condiciones de iluminación y pose como lo es el color de la piel y la textura.

3.- Métodos de correspondencia de plantillas. Estos métodos utilizan varios patrones o modelos del rostro que describen completa o parcialmente sus características. Para detectar un rostro se calcula la correlación entre una imagen de entrada y los patrones.

4.- Métodos basados en la apariencia. Se crean patrones o modelos del rostro a partir de un conjunto de imágenes de entrenamiento de rostros y no-rostros, tomando los valores de los píxeles.

A continuación se van a describir brevemente algunos de los métodos más sobresalientes para detectar rostros.

4.2.1. Métodos Basados en el Conocimiento

El desempeño de los métodos basados en el conocimiento depende de qué tan bien el conocimiento del investigador sobre el rostro sea codificado. Además, el problema se complica al tratar de crear reglas para las distintas poses del rostro.

Kongqiao (Kongqiao, 2003) convierte las imágenes de entrada a imágenes de mosaico para eliminar las texturas. Después extrae las características cóncavas horizontales de esa imagen que representan los ojos y las cejas de los posibles rostros. Entonces usa reglas que definen la relación entre los ojos y las cejas para determinar la presencia de posibles rostros. Algunas de estas reglas son tan simples como que las cejas están siempre por encima de los ojos. Por último, usa 3 plantillas de la distribución de los ojos, cejas y boca para eliminar la mayor cantidad de falsos positivos. El autor reporta un 89.3 % de detecciones correctas sobre un conjunto de 402 imágenes.

Ramírez et al. (Ramírez, Zanella, y Fuentes, 2003) presentan un método que se basa en dos etapas. En la primera etapa buscan los posibles rostros usando la heurística de que el promedio de la intensidad de los ojos es menor que la intensidad del puente de la nariz. En la segunda etapa evalúan cada posible rostro para determinar si es en realidad

un rostro usando dos discriminadores. El primer discriminador se basa en la heurística de que el histograma de un rostro con una iluminación uniforme tiene una forma similar a una distribución Pearson. Por último, utilizan un discriminador que se basa en la imagen de bordes del posible rostro. De esa imagen obtienen los valores promedio de siete regiones, los cuales son comparados con umbrales calculados estadísticamente. Este sistema alcanza un 87.3 % de detecciones correctas con el conjunto de prueba de BioID. El conjunto de prueba BioID consiste de 1521 imágenes en escala de grises con una resolusión de 384×286 píxeles. Cada imagen muestra la vista frontal de un rostro de 23 personas diferentes.

4.2.2. Métodos de Características Invariantes

Estos métodos se basan en encontrar las características del rostro que son invariantes al tamaño, la pose o las condiciones de iluminación. Su primer objetivo consiste en detectar las características faciales para después inferir la existencia de un rostro.

Características Faciales y Texturas

Las características faciales a buscar incluyen: ojos, boca, nariz y cejas. La principal desventaja de los métodos basados en características faciales y texturas radica en no poder encontrar las características faciales en imágenes con mucho ruido o muy degradadas, debido a que las características pueden no estar bien definidas.

Fröba y Ernst (Froba y Ernst, 2004) detectan rostros de frente extrayendo características locales en regiones de 3×3 píxeles; usando una versión modificada de la Transformada Census. Cada característica es codificada en un patrón binario que representa los bordes de la imagen. Estas características se evalúan usando un clasificador de cuatro etapas en cascada. Cada etapa es un clasificador lineal que usa una tabla de pesos calculada con un conjunto de imágenes de entrenamiento. Para entrenar se usaron 6,000 imágenes de rostros y 2,000 imágenes de no-rostros, también se utilizó la técnica de Bootstrapping para incrementar el número de ejemplos de no-rostros. Debido a que las características se evalúan localmente, no se ven afectadas drásticamente por una iluminación no uniforme. La detección de un posible rostro consiste en explorar una imagen usando pequeñas ventanas de 22×22 píxeles, para calcular las características asociadas a cada etapa. Después, se combinan linealmente usando los pesos de las tablas. Los autores reportan un 95.75 % de detecciones correctas sobre el conjunto de

entrenamiento de BioID.

Hamouz et al. (Hamouz et al., 2004) presentan un método para detectar y localizar rostros de frente. Primero buscan de forma independiente 10 características del rostro como las esquinas y centro de los ojos, nariz y boca usando filtros de Gabor en diferentes escalas. Las configuraciones de estas características son clasificadas usando un clasificador bayesiano que supone una densidad de probabilidad de modelo de mezcla Gausiana. Por último se usa un clasificador de textura de la piel para eliminar posibles falsos positivos. Los autores logran alcanzar un 91.3 % de detecciones correctas sobre el conjunto de prueba de BioID.

Color de la piel

A pesar de que en apariencia, las personas tienen diferente color de piel, varios estudios han mostrado que la principal diferencia está en la intensidad y no en el color mismo. Así, diversos métodos para detectar rostros se basan en el color de la piel usando diferentes espacios de color, incluyendo: RGB, HSV y YCbCr. Tienen la ventaja de no depender de la pose del rostro para poder detectarlo, así que pueden detectar rostros de frente o de perfil y rotados sobre el plano de la imagen. A pesar de funcionar bien en diferentes condiciones de iluminación, tienen la desventaja de no funcionar bien cuando hay variaciones en el color de la fuente de iluminación. También pueden no distinguir bien entre los verdaderos rostros y regiones de la imagen que contienen colores similares a los del rostro, como la ropa.

Kovac et al. (Kovac, Peer, y Solina, 2003), (Peer, Kovac, y Solina) presentan un método para detectar rostros basado en el color de la piel. Primero normalizan las diferencias de iluminación de la imagen usando métodos de compensación de color. Después segmentan la imagen buscando los píxeles del color de la piel usando su propio modelo del color de la piel, el cual está basado en reglas. Por último, utilizan algunas heurísticas para seleccionar las regiones que corresponden a los rostros.

Chai y Bouzerdoum (Chai y Bouzerdoum, 2000) utilizan reglas bayesianas de decisión para clasificar los píxeles de las imágenes. Su sistema trabaja sobre el espacio de color YCbCr, el cual tiene la ventaja de separar la imagen en su luminosidad y sus componentes de color. Los autores hicieron pruebas con imágenes de personas de diferentes razas obteniendo un 2.47 % de clasificaciones incorrectas de los píxeles de la piel. Kuchi et al. también trabajan con el espacio de color YCbCr y clasificadores bayesianos; además, su sistema puede rastrear un rostro en una secuencia de imágenes

en tiempo real. Los autores reportan un 82 % de detecciones correctas de rostros.

Seow et al. (Seow, Valaparla, y Asari, 2003) usan redes neuronales para generar un modelo del color de la piel y para detectar rostros en el espacio de color RGB. Reunieron un conjunto de 41,000 píxeles del color de la piel y entrenaron una red neuronal de tres capas con tres entradas correspondientes a cada color del espacio RGB. Después de segmentar la imagen de entrada por color, se evalúa cada región resultante dividiéndola en nueve partes, en donde con ayuda de nueve redes neuronales determinan la presencia o ausencia de un rostro.

4.2.3. Métodos de Correspondencia de Plantillas

Estos métodos se basan en una plantilla del rostro previamente definida de forma manual por un experto. Para verificar la existencia de un rostro se mide la correlación del posible rostro respecto a la plantilla.

Plantillas Predefinidas

En Jesorsky et al. (Jesorsky, Kirchberg, y Frischholz, 2001) se propone un método en el que se trabaja con los bordes de la imagen y la distancia Hausdorff. Se basa en una detección aproximada de los posibles rostros para luego usar una detección refinada de los rostros. En la detección aproximada se evalúa la similitud entre un modelo del rostro definido manualmente y los bordes del posible rostro. En la detección refinada se usa un modelo de la región de los ojos definida manualmente para tener un ajuste preciso del rostro. Los autores reportan un 91.8 % detecciones correctas con el conjunto de BioID. Algunos modelos utilizados fueron optimizados usando algoritmos genéticos, mejorando sus resultados para alcanzar un 92.8 % de detecciones correctas con un mejor ajuste del rostro.

Plantillas Deformables

Yuille et al. (Yuille, Hallinan, y Cohen, 1992) utilizaron plantillas deformables para detectar las características del rostro como los ojos, boca y nariz. Las características son descritas como plantillas parametrizadas. Una función de energía relaciona la plantilla y los bordes de las características a buscar. Para detectar y ajustar una plantilla a las características, la plantilla va alterando sus parámetros; lo cual hace que se deforme y se ajuste mejor a los bordes de las características mientras se minimiza la función de energía. Los parámetros finales se usan como un descriptor de las características

encontradas. Una de las desventajas de este método es que funciona bien siempre y cuando la plantilla se inicialice cerca de la característica a buscar.

Stephen et al. (Stephen et al., 2004) usan plantillas deformables para detectar y alinear un rostro en imágenes con iluminación variable. Para minimizar las diferencias de iluminación en el rostro, usan distintos filtros que trabajan localmente y así normalizar la intensidad de la imagen y detectar bordes.

4.2.4. Métodos basados en la Apariencia

A diferencia de los métodos de correspondencia de plantillas, en los cuales los patrones o modelos son predefinidos por expertos, los modelos en los métodos basados en la apariencia son definidos a partir de un conjunto de imágenes de entrenamiento de rostros y no-rostros. Utilizan análisis estadístico o algoritmos de aprendizaje automático.

Redes Neuronales

Las redes neuronales se han usando en muchos problemas relacionados con el reconocimiento de objetos y han mostrado ser una herramienta útil para la detección de rostros. Uno de los trabajos más significativos para detectar rostros usando redes neuronales es el de Rowley et al. (Rowley, Baluja, y Kanade, 1998a).

El sistema desarrollado por Rowley et al. (Rowley, Baluja, y Kanade, 1998a) se basa en dos etapas. En la primera etapa, se realiza una corrección y normalización de la iluminación, después se usan varias redes neuronales especializadas en diferentes regiones del rostro. En la segunda etapa se combinan las detecciones de la primer etapa usando heurísticas y otras redes neuronales, de esta manera se eliminan falsos rostros. El conjunto de ejemplos de rostros fue creado a partir de 1050 imágenes de rostros que fueron normalizadas. De cada imagen del conjunto de rostros fueron generadas 15 variaciones por medio de: rotar en forma aleatoria hasta 10^o , escalar entre el 90 % y 110 % el tamaño y trasladar horizontalmente hasta un 5 %. Para generar el conjunto de no-rostros fue usada la técnica de *Bootstraping*, la cual consiste en usar el conjunto de rostros y generar un conjunto inicial aleatorio de no-rostros para entrenar la red, después se usa la red entrenada para buscar rostros en imágenes que no contienen rostros. Todas las regiones que la red clasificó incorrectamente como rostros son agregadas al conjunto de entrenamiento y se entrena nuevamente a la red, para realizar el mismo

procedimiento con otras imágenes que no contienen rostros. Los autores reportan hasta un 90.5 % de detecciones correctas con un conjunto de prueba de 130 imágenes con 507 rostros. Rowley et al. (Rowley, Baluja, y Kanade, 1998b) extendieron su sistema para que también se detectaran rostros rotados sobre el plano de la imagen. Para ello usan una red neuronal que predice el ángulo de rotación del posible rostro y entonces se rota al posible rostro en sentido contrario para que tenga un ángulo de 0^o . Sin embargo, el porcentaje de detecciones correctas de rostros sin rotar bajó a un 76.9 % debido al error en la predicción del ángulo.

Otro método para la detección de rostros basado en redes neuronales es el de Garcia y Delakis (Garcia y Delakis, 2002), este método utiliza una arquitectura de red neuronal de convolución. El método que se presenta deduce automáticamente el filtro de convolución óptimo que actúa como extractor de características. La red neuronal fue entrenada con 12,976 imágenes de rostros sin normalizar su intensidad. El porcentaje de detecciones correctas de este método con un subconjunto de 104 rostros proveniente del conjunto de CMU+MIT (Rowley, Baluja, y Kanade, 1998a) fue del 98 % con un falso positivo.

Clasificador simple de Bayes

Schneiderman y Kanade (Schneiderman y Kanade, 1998) presentan un método para detectar rostros de frente usando un clasificador simple de Bayes y análisis de componentes principales (PCA). Su sistema calcula la probabilidad de la apariencia y posición de los patrones del rostro. También realizan una normalización de la intensidad de cada región. Reportaron un 92.5 % de detecciones correctas con el conjunto de CMU+MIT. También muestran cómo con algunas variaciones en su sistema pueden detectar rostros de perfil. Schneiderman y Kanade (Schneiderman y Kanade, 2000) utilizan características wavelet para detectar rostros y automóviles.

En cuanto a los rostros, su sistema puede detectar tanto rostros de frente como de perfil. Su sistema calcula histogramas que representan los coeficientes wavelet y su posición; estos histogramas son clasificados con reglas de decisión basadas en la estadística del conjunto de entrenamiento. Utilizan clasificadores especializados en las distintas poses u orientaciones de los objetos a detectar. Reportan un 90.2 % de detecciones correctas con el conjunto de CMU+MIT y un 92.8 % de detecciones correctas con el conjunto de rostros de perfil del CMU. Schneiderman (Schneiderman, 2004) presenta en otro trabajo un método general para detectar objetos. Se basa en las características wavelet y usa una arquitectura de cascada. Cada elemento de la cascada

consiste de un clasificador semi-simple Bayes. Además, realiza una corrección de la iluminación de cada región de la imagen. En cuanto a la detección de rostros de frente alcanza un 95.7 % de detecciones correctas con el conjunto de prueba de CMU+MIT.

4.2.5. AdaBoost

De manera general, el algoritmo *Boosting* es un método de clasificación que combina varios clasificadores básicos o débiles para al final formar un clasificador más complejo o fuerte que sea más preciso. Una de sus variantes más conocidas es el algoritmo *AdaBoost*.

AdaBoost es un método de aprendizaje iterativo que combina una serie de clasificadores base usando una combinación lineal para clasificar nuevos ejemplos. En cada iteración se genera un nuevo clasificador el cual trata de minimizar el error esperado asignándole más peso a los ejemplos que fueron mal clasificados, aumentando la probabilidad de que sean clasificados correctamente.

El algortimo AdaBoost, propuesto por Freud y Shapire en 1995 (Schapire, 2002), resuelve muchas de las dificultades prácticas de los algoritmos boosting iniciales. El algoritmo toma como entrada un conjunto de entrenamiento $(x_1, y_1), ..., (x_m, y_m)$ donde cada x_i pertenece a algún *dominio* o *ejemplos en el espacio X*, y cada etiqueta y_i es alguna etiqueta del conjunto *Y*. AdaBoost (Schapire, 2002) llama un *algoritmo de aprendizaje base o débil* dado, en una serie de etapas *t = 1, ..., T*. Una de las ideas principales del algoritmo es la de mantener una distribución o un conjunto de pesos sobre el conjunto de entrenamiento. El peso de esta distribución sobre el ejemplo de entrenamiento *i* en la etapa *t* es denotado $D_t(i)$. Inicialmente todos los pesos son iguales, pero en cada etapa los pesos de los ejemplos incorrectamente clasificados son incrementados de modo que el algoritmo base es forzado a enfocarse a los ejemplos difíciles en el conjunto de entrenamiento.

El trabajo de los algoritmos base es encontrar un *clasificador base* $h_t : X \to R$ apropiado para la distribución D_t. En el caso más simple, el rango de cada h_i es binario, es decir, está restringido de $\{$-1, +1$\}$; el trabajo del algoritmo base es minimizar el *error* ϵ_t.

$$\epsilon_t = Pr_{i \sim D_t}[h_t(x_i) \neq y_i].$$

Una vez que el clasificador base h_t ha sido recibido, AdaBoost escoge un parámetro $\alpha \in R$ que intuitivamente mide la importancia que es asignada a h_t. El *clasificador final*

o *combinado H* es el clasificador con mayor peso de los *T* clasificadores base donde α_t es el peso asignado a h_t.

En el algoritmo 1 se muestra el algoritmo general de *AdaBoost*

Algoritmo 1 AdaBoost

Entrada: $(x_1, y_1), ..., (x_m, y_m)$ donde cada $x_i \in X, y_i \in Y = \{-1, +1\}$
Salida: Clasificador final:

$$H(x) = sign(\sum_{t=1}^{T} \alpha_t h_t(x))$$

1: Inicializar $D_1 = 1/m$
2: **para** $t = 1, ..., T$: **hacer**
3: Entrenar el clasificador base usando la distribución D_t.
4: Obtener el clasificador base $h_t : X \rightarrow R$.
5: Elegir $\alpha \in R$.
6: Actualizar:

$$D_{t+1}(i) = \frac{D_t(i) exp(-\alpha_t y_i h_t(x_i))}{Z_t}$$

7: donde Z_t es un factor de normalización (se elige de manera que D_{t+1} sea una distribución).
8: **fin para**

4.3. Conclusiones

Los trabajos antes escritos, en general, presentan buenos resultados a la hora de detectar rostros en imágenes. Los resultados reportados están por arriba del 90 % de detecciones correctas. Aparentemente esto es un buen resultado, pero en realidad estos métodos tienen el problema de que no detectan la totalidad de los rostros o por el contrario, detectan rostros cuando en realidad no los hay.

La mayoría de los métodos analizados trabaja con imágenes en condiciones controladas, es decir, las personas están de frente o no hay oclusión del rostro por algún objeto. Aunque en algunos casos las imágenes presentan ciertos problemas de iluminación, no son de consideración. La ventaja de nuestro método es que considera características que con mucha frecuencia presentan las imágenes cuando son capturadas, como es la oclusion del rostro o problemas de iluminación.

El método de detección de rostros que proponemos se basa en el empleo de varios detectores basados en el algoritmo AdaBoost. Utilizamos un detector por cada elemento

del rostro que se definió. Por otra parte, representamos la gramática visual que se definió mediante un red bayesiana. Esta red bayesiana incluye tanto la información de los detectores, como información de las relaciones espaciales que cumplen los elementos del rostro que se definieron.

Capítulo 5
Método propuesto

De manera general, podemos dividir la solución propuesta en tres fases:

1.- Definición de la gramática visual.

2.- Representación de la gramática visual.

3.- Detección de rostros utilizando la gramática visual definida.

En la Figura 5.1 se observa el proceso de aprendizaje de los parámetros de la red bayesiana. Los parámetros de la red bayesiana se calcularon tomando en cuenta el resultado de los detectores de cada elemento del rostro. Por otra parte, las reglas de producción y el conocimiento previo de cómo está formado un rostro sirvió para modelar la estructura de la red bayesiana.

En la Figura 5.2 se observa el esquema general para la detección de rostros en condiciones difíciles utilizando el modelo propuesto. El resultado de los detectores es la evidencia que se le pasa a la red bayesiana para llevar a cabo la propagación en la red.

5.1. Definición de la Gramática Visual

Cada rostro es diferente y por tanto, cada rostro posee rasgos diferentes, pero en general todos los rostros tienen los mismos "elementos" (ojos, nariz, boca).

Para el método que proponemos, el primer paso para definir una gramática visual es *definirla de manera conceptual*, es decir, debemos establecer cuáles van a ser los elementos que vamos a tomar en cuenta para representar a los objetos además de establecer las relaciones que existen entre estos elementos.

Para nuestro caso los elementos del rostro que consideramos son:

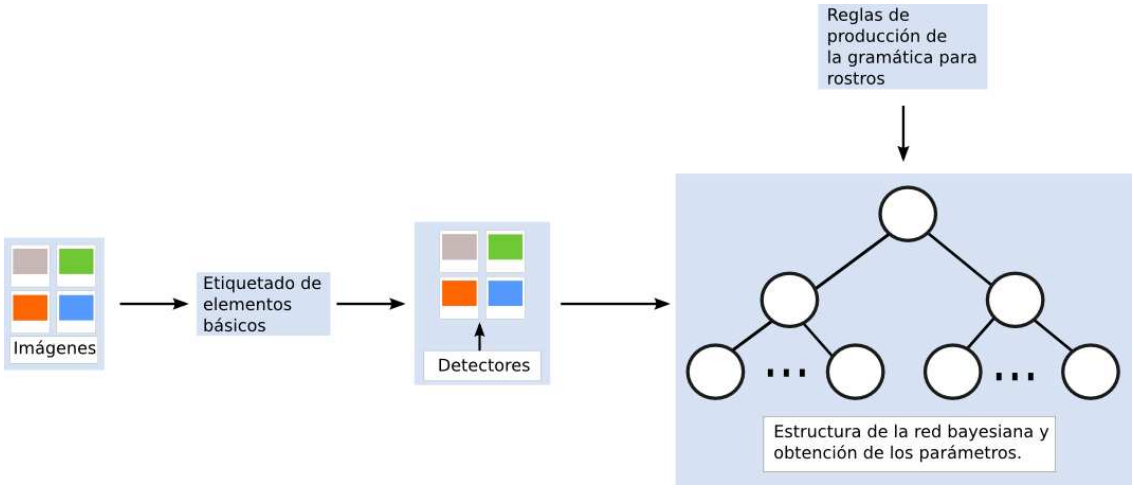


Figura 5.1: Esquema general para la obtención de los parámetros y la estructura de la red bayesiana. Basándonos en los detectores de cada elemento del rostro y las reglas de producción de la gramática del rostro se obtiene la red bayesiana que nos va a servir para la detección de rostros en condiciones difíciles.

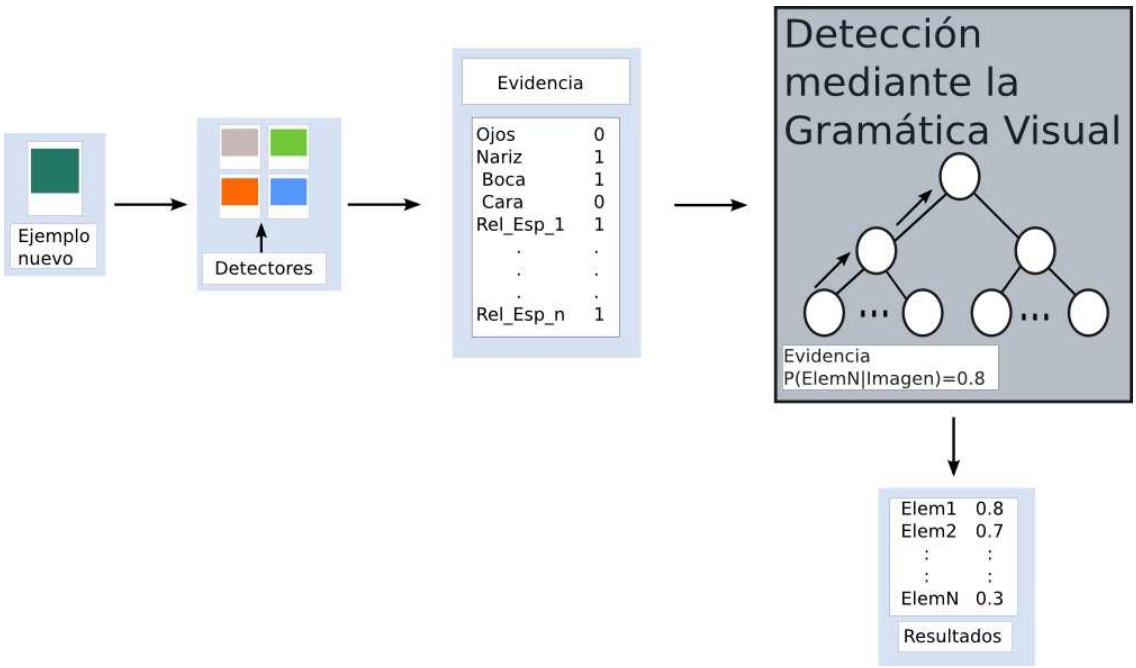


Figura 5.2: Esquema general para la detección de rostros utilizando la red bayesiana que representa la gramática definida. Se aplican los detectores de los elementos del rostro al ejemplo nuevo, el resultado de la detección de cada elemento es la evidencia de la red bayesiana, se realiza el proceso de inferencia en el modelo y se obtiene el resultado final de la probabilidad de que la imagen sea un rostro.

- Ojos.
- Nariz.
- Boca.
- Cara.

Para poder mejorar un primer modelo que también fue propuesto por nosotros, se integraron relaciones espaciales que pudieran ayudar en la detección del rostro.

A continuación se describen de manera general los tipos de relaciones espaciales, porteriormente se definen las relaciones espaciales que fueron consideradas para integrarse a la gramática visual que se definió para representar al rostro.

5.2. Relaciones espaciales

Se entiende por relaciones espaciales aquellas relaciones que se pueden determinar de un objeto con respecto a otro (conocido como objeto de referencia), y que nos son de utilidad para conocer información acerca de su posición relativa en una escena. Comúnmente, las relaciones espaciales se establecen de manera binaria; sin embargo, ciertas relaciones como: *entre, rodeado por, más cercano a*, entre otras, son mejor entendidas cuando se definen con respecto a más de un objeto de referencia.

Las relaciones espaciales más comúnmente utilizadas son:

1.- *Relaciones topológicas*. Estas relaciones se caracterizan por la propiedad de que se preservan ante transformaciones topológicas como traslación, rotación y escalamiento. Algunos ejemplos de este tipo de relaciones son: *dentro de, traslapado con* y *vecino de*.

2.- *Relaciones de orden*. Este tipo de relaciones es variable ante transformaciones de rotación, pero se preserva ante transformaciones de escalamiento y traslación. Estas relaciones generalmente se presentan en forma dual como el caso de: *delante - detrás, debajo - arriba*.

3.- *Relaciones métricas*. Estas relaciones hacen uso de medidas tales como distancia y dirección. Este tipo de relaciones son afectadas por transformaciones de escalamiento, pero no son afectadas por transformaciones de rotación o traslación. Un ejemplo de este tipo de relación es: *a 10 kilómetros de distancia*.

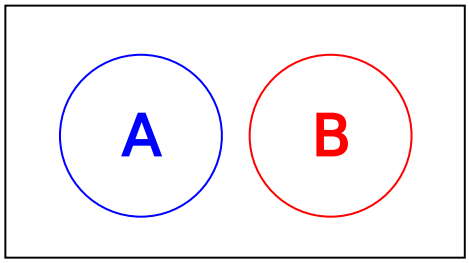


Figura 5.3: Las regiones A y B no se tocan.

También existen *relaciones difusas*, por ejemplo, *cerca, lejos*, pero no son consideradas debido a que su medida no es precisa y dependen básicamente de criterios subjetivos, por lo que conceptos como *cerca* o *lejos* pueden tener valores completamente diferentes dependiendo del problema al que sean aplicados, la persona que define los criterios, etc.

Es importante mencionar que en el caso de las relaciones espaciales en imágenes, algunas no son aplicables, donde hablar de que un objeto está oculto *detrás de* otro no tiene sentido, debido a que el objeto oculto no podría ser apreciado en la imagen y dicha relación sería prácticamente imposible de comprobar. A continuación se describirá de manera breve cada una de las relaciones espaciales antes mencionadas.

Relaciones topológicas

Este tipo de relaciones existen entre dos objetos y se mantienen de forma constante aún cuando se aplican transformaciones de isometría a los objetos de interés. La presencia de dichas relaciones entre un par de objetos en una imagen representa información de gran relevancia, ya que al no ser afectadas por tales transformaciones, será posible comparar un par de imágenes sin que factores como ruido, ligeras variaciones en la posición de la cámara y distancia de los objetos con respecto al sensor sean determinantes en el resultado.

Gracias a su característica de invariabilidad, estas relaciones son las más frecuentemente modeladas para su uso en diversos dominios. Para el caso del uso en imágenes, es de interés trabajar con relaciones entre dos objetos, en un espacio bidimensional (la imagen). Las relaciones de este tipo para dos regiones A y B en una imagen, son:

- *Disjunto*. Ni A ni B se tocan. Ver Figura 5.3.
- *Contiene a*. El objeto A tiene dentro de sí al objeto B en su totalidad sin que algún

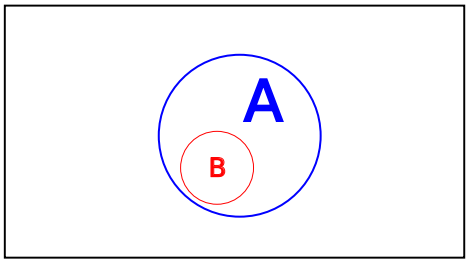


Figura 5.4: La región A contiene a la región B.

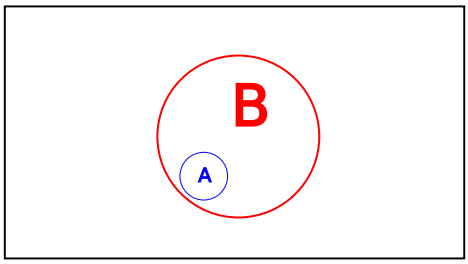


Figura 5.5: Las región B contiene a la región A.

borde de *A* toque algún borde de *B*. Ver Figura 5.4.

- *Dentro de*. Es el dual de *contiene a*, sólo que en este caso es *A* el que se encuentra dentro de *B*, los bordes tampoco se tocan. Ver Figura 5.5.
- *Tocando*. Únicamente los bordes de *A* y *B* se tocan. Ver Figura 5.6.
- *Cubre*. Similar a *contiene a*, pero en esta relación los bordes de *A* y *B* sí se tocan. Ver Figura 5.7.
- *Cubierto por*. Similar a *dentro de*, pero en esta relación los bordes de *A* y *B* sí se tocan. Ver Figura 5.8.

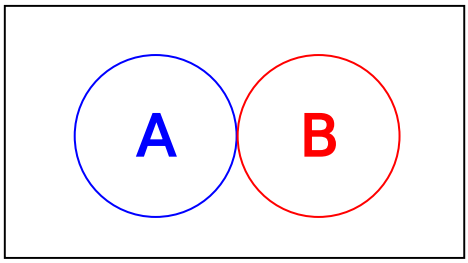


Figura 5.6: Los bordes de las regiones A y B se tocan.

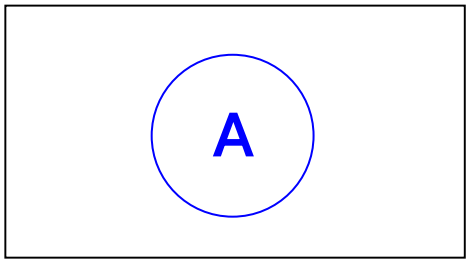


Figura 5.7: En este ejemplo la región A es la cubre a la región B.

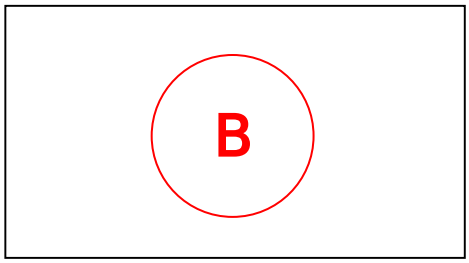


Figura 5.8: En este ejemplo la región A es cubierta por la región B.

- *Traslapado*. Los objetos *A* y *B* se intersectan, sin que ninguno de ellos contenga al otro. Ver Figura 5.9.
- *Iguales*. Los objetos *A* y *B* se encuentran ubicados exactamente en la misma posición. Ver Figura 5.10.

Relaciones de orden

Las relaciones de orden son variables ante transformaciones de rotación, pero se preserva ante transformaciones de escalamiento y traslación.

Estas relaciones se dividen en dos grupos que son las *relaciones de orden parcial* y *las relaciones de orden estricto*.

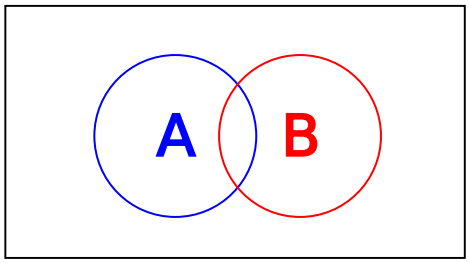


Figura 5.9: En este ejemplo las regiones A y B están traslapadas.

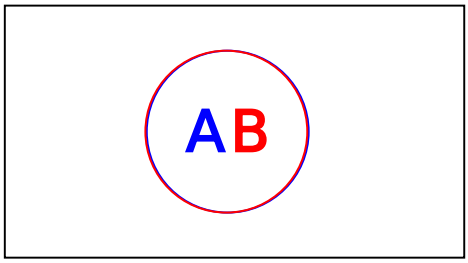


Figura 5.10: Las regiones A y B son exactamente iguales.

1.- *Relaciones de orden parcial*.
Se dice que una relación *R* entre los elementos $a, b, c \in S$, es de orden parcial si cumple con las siguientes propiedades:

- *Antisimétrica*. Si aRb y bRa, entonces $a = b$. O bien, si existe la relación aRb y $a \neq b$, entonces bRa no se cumple.
- *Transitiva*. Si aRb y bRc, entonces se cumple aRc.
- *Reflexiva*. $\forall a \in S, aRa$ se cumple.

2.- *Relaciones de orden estricto*.
Toda relación *R* entre los elementos $a, b, c \in S$, es de orden estricto si cumple con las siguientes propiedades:

- *Antisimétrica*. Si aRb y bRa, entonces $a = b$. O bien, si existe la relación aRb y $a \neq b$, entonces bRa no se cumple.
- *Transitiva*. Si aRb y bRc, entonces se cumple aRc.
- *Irreflexiva*. aRa no se cumple para ninguna $a \in S$.

La posición y el orden de los objetos en una imagen proporcionan información valiosa respecto a cómo estos interactúan en la escena, si sabemos qué relación o relaciones de orden existen entre un par de objetos podemos determinar características como el grado de semejanza entre dos imágenes que contienen mismos objetos.

Relaciones métricas

Las relaciones métricas especifican qué tan lejos se encuentra un objeto del objeto de referencia, medido en términos específicos del dominio en que se aplica.

Relaciones como: *a 100 metros de, diez pasos a la derecha de, 1 kilómetro a la redonda,* son ejemplos de ellas. Las relaciones métricas toman sentido cuando se cuenta con un punto de referencia conocido con el que podamos realizar una medición.

Relaciones consideradas

Dentro de la Gramática de rostro se consideraron relaciones topológicas y de orden estricto:

- Topológicas: "$dentro-de$".
- Orden: "$arriba-de$".

De manera particular, las relaciones espaciales que se definieron para la gramática del rostro son:

- *ojos ARRIBA DE boca*
- *nariz ARRIBA DE boca*
- *ojos DENTRO DE cara*
- *nariz DENTRO DE cara*
- *boca DENTRO DE cara*

5.3. Gramática para rostros

Es importante mencionar que tanto los elementos del rostro, como las relaciones espaciales que se establecieron fueron tomados en cuenta considerando tanto el conocimiento intuitivo como los detectores de los elementos básicos disponibles.

Utilizando el formalismo para definir una *gramática simbólica relacional* se definió la siguente gramática para la representación del rostro:

$$GR = (\{ROSTRO\}, \{ojos, nariz, boca, cara\}, \{arriba\,de, dentro\,de\}, ROSTRO, P)$$

donde:

- $ROSTRO$ es un símbolo no terminal que representa el *rostro completo* y también es el símbolo inicial de la gramática.

- $\{ojos, nariz, boca, cara\}$ son símbolos terminales para representar a los elementos correspondientes.

- $\{arriba\ de, dentro\ de\}$ son los símbolos relacionales que representan las relaciones espaciales.

- P son las reglas de producción s-item, definidas como:
  1: $ROSTRO^0 \rightarrow < ROSTRO, \phi >$
  2: $ROSTRO^0 \rightarrow < ojos, \phi >$
  3: $ROSTRO^0 \rightarrow < nariz, \phi >$
  4: $ROSTRO^0 \rightarrow < boca, \phi >$
  5: $ROSTRO^0 \rightarrow < cara, \phi >$
  6: $ROSTRO^0 \rightarrow < \{ojos, boca\}, arriba\ de\ (ojos, boca) >$
  7: $ROSTRO^0 \rightarrow < \{nariz, boca\}, arriba\ de\ (nariz, boca) >$
  8: $ROSTRO^0 \rightarrow < \{ojos, cara\}, dentro\ de\ (ojos, cara) >$
  9: $ROSTRO^0 \rightarrow < \{nariz, cara\}, dentro\ de\ (nariz, cara) >$
  10: $ROSTRO^0 \rightarrow < \{boca, cara\}, dentro\ de\ (boca, cara) >$

El elemento *R* que contienen las r-producciones no son consideradas debido a lo restringuido del modelo que definimos para representar al rostro.

Para poder considerar el elemento *R* en nuestra gramática de rostro, deberiamos incluir otros elementos no terminales que estuvieran relacionados con algún otro elemento no terminal o con algún elemento terminal.

Por ejemplo, si tuvieramos una relación del estilo

$$11 : OJOS^0 \rightarrow < \{OJO\_IZQ, OJO\_DER\}, \phi >$$

la producción en el conjunto *R* podría expresarse como sigue:

$$arriba\ de\ (OJOS, boca) \rightarrow [11]\{arriba\ de(OJO\_IZQ, boca), arriba\ de(OJO\_DER, boca)\}$$

Lo que nos indica esta relación, es que si sustituimos la producción 11, en la relación *OJOS arriba de boca*, tenemos de manera explícita, las relaciones que indican que tanto el *ojo izquierdo* como el *ojo derecho* están *arriba de la boca*.

Podemos decir que las *r-producciones* expresan de manera explícita las relaciones entre los elementos terminales o no terminales de la gramática.

5.4. Representación de la Gramática Visual

Los detectores que utilizamos para encontrar los elementos terminales no son confiables y en general, hay mucha incertidumbre tanto en la detección de los símbolos terminales de la gramática, como en las relaciones espaciales en las imágenes. Por lo tanto, necesitamos una representación que pueda integrar la incertidumbre existente, y una red bayesiana cumple estas características.

Con una **Red bayesiana** podemos representar las dependencias (relaciones espaciales) entre atributos (elementos de la imagen).

A continuación se describen los modelos que se propusieron para representar la gramática que se definió.

Para transformar la gramática de rostro a una primera representación utlizando una red bayesiana tomamos las siguientes consideraciones:

- Cada símbolo (terminal o no terminal) es representado como una variable (nodo) binaria.
- Cada relación es representada como una variable (nodo) binaria.
- Las producciones son representadas como un conjunto de relaciones de dependencia (arcos) en la red bayesiana. Un arco es incorporado entre los nodos que representan un símbolo no terminal y un símbolo terminal.

Además, incorporamos *nodos virtuales* que representan la información que obtenemos de los detectores para los símbolos terminales. Estos nodos representan la incertidumbre inherente a todos los detctores en una tabla de probabilidad condicional (TPC), y son incorporados al modelo agregando un arco del nodo que representa un *símbolo en la gramática* al nodo que representa un *detector en el sistema*. Por ejemplo, tenemos un nodo que representa el símbolo *nariz* con un arco al nodo que representa el resultado del detector de *nariz*; que podría ser una nariz “real” o una detección falsa.

En la Figura 5.11 se observa la red bayesiana que obtenemos para la primer representación de la gramática de rostro tomando las consideraciones anteriores. El nodo raíz representa el símbolo no terminal *Rostro* como la unión de los diferentes elementos que definen al rostro. En el siguiente nivel se observan los nodos que representan los símbolos terminales del rostro (*Boca, Nariz, Ojos, Cara*), estos nodos representan el grado de probabiliad que tiene cada elemento de aparecer en una imagen dada. En el último

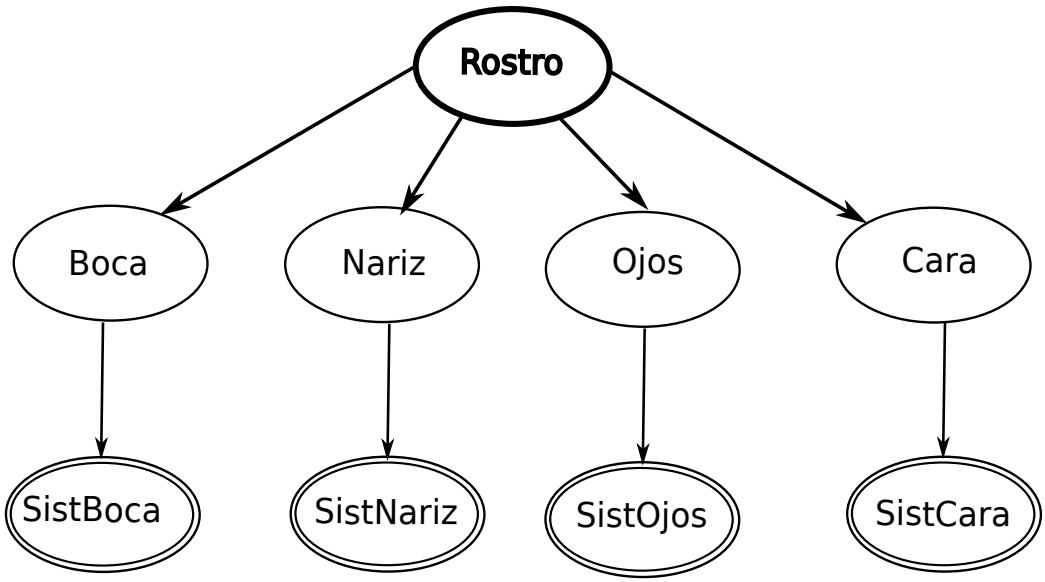


Figura 5.11: Red bayesiana sin relaciones espaciales que representa la gramática del rostro. Los óvalos con línea simple representan los símbolos no terminales y los óvalos con doble línea representan las variables de los detectores.

nivel, los nodos representan los detectores para cada elemento del rostro, estos nodos representan las probabilidades de detección de los elementos definidos, considerando el funcionamiento de los detectores que se utilizaron.

En el segundo modelo que se definió se incorporaron las relaciones espaciales entre ciertos elementos del rostro. Las relaciones espaciales definidas son:

- *ojos ARRIBA boca*
- *nariz ARRIBA boca*
- *ojos DENTRO cara*
- *nariz DENTRO cara*
- *boca DENTRO cara*

Para esta segunda representación, las relaciones espaciales son representadas como una variable (nodo) binaría. Se agregaron arcos entre los nodos de los símbolos no terminales en la relación y el nodo que representa la relación espacial.

El modelo que integra las relaciones espaciales antes mencionadas se puede observar en la Figura 5.12.

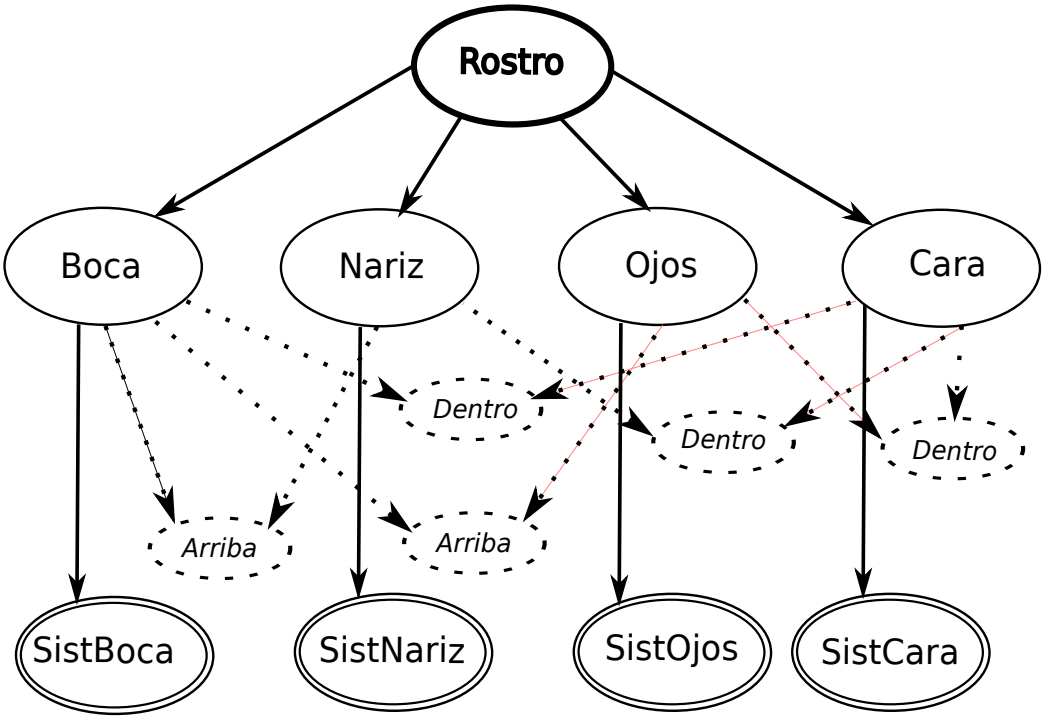


Figura 5.12: Red bayesiana con relaciones espaciales que representa la gramática del rostro. En este modelo se integran las relaciones espaciales entre ciertos elementos del rostro, estas relaciones están representadas con nodos con líneas punteadas.

5.5. Obtención de los parámetros del modelo

Una vez que la estructura de la red bayesiana para la gramática de rostros es definida, tenemos que obtener los parámetros correspondientes, es decir, las TPC para cada variable de la red bayesiana. Se aprendieron estas probabilidades a partir de datos (imágenes de rostros) usando aprendizaje parámetrico estándar para redes bayesianas (Neapolitan, 2004), tomando en cuenta la estructura que se eligió y porque deseamos saber que objeto es más probable que se esté reconociendo dado un conjunto de elementos básicos de la imagen.

Para esto consideramos un conjunto de imágenes como ejemplos positivos y negativos de rostros. Los ejemplos positivos fueron etiquetados manualmente indicando los símbolos terminales en cada imagen. Después se aplicaron los detectores para cada elemento terminal y contamos todas las detecciones correctas e incorrectas. Para reconocer cada elemento del rostro se van a utilizar reconocedores de objetos que se basan en el algoritmo AdaBoost (Viola y Jones, 2004).

Basados en estas estadísticas, obtuvimos las TPC usando un estimador de máxima

verosimilitud. Las imágenes que se utilizaron tienen las siguientes características:

- Personas en posición frontal y de perfil.
- Personas que presenten alguna oclusión por algun objeto en el rostro.
- Personas cuyo rostro no esté completo en la imagen.

Los detalles de este proceso se describen en el capítulo de experimentos.

Los parámetros de las relaciones espaciales entre los elementos fueron estimados subjetivamente, es decir, en general esperamos que las relaciones siempre se satisfagan en un rostro. Sin embargo, podría haber situaciones especiales en las que algunas relaciones no se satisfagan, como por ejemplo, si uno de los elementos de la relación está ocluido o el rostro está inclinado casi 90 grados, es por eso que asignamos probabilidades cercanos a 1 para cada relación que sea verdadera.

5.6. Detección de rostros utilizando la Gramática Visual

El *reconocimiento de objetos* se lleva a cabo realizando el proceso de *propagación de probabilidades* en la red bayesiana que representa la gramática visual.

Para este trabajo nos interesa saber la probabilidad posterior de la variable *Rostro* de los modelos que definimos en la Figura 5.11 y en la Figura 5.12. Para obtener estos valores de probabilidad debemos darle al modelo *evidencia* de la información con que se cuenta. En el caso de la gramática visual tenemos la información que nos dan los detectores de cada elemento del rostro. Si se detecta algún elemento en la imagen, se pasa como evidencia positiva al modelo, en caso contrario se indica que no se detectó el elemento deseado y también se pasa esta información al modelo.

En cuanto a la información de las relaciones espaciales entre los elementos de interés, la evidencia que se le pasa al modelo consiste en la ausencia o presencia de la relación espacial entre los elementos de una imagen dada. Es decir, después de analizar los elementos detectados en la imagen se verifica si se cumple alguna relación, si alguna relación se cumple se considera como evidencia positiva, si no se cumple se pasa como evidencia negativa. En el caso de que alguno de los elementos en la relación no se detecte en la imagen, no se da evidencia a esta relación.

Una vez asignada la evidencia, el proceso de propagación se realiza utilizando la herramienta ProBT (Kamel Mekhnacha et al.).

5.7. Conclusiones

Se ha desarrollado un método novedoso para la detección de rostros basado en una gramática simbólica relacional para rostros. Definimos la gramática para un rostro que fue transformada en una red bayesiana, cuyos parámetros fueron obtenidos a partir de datos (ejemplos positivos y negativos de rostros).

En el siguiente capítulo se describen los experimentos que se realizaron para probar los modelos propuestos.

Capítulo 6

Experimentos y resultados

En el presente capítulo se muestran los experimentos y resultados obtenidos al utilizar la gramática visual definida para la detección de rostros en imágenes. Para ello se prueba el modelo propuesto con diversas imágenes de rostros que fueron adquiridas en condiciones "difíciles" (oclusión parcial, variación en la posición, iluminación, etc).

Se realiza una comparación entre los resultados proporcionados al utilizar la gramática y los detectores de rostros proporcionados por la librería OpenCV (Yu et al., 2004), los cuales son variaciones de los detectores propuestos por Viola-Jones (Viola y Jones, 2004). También se compara el modelo sin relaciones espaciales (ver Figura 5.11) contra el modelo que incluye relaciones espaciales (ver Figura 5.12).

6.1. Características de las imágenes utilizadas

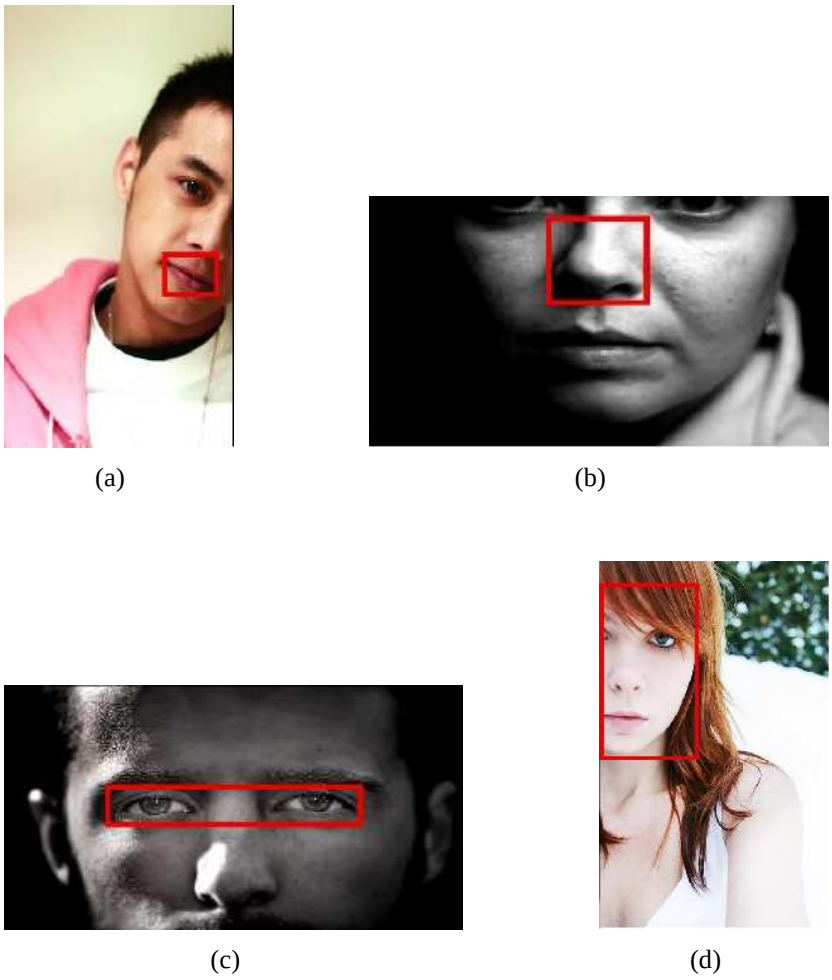
Existen varias bases de datos estándar de rostros que en general contienen imágenes en "buenas" condiciones, es decir, sin oclusión, rotaciones u otros objetos que puedan dificultar su detección. Como nuestro método esta enfocado a imágenes que fueron adquiridas en condiciones "difíciles", las bases de datos existentes no son un buen banco de prueba, por esta razón construimos nuestra propia base de datos.

Las imágenes utilizadas para obtener los parámetros de la red bayesiana y llevar a cabo los experimentos fueron descargadas de Internet utilizando búsquedas en Google Image[1]. Se buscó que estas imágenes presentaran condiciones difíciles, es decir, variación en la posición de la persona, oclusión parcial de la persona, etc., también se obtuvieron ejemplos negativos, es decir, imágenes donde no se visualizaran personas, por ejemplo, recamara, sala, cocina, etc.

[1] www.googel.com

Se descargaron un total de 200 imágenes para poder estimar los parámetros de la red bayesiana, 100 de estas imágenes contienen personas y 100 no contienen personas. Las imágenes en donde se pueden visualizar personas fueron etiquetadas ubicando los elementos básicos que constituyen un rostro (ojos, nariz, boca, cara). Después se aplicaron los detectores para cada elemento que se etiquetó, con el fin de estimar el número de verdaderos positivos y falsos positivos de cada detector.

En las Figuras 6.1 y 6.2 se puede observar algunos ejemplos de las imágenes utilizadas para la estimación de los parámetros de la red bayesiana.



(a) (b)

(c) (d)

Figura 6.1: Ejemplos de imágenes utilizadas para la obtención de los parámetros de las redes bayesianas. Los elementos del rostro fueron etiquetados de manera manual, señalando dónde se ubican los elementos terminales de la gramática. a) Boca etiquetada. b) Nariz etiquetada. c) Ojos etiquetados. d) Cara etiquetada.

Para la prueba se descargaron otras 60 imágenes, 30 de estas imágenes contienen personas, es decir son ejemplos positivos, las otras 30 imágenes no contienen personas,

(a)

(b)

(c)

Figura 6.2: Ejemplo de imágenes de entrenamiento que no contienen rostros.

es decir son ejemplos negativos. El tamaño de las imágenes es variante (entre 300×200 píxeles y 300×600 píxeles), por tal motivo, tanto el tamaño como la posición de las personas en las imágenes son variados. La posición de las personas en las imágenes son: frontal, perfil izquierdo y perfil derecho. En las Figuras 6.3 y 6.4 se pueden observar algunos ejemplos de las imágenes utilizadas en la fase de experimentos en donde se probaron los modelos propuestos.

(a)

(b)

(c)

Figura 6.3: Ejemplo de imágenes de prueba que contienen rostros en condiciones difíciles.

6.2. Descripción de los experimentos

Podemos dividir la etapa de experimentación en dos fases, en la primera fase se lograron establecer los parámetros de la red bayesiana que utilizamos para representar la gramática, y en la segunda fase se prueba el modelo propuesto, estas etapas serán detalladas a continuación.

6.2.1. Estimación de los parámetros del modelo

Para poder estimar los parámetros del modelo se etiquetaron las 100 imágenes en donde se visualiza una persona, los elementos etiquetados fueron: ojos, nariz, boca y cara. Cada uno de los elementos fue encerrado en un rectángulo y se almacenaron en un

(a)

(b)

(c)

Figura 6.4: Ejemplo de imágenes de prueba que no contienen rostros.

archivo las coordenadas de dos esquinas opuestas de cada rectángulo, en este archivo también están almacenados el número de elementos que fueron etiquetados en cada imagen. Después se calcularon los centroides de cada elemento etiquetado. Para cada una de las 200 imágenes se le aplicó los detectores de elementos y contabilizar tanto los verdaderos positivos como los falsos positivos para cada detector. Para saber si se trataba de un verdadero positivo o un falso positivo se tomó en cuenta la distancia entre los centroides del elemento etiquetado y el elemento detectado. Si la distancia entre los centroides es menor a 25 píxeles, se considera como un verdadero positivo, en caso contrario el elemento se considera como un falso positivo. Con esto se hizo una estimación de detección para cada elemento. Con esta estimación se calcularon las probabilidades condicionales para cada nodo de la red bayesiana.

En la Tabla 6.3 se muestran la tabla de probabilidad de que el *Sistema detecte una boca, dado que en la imagen haya una boca* es decir, $P(SistBoca|Boca)$. La tabla de probabilidad de que *Haya una nariz, dado que en la imagen exista un rostro,* osea

$P(Nariz|Rostro)$ se muestra en la Tabla 6.2.

Como ya se mencionó con anterioridad, las probabilidades de las relaciones espaciales se asignaron de manera subjetiva, variando los valores de que se cumpliera alguna relación espacial o de que no se cumpliera, tomando en consideración si los elementos de la relación fueron detectados. En la Tabla 6.1 se muestran las tablas de probabilidad de la relación *ojos DENTRO cara*, en ella los valores 0.6 y 0.4 son los que fueron variando, desde 0.6 hasta 0.9, para el caso de que la relación se cumpliera y de 0.4 a 0.1, para el caso en el que la relación no se cumpliera. Las tablas de probabilidad completas se pueden observar en el Apéndice A.

Tabla 6.1: Tabla de probabilidad condicional para la relación "ojos DENTRO cara"

	Cara		NoCara	
	Ojos	NoOjos	Ojos	NoOjos
RE = True	0.9	0.5	0.5	0.5
RE = False	0.1	0.5	0.5	0.5

Tabla 6.2: Tabla de probabilidad condicional para "nariz dado rostro"

	Nariz	NoNariz
Rostro	0.7941	0.2059
NoRostro	0.1451	0.8549

Tabla 6.3: Tabla de probabilidad condicional para "sistema detecte boca dado boca"

	Boca	NoBoca
SistBoca	0.8947	0.4784
SistNoBoca	0.1052	0.5215

El problema con los detectores de los elementos básicos del rostro es que localiza muchos objetos los cuales considera como algún elemento básico del rostro, pero que en realidad son objetos diferentes al elemento que se desea detectar, es decir, genera muchos falsos positivos. En la Figura 6.5 podemos observar un ejemplo del funcionamiento de los detectores para cada elemento del rostro y cómo es que se detectan muchos falsos positivos en algunas imágenes que se utilizaron para probar el modelo.

Se realizó un proceso de reducción de falsos positivos. En la Figura 6.6 se puede observar el resultado de este proceso de reducción de falsos positivos en las imágenes.

Así como en las imágenes que contienen rostros se localizan muchos falsos positivos, en las imágenes que no contienen rostros también se localizan falsos positivos. En la Figura 6.7 se puede observar algunos ejemplos de falsos positivos localizados en imágenes que no contienen rostros.

6.2.2. Prueba del modelo

En esta fase se llevaron a cabo dos tipos de experimentos. Los primeros fueron utilizando el modelo sin relaciones espaciales y los segundos utilizando el modelo que incluye las relaciones espaciales. Con esto comprobamos que las relaciones espaciales entre los elementos del rostro ofrecen información importante para saber si en la imagen hay un rostro o no. Ambos modelos fueron implementados utilizando la librería ProBT (Kamel Mekhnacha et al.), esta librería permite tanto definir el modelo como hacer la propagación de la red bayesiana y obtener las probabilidades finales de detección del rostro para cada imagen, para ambos casos se utilizaron las 120 imágenes antes mencionadas.

Cada prueba se repitió 10 veces ya que se observó que los detectores no siempre se comportaban de la misma forma; es decir, aunque se detectaba el mismo número de elementos en la imagen, estos elementos variaban en la posición de detección, por lo tanto, se calculó el promedio de detección de estas 10 veces, con esto se trató de cubrir todas las posibles formas de comportamiento de los detectores. Se comparó contra tres variaciones de detectores de rostros basados en AdaBoost. En la Figura 6.8 se muestra la curva ROC para el experimento realizado. Para el caso del modelo con relaciones espaciales se tomó un valor de 0.9 de que la relación se cumpliera, y de 0.1 de que alguna relación no se cumpliera. Se manejó un umbral de tolerancia de 40 píxeles en la alineación de los elementos del rostro.

Para determinar si las relaciones espaciales se cumplían se tomaron en cuenta los centroides de los elementos detectados. Tomando en cuenta la posición de cada centroide se pudo establecer la ubicación de cada elemento para con esto establecer la relación espacial que se cumple, también se consideró que los elementos estuvieran alineados, los umbrales de tolerancia de alineación fueron, 10, 20, 30 y 40 píxeles.

Para evaluar el modelo se fueron variando los umbrales de decisión (entre 0.5 y 0.85) y comparamos el número de verdaderos positivos contra el número de falsos positivos detectados. En la Figura 6.8 se puede observar una mejora significativa cuando las relaciones espaciales son incorporadas a la red bayesiana. Esta mejora es con relación

en el aumento de los *verdaderos positivos* y la reducción de los *falsos positivos* en la detección del rostro. Esto es importante debido a que se tiene un grado de seguridad más alto de que la detección es de un rostro y no de algún otro objeto.

También comparamos nuestro método contra tres variantes de detectores de rostros de Viola y Jones que están implementados en OpenCV (Yu et al., 2004). La Figura 6.9 muestra los resultados de esta comparación. Para los detectores de Viola y Jones no tenemos el control de los umbrales, así que mostramos sólo unos puntos como resultado. Para el tipo de imágenes que se utilizaron nuestro método supera claramente los detectores de rostros. Creemos que la mejora en los resultados de detección es debido a la fusión de información de los detectores y las relaciones espaciales que se establecieron, aunque los detectores de los diferentes elementos del rostro no tienen un desempeño bueno, los resultados mejoran al incluir las relaciones espaciales.

6.3. Discusión

Los resultados obtenidos en la detección de rostros utilizando la gramática para rostros que se definió, son mejores en cuanto a la reducción de falsos positivos.

Esta reducción en el número de falsos positivos es importante debido a que se tiene una mayor certeza de que se está detectando en verdad un rostro. El hecho de descomponer al rostro en diferentes elementos, ayudó a mejorar su detección en las imágenes adquiridas en condiciones difíciles. La ventaja de descomponer al rostro en diferentes elementos nos permitió tener información separada de los elementos que componen al rostro, por tanto, si lográbamos detectar alguno de los elementos del rostro, teníamos información de que en esa imagen había un rostro, sin necesidad de tener el rostro completo.

Por otra parte, al incorporar relaciones espaciales entre los elementos del rostro, nos permitió tener información adicional que nos ayudó a disminuir el número de falsos positivos que aún teníamos en la detección. Con estas relaciones espaciales, pudimos discriminar elementos que eran detectados como parte del rostro y que afectaban los resultados.

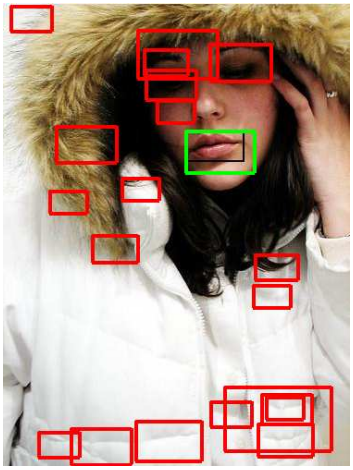
6.4. Conclusiones

De manera general, la detección de rostros consiste en ubicar las regiones en una imagen que pertenezcan a un rostro. Para realizar esta tarea se han desarrollado diversos métodos que utilizan diversas características de la imagen y diversas técnicas. En la mayoría de los métodos propuestos las imágenes con las que se prueban, están en "buenas" condiciones, es decir, no presentan problemas de iluminación o de oclusión del rostro en la imagen y en la mayoría de las imágenes el rostro está en posición frontal.

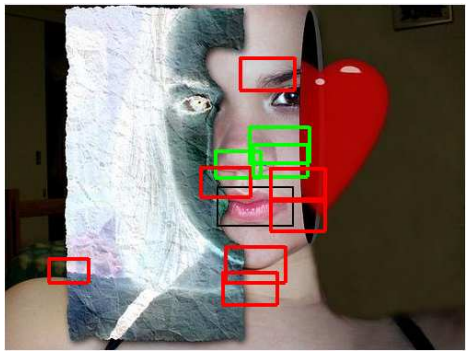
Los modelos que se propusieron en este trabajo, por el contrario, trabajan con imágenes en condiciones "dificiles", como oclusión de los rostros, variación en la posición del rostro, problemas de iluminación, etc.

Los detectores utilizados para ubicar los elementos del rostro generaban muchos falsos positivos, por lo que se llevó a cabo un proceso de reducción de falsos positivos para obtener mejores resultados a la hora de aplicar los modelos definidos a las imágenes de rostros.

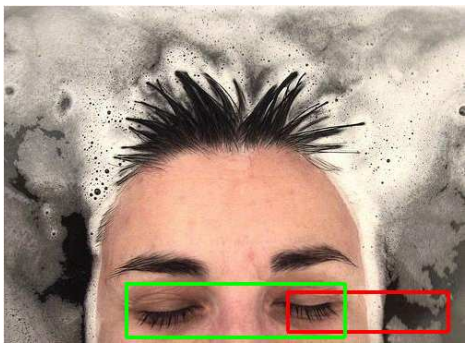
Después de realizar los experimentos se puede observar que al utilizar la grámatica visual para la detección de rostros, se mejora con respecto a los detectores basados en AdaBoost, esta mejora se ve reflejada en la disminución de los falsos positivos y en el aumento de los verdaderos positivos que son detectados en las imágenes. Se pudo observar también que el modelo que incorpora las relaciones espaciales obtiene mejores resultados que el modelo sin relaciones espaciales, lo cual nos indica que las relaciones espaciales aportan información importante a la hora de llevar a cabo el proceso de detección de rostros.



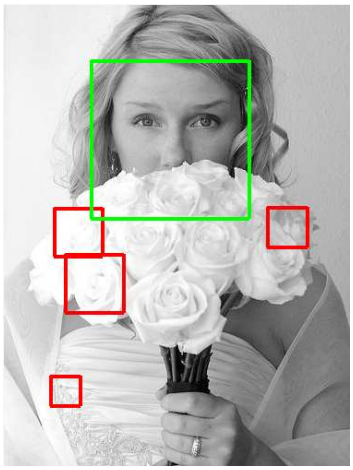
(a)



(b)

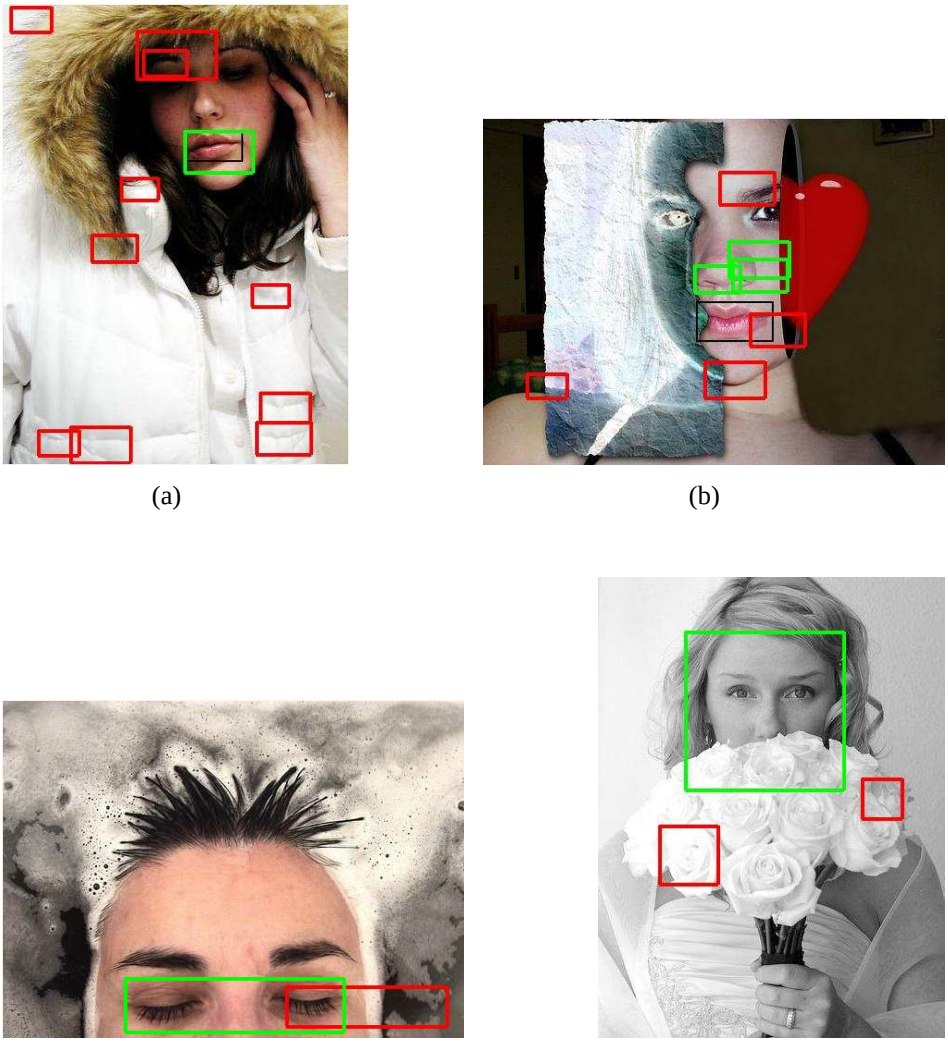


(c)



(d)

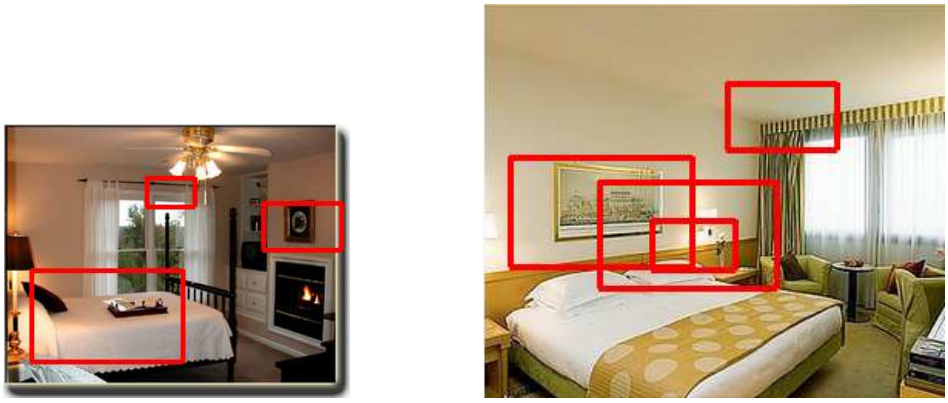
Figura 6.5: Aplicación de los detectores a algunas de las imágenes de entrenamiento. a) Detector para boca. b) Detector para nariz. c) Detector para ojos. d) Detector para cara.



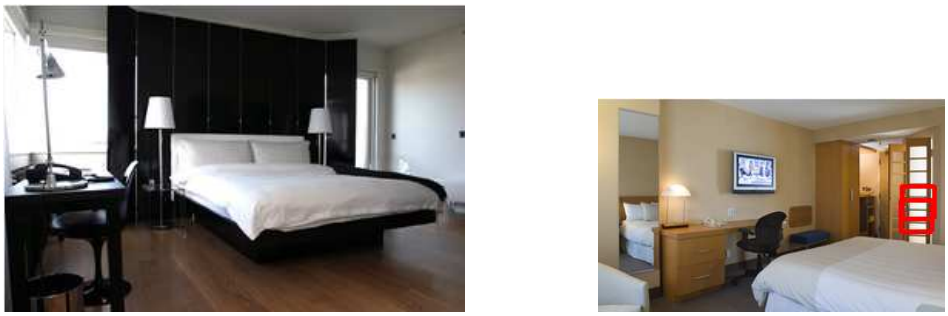
(a) (b)

(c) (d)

Figura 6.6: Reducción del número de falsos positivos en las imágenes. a) Detector para boca. b) Detector para nariz. c) Detector para ojos. d) Detector para cara.



(a) (b)



(c) (d)

Figura 6.7: Funcionamiento de los detectores de los elementos del rostro en imágenes que no continen rostros. a) Detector para boca. b) Detector para nariz. c) Detector para ojos. d) Detector para cara.

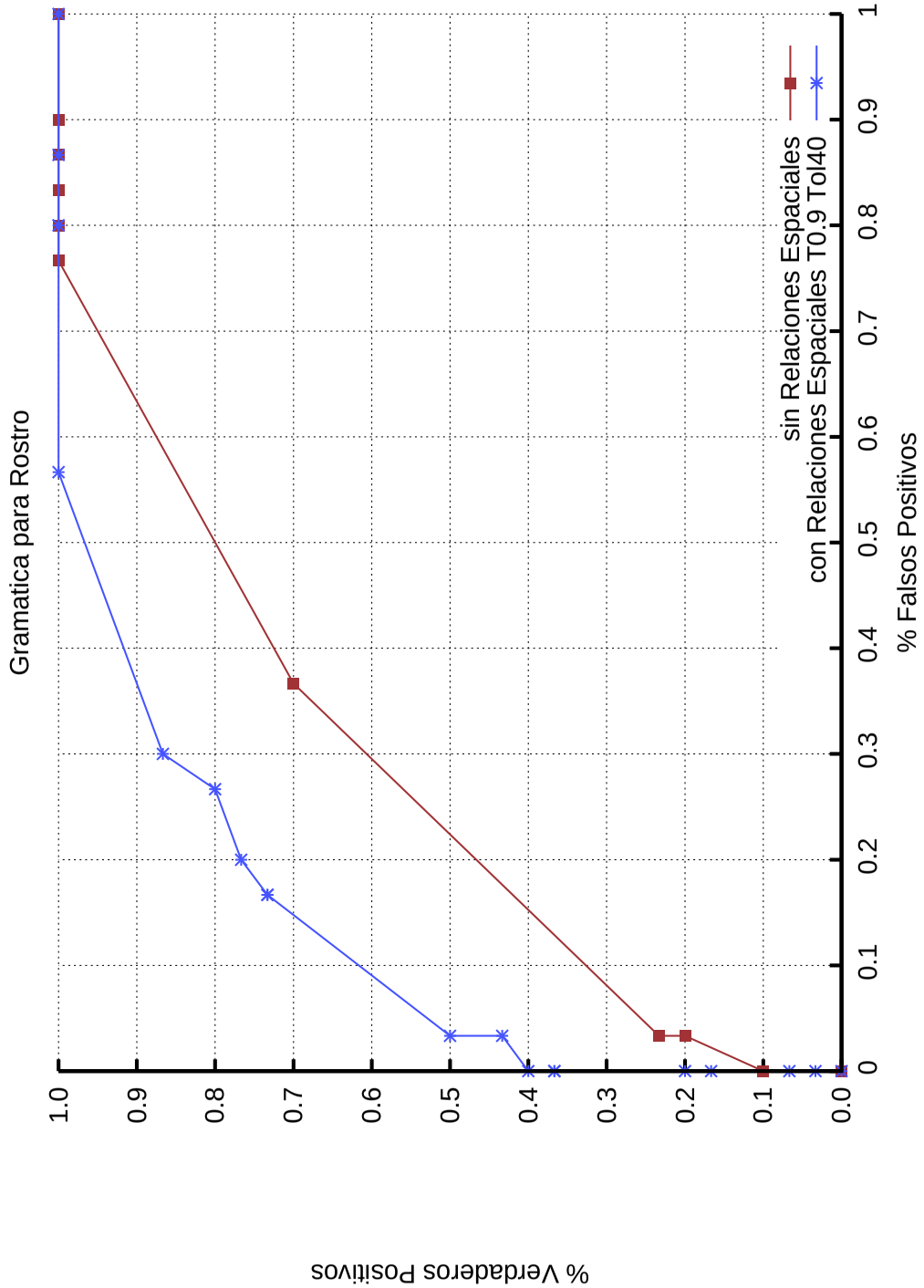


Figura 6.8: Tasa de detección en términos de verdaderos positivos vs. falsos positivos variando el umbral de decisión para el método propuesto. Comparamos la gramática visual con el modelo completo (cruces azules) y la gramática visual con el modelo parcial sin relaciones espaciales (cuadros rojos).

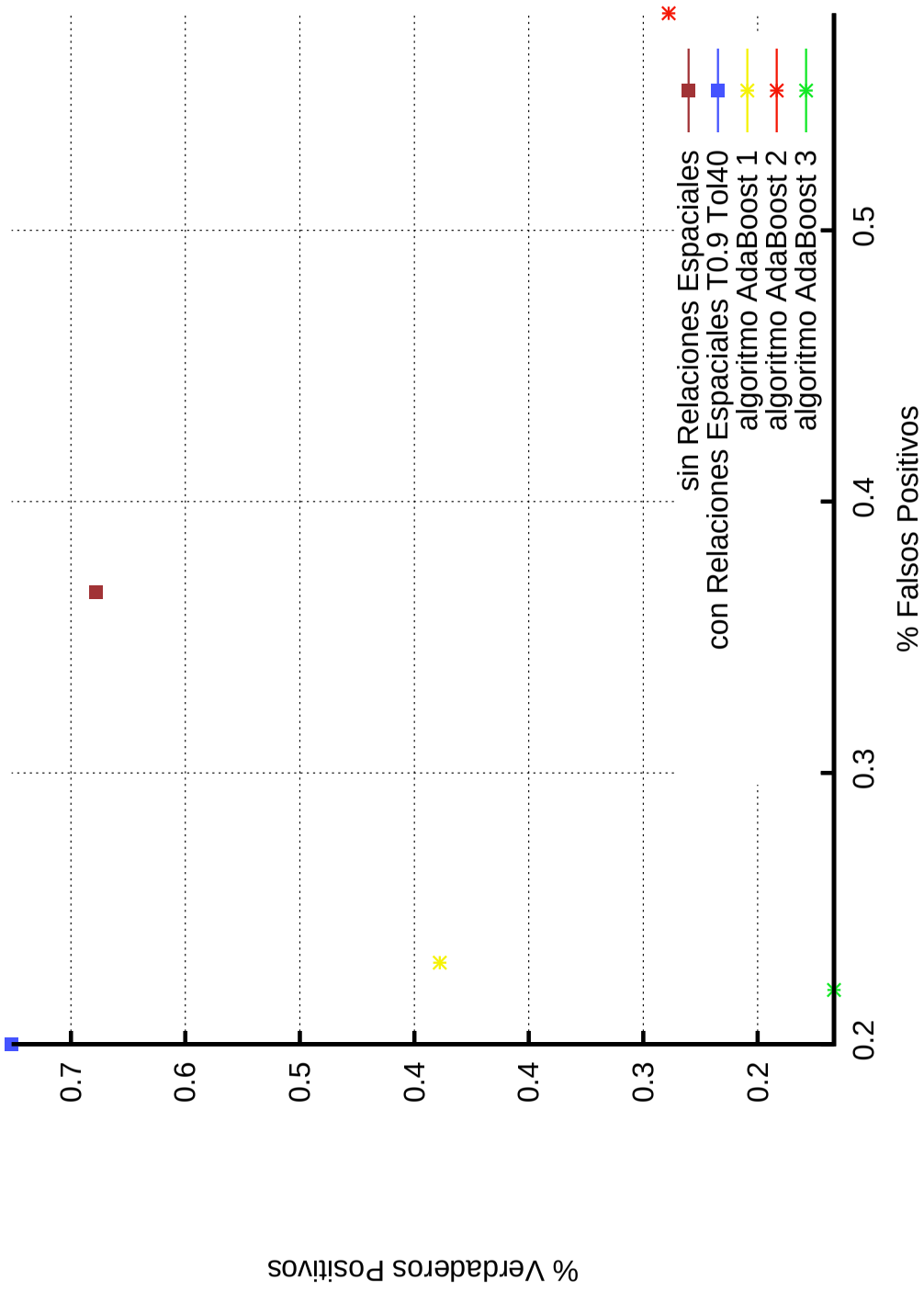


Figura 6.9: Grafica comparativa entre puntos representativos de los modelos propuestos contra tres variantes del detector de rostros de Viola y Jones.

Capítulo 7

Conclusiones y trabajo futuro

7.1. Síntesis de la tesis

Se han desarrollado varios métodos para la detección de rostros con cierto éxito, no obstante, éstos métodos tienden a fallar bajo condiciones difíciles tales como oclusiones parciales y cambios en la orientación e iluminación.

En esta tesis se presentó un método novedoso para la detección de rostros basado en una *gramática visual* para rostros. Hemos definido una gramática para rostros que después fue transformada en una red bayesiana, cuyos parámetros se obtienen a partir de datos. Se aplicó este modelo para la detección de rostros en condiciones difíciles con resultados muy buenos en comparación con otros detectores de rostros reportados en el estado del arte.

Se evaluó el modelo basado en una gramática visual que incorpora relaciones espaciales y se comparó contra un modelo simplificado sin relaciones espaciales y contra otros detectores de rostros.

7.2. Conclusiones

Se pudo observar que el detector basado en la gramática visual tiene resultados satisfactorios en imágenes con condiciones difíciles. Por otra parte y de manera particular, se logra una importante reducción en el número de falsos positivos generados por los detectores de los diferentes elementos que componen un rastro.

Se pudo comprobar que la incorporación de las relaciones espaciales al modelo mejora la detección del rostros en las imágenes utilizadas.

Se propuso un modelo para la detección de rostros basado en una gramática

simbólica-relacional, la cual fue transformada en una red bayesiana para poder llevar a cabo la detección de rostros en condiciones difíciles.

7.3. Trabajo futuro

Aunque la gramática actual es restringida, consideramos que esto podría ser ampliado para proporcionar una descripción más completa de la cabeza desde puntos de vista diferentes, y también para la representación de otras clases de objetos. En el futuro queremos explorar el aprendizaje de la gramática a partir de ejemplos. Por otra parte, se podría pensar en aprender la estructura de la red bayesiana que define a la grámatica a partir de las reglas que definan a algún objeto.

Tomando en cuenta que las relaciones espaciales entre los elementos definidos proporcionan información importante para obtener buenos resultados, se puede pensar en incluir otras relaciones espaciales que hagan más completa la gramática y por tanto un modelo más robusto que el presentado en este trabajo, en el que los resultados de la detección sean mejores.

Se podría utilizar algún otro formalismo de los mencionados en (Marriot y Meyer, 1998), para definir de manera conceptual la gramática. Quizá utilizando otro formalismo se pueda integrar informacion adicional como, color de piel o posición del rostro en la imagen que puede dar información adicional importante para la detección del rostro.

Apéndices

Apéndice

Tablas de probabilidad para los modelos definidos

Tabla A.1: Tablas de probabilidad de los elementos del rostro y de las relaciones espaciales.

Rostro	NoRostro
0.5	0.5

$P(Rostro)$

	Boca	NoBoca
Rostro	0.8823	0.1177
NoRostro	0.1612	0.8388

$P(Boca|Rostro)$

	Nariz	NoNariz
Rostro	0.7941	0.2059
NoRostro	0.1451	0.8549

$P(Nariz|Rostro)$

	Ojos	NoOjos
Rostro	0.7058	0.2942
NoRostro	0.1291	0.8709

$P(Ojos|Rostro)$

	Cara	NoCara
Rostro	0.4657	0.5342
NoRostro	0.1827	0.8173

$P(Cara|Rostro)$

	Boca	NoBoca
SistBoca	0.8947	0.4784
SistNoBoca	0.1052	0.5215

$P(SistBoca|Boca)$

	Nariz	NoNariz
SistNariz	0.7894	0.4555
SistNoNariz	0.2105	0.5444

$P(SistNariz|Nariz)$

	Ojos	NoOjos
SistOjos	0.4102	0.1489
SistNoOjos	0.5897	0.8510

$P(SistOjos|Ojos)$

	Cara	NoCara
SistCara	0.4852	0.3024
SistNoCara	0.5147	0.6975

$P(SistCara|Cara)$

	Elemento 1		NoElemento 1	
	Elemento 2	NoElemento 2	Elemento 2	NoElemento 2
RE = Verdadero	0.9	0.5	0.5	0.5
RE = Falso	0.1	0.5	0.5	0.5

$P(RelacionEspacial|elemento1, elemento2)$

Bibliografía

Chai, D., y Bouzerdoum, A. 2000. A bayesian approach to skin color classification in ycbcr color space. 2:421–424.

Chow, C. I.; Member, S.; y Liu, C. N. 1968. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory* 14:462–467.

Froba, B., y Ernst, A. 2004. Face detection with the modified census transform. *Automatic Face and Gesture Recognition, IEEE International Conference on* 0:91.

Garcia, C., y Delakis, M. 2002. A neural architecture for fast and robust face detection. *Pattern Recognition, International Conference on* 2:20–44.

García, G. A. R. 2006. *Deteccion de Rostros con Aprendizaje Automatico*.

Hamouz, M.; Kittler, J.; Kamarainen, J.-K.; Paalanen, P.; y Kalviainen, H. 2004. Affine-invariant face detection and localization using GMM-based feature detector and enhanced appearance model. *Automatic Face and Gesture Recognition, IEEE International Conference on* 67.

Han, F., y Zhu, S.-C. 2005. Bottom-up/top-down image parsing by attribute graph grammar. *Computer Vision, IEEE International Conference on* 2:1778–1785.

Hjelmas, E., y Low, B. K. 2001. Face detection: A survey. *Computer Vision and Image Understanding* 83:236–274.

Jesorsky, O.; Kirchberg, K. J.; y Frischholz, R. 2001. Robust face detection using the hausdorff distance. 90–95.

Kamel Mekhnacha; Linda Smail; Juan-Manuel Ahuactzin; Pierre Bessière; y Emmanuel Mazer. A unifying framework for exact and approximate Bayesian inference. Technical report.

Kongqiao, W. 2003. Automatical face detection in images with complex background. *International Conference on Neural Networks and Signal Processing* 2:1027–1030.

Kovac, J.; Peer, P.; y Solina, F. 2003. Illumination independent color-based face detection. 510–515.

Li, S. Z.; Jain, A. K.; y Li, S. Z. 2005. *Face Detection*. Springer New York.

Marriot, K., y Meyer, B. 1998. *Visual Language Theory*. New York: Springer.

Neapolitan, R. E. 2004. *Learning Bayesian Networks*. Chicago, Illiniois: Pearson, Prentice Hall.

Park, S., y Aggarwal, J. K. 2004. A hierarchical bayesian network for event recognition of human actions and interactions. *Multimedia Systems* 10.

Pearl, J., y Dechter, R. 1989. Learning structure from data: a survey. 230–244.

Peer, P.; Kovac, J.; y Solina, F. Human skin colour clustering for face detection.

Ramírez, G. A.; Zanella, V.; y Fuentes, O. 2003. Heuristic-based automatic face detection. 267–272.

Rowley, H. A.; Baluja, S.; y Kanade, T. 1998a. Neural network-based face detection. 23–38.

Rowley, H. A.; Baluja, S.; y Kanade, T. 1998b. Rotation invariant neural network-based face detection. 38–44.

Schapire, R. E. 2002. The boosting approach to machine learning: An overview.

Schneiderman, H., y Kanade, T. 1998. Probabilistic modeling of local appearance and spatial relationships for object recognition. In *CVPR '98: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 45. Washington, DC, USA: IEEE Computer Society.

Schneiderman, H., y Kanade, T. 2000. A statistical method for 3d object detection applied to faces and cars. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on* 1:1746.

Schneiderman, H. 2004. Feature-centric evaluation for efficient cascaded object detection. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on* 2:29–36.

Seow, M.-J.; Valaparla, D.; y Asari, V. K. 2003. Neural network based skin color model for face detection. *Applied Image Pattern Recognition Workshop,* 0:141.

Sierra, A. B. 2006. *Aprendizaje Automático : Conceptos básicos y Avanzados*. Mexico: Pearson.

Stephen, Y. H.; Lin, S.; Li, S. Z.; Lu, H.; y yeung Shum, H. 2004. Face alignment under variable illumination. 85–90.

Storkey, A. J., y Williams, C. K. I. 2003. Image modeling with position-encoding dynamic trees. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 25(7):859–871.

Tan, K.-H., y Ahuja, N. 2001. A representation of image structure and its application to object selection using freehand sketches. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on* 2:677.

Viola, P., y Jones, M. J. 2004. Robust real-time face detection. *International Journal of Computer Vision* 57:137–154.

Wu, T.-F.; Xia, G.-S.; y Zhu, S.-C. 2007. Compositional boosting for computing hierarchical image structures. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on* 0:1–8.

Yang, M.-H.; Kriegman, D. J.; y Ahuja, N. 2002. Detecting faces in images: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24:34–58.

Yu, Q.; Cheng, H. H.; Cheng, W. W.; y Zhou, X. 2004. Ch opencv for interactive open architecture computer vision. *Advances in Engineering Software* 35(8-9):527–536.

Yuille, A. L.; Hallinan, P. W.; y Cohen, D. S. 1992. Feature extraction from faces using deformable templates. *International Journal of Computer Vision* 8(2):99–111.

Zhu, L.; Chen, Y.; y Yuille, A. 2009. Unsupervised learning of probabilistic grammar-markov models for object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31(1):114–128.

Zhu, S.-C., y Mumford, D. 2006. A stochastic grammar of images. *Found. Trends. Comput. Graph. Vis.* 2(4):259–362.