Reconocimiento visual con técnicas de aprendizaje de máquina

Pablo Espinace
Doctor en Ciencias de la Ingeniería y Magíster en Ciencias de la Ingeniería de la Pontificia Universidad Católica de Chile (2011). Actualmente es investigador en el Departamento de Ciencia de la Computación de la Escuela de Ingeniería de la Pontificia Universidad Católica de Chile. Sus intereses de investigación incluyen Robótica Móvil, Inteligencia de Máquina y Visión por Computador.
pespinac@ing.puc.cl

Billy Peralta
Candidato a Doctor en Ciencias de la Ingeniería de la Pontificia Universidad Católica de Chile. Magíster en Ciencias de la Ingeniería (2007). Sus intereses de investigación principales son Inteligencia de Máquina y Visión por Computador.
bmperalt@uc.cl

Álvaro Soto
Profesor Asociado del Departamento de Ciencia de la Computación de la Escuela de Ingeniería de la Pontificia Universidad Católica de Chile. Ph.D. en Ciencias de la Computación Carnegie Mellon University (2002); Magíster en Ingeniería Eléctrica y Computación en Louisiana State University (1997). Sus intereses principales en investigación son: Aprendizaje de Máquina, Robótica Cognitiva y Reconocimiento Visual.
asoto@ing.puc.cl

La gran versatilidad de nuestro propio sistema de percepción visual, es un claro ejemplo de la gran relevancia de poder contar con sistemas artificiales de reconocimiento visual que faciliten la operación de robots y máquinas embebidas en nuestros ambientes cotidianos. Sin embargo, complejidades del mundo visual como cambios de pose, escala o condiciones de iluminación, así como situaciones de oclusión o cambios en la configuración de objetos deformables, han resultados desafíos de envergadura mayor, que han complicado el desarrollo de posibles soluciones.

Afortunadamente, la exitosa reciente aplicación de técnicas de aprendizaje de máquina al área de reconocimiento visual ha abierto nuevas puertas que han llenado de optimismo a esta área [1,2,3]. La Figura 1 ejemplifica el aporte medular que ha ofrecido el aprendizaje de máquina al reconocimiento visual. En la Figura, pese a la mala calidad de la imagen, nuestro sistema visual es capaz de distinguir que se trata de una mesa con sillas a su alrededor, sobre la cual hay un objeto decorativo. En forma notable, mediante la integración de conocimiento aprendido en nuestro largo interactuar con el mundo natural, nuestro sistema visual es capaz de realizar una inferencia que va mucho más allá de la información contenida en los ruidosos píxeles de esta imagen. Esta capacidad, de ir más allá de la información en una imagen determinada, es el ingrediente clave que ha aportado la inteligencia de máquina al reconocimiento visual. De esta manera, los algoritmos de aprendizaje de máquina están permitiendo extraer desde miles o millones de imágenes, patrones de apariencia visual y relaciones contextuales relevantes, las cuales luego pueden ser usadas para reconocer elementos en una imagen particular. Este tipo de aprendizaje

visual está permitiendo por primera vez crear sistemas de reconocimiento visual capaces de operar exitosamente en ambientes naturales [1,4].

En este artículo, mediante dos casos ilustrativos, queremos mostrar algunos de los aportes realizados por nuestro Grupo de Investigación en Inteligencia de Máquina, GRIMA [5], en el área de reconocimiento visual. Nuestro primer ejemplo corresponde a un sistema de reconocimiento de escenas de ambientes interiores que desarrollamos para nuestro robot móvil. Este sistema utiliza relaciones contextuales entre objetos y escenas, de manera que si nuestro robot encuentra una sala con refrigerador y microondas pensará que se encuentra en una cocina. Interesantemente, nunca entregamos esta información directamente a nuestro modelo, sino que nuestro sistema es capaz de inferir (aprender) este tipo de relaciones por sí mismo, analizando la información textual contenida en millones de imágenes del popular sitio web Flickr. En nuestro segundo ejemplo, exploramos la relación inversa entre objetos y escenas, es decir, cómo conocimiento holístico sobre la escena condiciona las relaciones contextuales entre objetos. De esta manera, podemos construir un modelo capaz de distinguir que en una escena de parque es altamente probable ver personas caminando al lado de un perro, pero que esa misma relación es altamente improbable para el caso de un edificio de oficinas.

RECONOCIMIENTO DE ESCENAS A TRAVÉS DE DETECCIÓN DE OBJETOS

En el ámbito de reconocimiento de escenas, los primeros métodos buscaban extraer características y categorizar la escena en la cual se tomó una imagen mediante representaciones "holísticas", esto es, analizando la información contenida dentro de la imagen como un todo [6,7]. Luego, algunos trabajos buscaron utilizar representaciones intermedias [8,9,10], en algunos casos incluyendo información espacial [11]. Todos estos métodos tuvieron relativo éxito, especialmente en categorización de imágenes de exterior, sin embargo, su desempeño en escenas de interior resulta ser muy pobre. Las razones para este bajo rendimiento saltan a la vista: una escena de interior (o habitación) puede ser muy similar, o incluso igual a otra, excepto por los objetos que contiene, los cuales son muy diversos en apariencia y posibles poses.

A modo de ejemplo, consideremos una habitación de 3x3 metros inicialmente vacía. Sin objetos en su interior, muy difícilmente se podrá distinguir de qué tipo de habitación se trata. Ahora, si en esta habitación ponemos una silla y un escritorio, pasa de inmediato a tomar la forma de una oficina. Luego, si sacamos la silla y el escritorio, y en su lugar ponemos una cama y una lámpara, la habitación tomará la forma de un dormitorio. Como se ve, en términos globales la habitación es la misma, pero son los objetos dentro de ella los que hacen la diferencia.

Considerando lo anterior, diversos trabajos han comenzado a utilizar la detección de objetos como parte central de la detección de escenas [12,13,14]. La idea principal es utilizar los objetos como partes componentes de una escena, además de aprovechar la configuración espacial de estos para mejorar el rendimiento y los resultados obtenidos. A continuación presentamos el trabajo realizado al respecto en una investigación conjunta entre GRIMA y el Laboratorio de Ciencia de la Computación e Inteligencia Artificial del Instituto Tecnológico de Massachusetts (CSAIL-MIT).

Figura 1

A pesar de lo borroso de la imagen, nuestro sistema visual es capaz de inferir información a partir de conocimiento aprendido previamente.

Modelamiento matemático del problema

El objetivo de nuestro método es encontrar la distribución de probabilidad de los valores que puede tomar una variable ξ, que representa las distintas etiquetas que se pueden asignar a una escena. En el caso de escenas de exterior estas etiquetas podrían ser bosque, playa, ciudad, etc., mientras que en escenas de interior podrían ser oficina, dormitorio, cocina, etc. Como punto de

partida, se tienen las características visuales que se pueden extraer directamente desde los píxeles de la imagen correspondiente, bajo un método de ventana deslizante. Así, si asumimos que se tienen w_L ventanas, a cada una de las cuales se le extrae un conjunto $\vec{f}$ de características, tendríamos un conjunto total de $\vec{f}_{1:w_L}$ características extraídas a todas las ventanas. De esta forma, lo que buscamos es la distribución de probabilidad de ξ dada la información $\vec{f}_{1:w_L}$, esto es, $p(\xi \mid \vec{f}_{1:w_L})$.

El hecho de ocupar un método de ventana deslizante nos permite obtener características locales en diversas partes de la imagen, en lugar de características globales. Dado que nuestro objetivo es reconocer una escena a través de los objetos presentes en ella, nuestro método implementa un clasificador que determina la probabilidad de que cada uno de S objetos distintos esté presente en cada una de las ventanas, a partir del conjunto de características $\vec{f}$ extraído a cada ventana. Se representa la salida de este clasificador para el conjunto de w_L ventanas como $c_{1:w_L}$. De esta forma, las características locales extraídas en diversas partes de la imagen (ventanas), se transforman en probabilidades de que distintos objetos estén presentes en la escena donde fue tomada la imagen.

Teniendo información probabilística sobre los objetos que están presentes en la escena donde fue tomada la imagen, se debe establecer una relación entre el conjunto de objetos presentes y la etiqueta a asignar a esta escena. De esta manera se podrá saber cuál es la escena más probable dados los objetos presentes en ella. Así, nuestro método implementa una relación contextual, la cual determina a partir de información de frecuencia la probabilidad de que una escena tome cada uno de los valores posibles para sus etiquetas. Se representa la presencia o ausencia de cada uno de los objetos disponibles a través de la variable conjunta $o_{1:s}$, la cual tiene un valor por cada categoría de objeto.

Las variables asociadas a las salidas de los clasificadores de objetos, $c_{1:w_L}$, y la combinación de objetos presente en la escena, $o_{1:s}$, son incorporadas al cálculo de $p(\xi \mid \vec{f}_{1:w_L})$ a través de la Ley de Probabilidad Total:

$$p(\xi \mid \vec{f}_{1:w_L}) = \sum_{o_{1:s}} \sum_{c_{1:w_L}} p(\xi \mid o_{1:s}, c_{1:w_L}, \vec{f}_{1:w_L}) p(o_{1:s}, c_{1:w_L} \mid \vec{f}_{1:w_L})$$

Simplificando y aplicando la Ley de Probabilidad Conjunta al segundo término, obtenemos:

$$p(\xi \mid \vec{f}_{1:w_L}) = \sum_{o_{1:s}} \sum_{c_{1:w_L}} p(\xi \mid o_{1:s}) p(o_{1:s} \mid c_{1:w_L}) p(c_{1:w_L} \mid \vec{f}_{1:w_L})$$

De esta manera, se construye nuestro método de reconocimiento de escenas a través de detección de objetos en base a tres términos principales:

- $p(c_{1:w_L} \mid \vec{f}_{1:w_L})$, que transforma las características extraídas en las ventanas a salidas de un clasificador de objetos.
- $p(o_{1:s} \mid c_{1:w_L})$, que representa la confianza que se tiene en el clasificador de objetos utilizado.
- $p(\xi \mid o_{1:s})$, que relaciona una combinación de objetos encontrados a través de la clasificación a una escena.

Podemos ver en estos términos que las variables incorporadas funcionan como intermediarios entre las características extraídas a partir de la información de los píxeles de la imagen, de bajo nivel semántico, y la escena donde fue tomada la imagen, de alto nivel semántico. Como se verá en la siguiente sección, el cálculo de los términos expuestos anteriormente requiere de información externa a la imagen misma desde donde se extrae la información de los píxeles, requiriéndose ir a buscar información más allá de la propia imagen.

Adicionalmente, nuestro modelo agrega características asociadas a información tridimensional extraída a través de un sensor de profundidad. Matemáticamente, esta información tridimensional se representa a través de un conjunto de características tridimensionales $\vec{d}$, extraídas en cada ventana las que se agregan a las características visuales ya disponibles. Así, se incorpora el conjunto de características tridimensionales extraídas en todas las ventanas $\vec{d}_{1:w_L}$, al término asociado al clasificador de objetos:

$$p(c_{1:w_L} \mid \vec{f}_{1:w_L}) \quad \rightarrow \quad p(c_{1:w_L} \mid \vec{f}_{1:w_L}, \vec{d}_{1:w_L})$$

Como se verá a continuación, la incorporación de información tridimensional no sólo permite tener información más rica de la imagen en cuestión, sino que también permite implementar nuestro modelo de manera eficiente.

Implementación

Como se vio en la sección anterior, nuestro modelo depende de tres términos principales. A continuación, veremos cómo se implementa cada uno de estos tres términos para lograr el reconocimiento de una escena a través de la detección de los objetos presentes en ella.

El término $(c_{1:w_L} \mid \vec{f}_{1:w_L}, \vec{d}_{1:w_L})$ relaciona las características extraídas de la información visual y 3D a la salida de un clasificador de objetos. En nuestro caso, y de modo de aprovechar la información tridimensional de la mejor forma posible, se usa esta información como un primer filtro, construyendo un foco de atención que define sectores de la imagen donde luego se utilizará un clasificador con las características visuales. La Figura 2 muestra los resultados de este foco de atención aplicado al caso de la detección de un monitor de computador.

Para el caso de la información tridimensional, se construye un modelo en base a ejemplos de cada objeto, el que define distintas características geométricas sobre una categoría de objetos, por ejemplo tamaños típicos, los cuales se utilizan para calcular la probabilidad de que un objeto esté presente en una ventana en base sólo a información tridimensional. Para el caso de la información visual, se construye por cada objeto un clasificador bajo el método de AdaBoost, similar a lo hecho en [1], usando como datos de entrenamiento

Figura 2

Ejemplo de ejecución del foco de atención con información 3D. (a) Imagen donde dos monitores son detectados. (b) Resultados del foco de atención, ejecutado como paso previo a la clasificación visual y que acota el espacio donde se buscan monitores.

imágenes etiquetadas que contienen a los objetos, extraídas de distintos sets de datos disponibles en la Web.

El término $p(o_{1:S}|c_{1:w_L})$ representa la confianza que se tiene en las salidas de la clasificación de objetos, esto es, qué tan probable es que una cierta configuración de objetos esté presente en la escena dado que la clasificación dijo que esa configuración estaba presente. Esta confianza se mide a través de probar los clasificadores con conjuntos de imágenes que contienen los objetos respectivos, distintas a las utilizadas para el entrenamiento, las que una vez más son extraídas desde distintos sets de datos disponibles en la Web.

El término $p(\xi\,|o_{1:S})$ relaciona una combinación de objetos encontrados a través de la clasificación a una escena. Esto se puede ver como una relación de contexto objeto-escena, donde se busca saber qué tan probable es la aparición conjunta de un objeto, o un conjunto de estos, con cada escena particular. Para calcular este término nuevamente se utiliza un gran conjunto de imágenes extraídas de la Web, en este caso de Flickr, con objetos y escenas etiquetadas. Luego, se calcula utilizando frecuencias la probabilidad de que una combinación de objetos esté en una escena dada, a través de analizar cuántas veces aparecen etiquetas asociadas a todos estos objetos en imágenes etiquetadas como la escena en cuestión.

Como se puede ver, para calcular cada uno de estos tres términos debemos recurrir a fuentes de datos externas a la imagen misma, lo que muestra lo indispensable que es recurrir a información que va más allá de la imagen analizada. Las técnicas de aprendizaje de máquina han resultado efectivas en esta tarea, renovando los aires en el ámbito del reconocimiento visual.

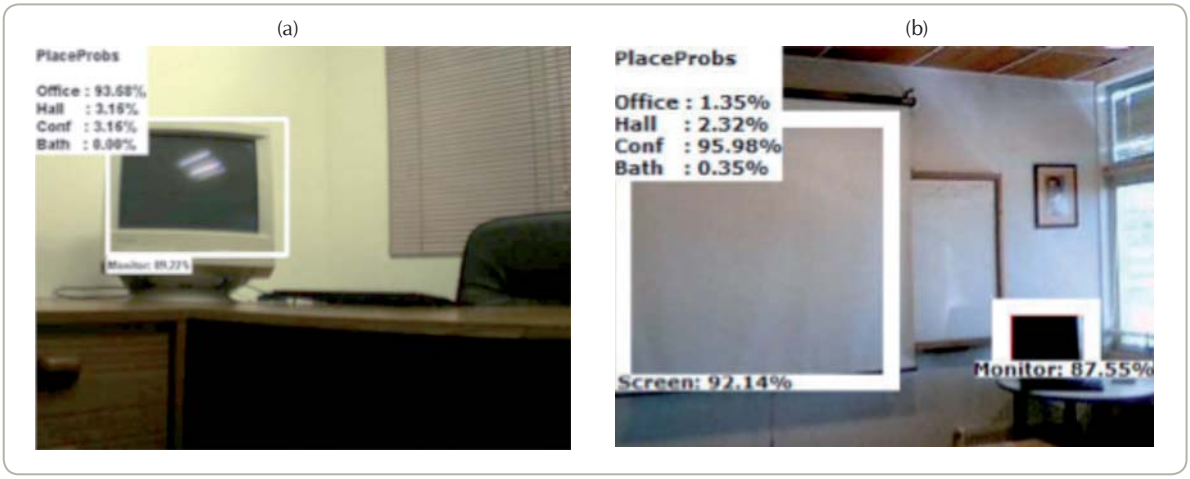
Resultados

A continuación presentamos algunos resultados obtenidos con nuestro método, en pruebas realizadas en un ambiente de oficinas, utilizando siete objetos entre los que se encontraban monitor y pantalla de proyector, y cuatro posibles escenas entre las que se encontraban oficina y sala de conferencias.

La Figura 3 muestra cómo el encontrar distintos objetos, o combinaciones de estos, define la detección de distintas escenas. En la Figura 3(a) se puede ver que al detectarse sólo un monitor en la escena, la detección más probable es una oficina, ya que de acuerdo a los datos utilizados es en las oficinas donde en mayor cantidad se encuentran monitores, en relación a las otras escenas utilizadas. Aún así, se mantiene cierta probabilidad de que sean otras las escenas reales en las que se tomó la imagen, debido a que pueden haber también monitores en ellas, además de existir cierta incerteza en la detección del monitor. Resultados similares se pueden ver en la Figura 2(a).

En la Figura 3(b) se puede ver un segundo caso en el que también se ha detectado un monitor, sin embargo en este caso, al estar acompañado de una pantalla de proyector, la escena que se detecta es una sala de conferencias, ya que la combinación de objetos hace que ésta sea la escena más probable. Nuevamente, se mantiene cierta probabilidad de que sea otra la escena, dada la incerteza en las detecciones.

Figura 3



Detecciones en ambiente de oficina. (a) Sólo un monitor es detectado, llevando al reconocimiento de una oficina. (b) Un monitor junto a una pantalla de proyector son detectados, llevando a la detección de una sala de conferencias.

La Tabla 1 muestra una matriz de confusión de las detecciones encontradas en imágenes tomadas en las distintas escenas utilizadas, haciendo una comparación de nuestro método (OM) con tres métodos alternativos que no utilizan objetos para la detección de escenas: OT-G [7], LA-SP [11] y pLSA [10]

Se puede ver que los resultados de nuestro método muestran una clara inclinación hacia la diagonal, mejorando bastante los resultados de los métodos alternativos. Aún así, existen ciertos errores, derivados principalmente de falsos positivos o falsos negativos en las detecciones de objetos, además del hecho de que no todos los objetos en una escena se pueden observar en una imagen tomada sólo en un área parcial de la misma.

RECONOCIMIENTO DE OBJETOS A TRAVÉS DE CONTEXTOS JERÁRQUICOS ADAPTIVOS

En el ámbito del reconocimiento de objetos, existen diversos métodos que han tenido éxito a través del uso de distintas características visuales y algoritmos asociados. Hasta hace algunos años, prácticamente la totalidad de estos métodos se basaba en extraer características a un objeto como un todo [1,15], mientras que recientemente se han comenzado a utilizar representaciones basadas en partes de los objetos y las relaciones espaciales entre éstas, las que han tenido gran éxito [3].

Tabla 1

	OM				OT-G			
Scene	Off.	Hall	Conf.	Bath.	Off.	Hall	Conf.	Bath.
Office	**91%**	7%	2%	0%	**83%**	4%	13%	0%
Hall	7%	**89%**	4%	0%	7%	**86%**	5%	2%
Conference	7%	7%	**86%**	0%	17%	3%	**79%**	1%
Bathroom	0%	6%	0%	**94%**	3%	5%	3%	**89%**
	LA-SP				**pLSA**			
Scene	Off.	Hall	Conf.	Bath.	Off.	Hall	Conf.	Bath.
Office	**72%**	6%	22%	0%	**88%**	4%	7%	1%
Hall	19%	**71%**	9%	1%	4%	**87%**	5%	4%
Conference	24%	6%	**67%**	3%	11%	2%	**85%**	2%
Bathroom	2%	4%	3%	**91%**	1%	4%	3%	**92%**

Comparación de nuestro método con métodos alternativos.

Una idea atractiva es la de incorporar información de contexto, esto es, información de las relaciones entre distintos objetos, o entre estos y la escena donde se encuentran, con lo cual se ha logrado mejorar el rendimiento de diversos clasificadores de objetos que no incorporan estas relaciones [16,17]. En particular, la información de contexto local, donde se cuantifica la relación entre partes de la imagen o entre ubicaciones particulares de objetos en ella, ha tenido varios casos de éxito [18,19].

Dentro de los trabajos mencionados anteriormente, uno que resulta muy relevante para modelar eficientemente la relación entre objetos es el de Choi et al. [16]. En este caso, se usa una red bayesiana en forma de árbol para representar las probabilidades condicionales entre objetos. Una gran ventaja del uso de una red bayesiana es que permite hacer inferencia de forma eficiente aprovechando la factorización del modelo gráfico. Sin embargo, una restricción de este modelo es que las interrelaciones entre objetos están fijas, es decir, no dependen de la escena donde éstas hipotéticamente ocurren. Por ejemplo, consideremos el caso de los objetos "perro" y "persona". Si consideramos la escena "oficina", la co-ocurrencia de ambos objetos tiende a ser bastante baja. Por otro lado, al analizar la misma relación en la escena "parque", la co-ocurrencia de ambos objetos tiende a ser mucho más alta.

A continuación, presentamos un trabajo realizado en GRIMA que incorpora conocimiento holístico sobre la escena en que fue tomada una imagen, condicionando las relaciones contextuales entre objetos, de modo de poder construir un modelo adaptivo que mejore las debilidades del algoritmo de Choi et al.

Modelamiento matemático del problema

Con el objetivo de incorporar información adaptiva acerca de los distintos contextos en los que podría haber sido tomada una imagen, nuestro método modela el problema de reconocimiento de objetos en el marco de una mezcla de expertos [20]. En particular, el método aprende relaciones adaptivas condicionales entre objetos de acuerdo a distintos árboles, los que son ponderados según la información de la escena.

Cada árbol sigue la estructura dada en [16], donde se acoplan los modelos a priori y de verosimilitud. El modelo a priori representa el conocimiento previo acerca de la ocurrencia y ubicación de los objetos. Se denomina b_i a la variable que representa la presencia de un objeto de categoría i y L_i a la variable que representa la ubicación y escala de los objetos de esta categoría. El modelo de verosimilitud predice la presencia de un objeto de categoría i en una ventana w, utilizando la información de la imagen misma. Aquí se utiliza un clasificador de objetos para cada categoría, el cual corresponde a una implementación del clasificador presentado en [3], además de información holística de la escena calculada a través del método Gist [7].

Debido a que posiblemente hay múltiples detecciones de un objeto, el modelo se simplifica al considerar las K mejores detecciones para cada objeto, ordenadas por score. Para cada detección $k = 1...K$, de cada objeto i, se incorpora en el modelo la validez o confianza de la detección C_{ik}, el score de la clasificación S_{ik}, la ubicación y escala de cada detección W_{ik} y la información global de la imagen g_L.

Es importante notar que la información global de la escena g_L se considera en forma individual para la detección de cada objeto y no considera la co-ocurrencia de los mismos, de modo de alcanzar una detección adecuada que incorpore la información del conjunto de objetos como un todo.

Dado que lo que buscamos es incluir adaptivamente la información de varios contextos, nuestro trabajo se centra en el modelo a priori de la ocurrencia de los objetos. En el trabajo de Choi et al. esta variable se representa por medio de una red bayesiana en forma de árbol. Nosotros planteamos una mezcla de expertos de redes bayesianas, donde el peso de cada una es función de las características de la imagen de entrada x_G.

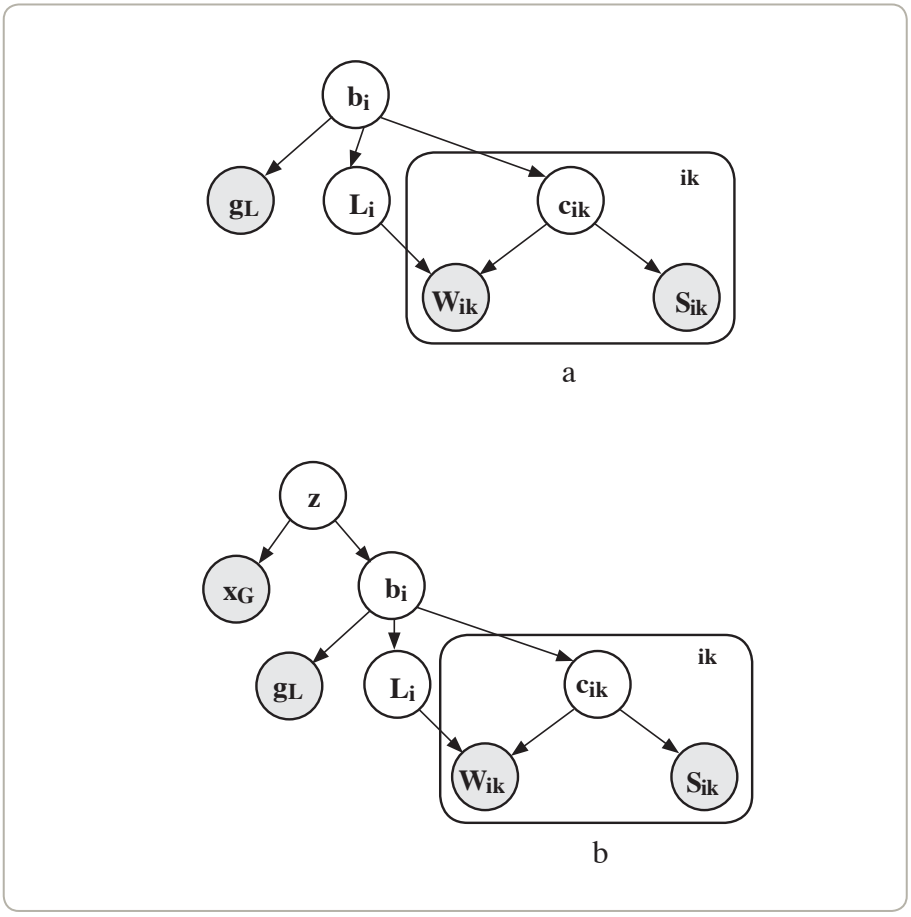
Asumiendo un conjunto de N imágenes de entrenamiento y D categorías, podemos construir N pares (x_G, b) donde $b \in 2^D$. Luego, relacionamos ambas variables usando una variable latente z que representa la escena a la cual pertenece la imagen (Figura 4.b). Asumiremos que hay K valores para z . En este punto sólo consideramos las variables x_G, b y z en forma conjunta por medio de la mezcla de expertos, con lo cual tenemos:

$$p(b|x_G) = \sum_{i=1}^{K} p(b, z_i|x_G) \quad =$$

$$\sum_{i=1}^{K} p(b|z_i, x_G)p(z_i|x_G) \quad =$$

$$\sum_{i=1}^{K} p(b|z_i)p(z_i|x_G)$$

Figura 4



Modelos de árbol para reconocimiento de objetos. (a) Sólo un árbol, equivalente a la implementación en [16]. (b) Mezcla de árboles propuesta en nuestro trabajo.

Aquí podemos distinguir dos componentes correspondientes al modelo de la mezcla de expertos. La función de compuerta, que corresponde al término $p(z_i|x_C)$, indica la influencia de cada escena de acuerdo a la información global y está representada por una función de kernel Gaussiano [21]. La función de experto, que corresponde al término $p(b|z_i)$, representa la probabilidad de ocurrencia de los objetos de acuerdo a la escena. En nuestro caso, esto es representado por una red bayesiana. Este modelo es similar al modelo de mezcla de árboles de Meila y Jordan [22], sin embargo la diferencia es que su modelo usa un conjunto fijo de pesos, mientras que en nuestro caso los pesos se adaptan a la información de entrada.

Implementación

Para resolver los parámetros del modelo aplicamos el algoritmo EM [23]. Asumiendo conocida la responsabilidad de cada componente en los datos podemos maximizar (paso M) los parámetros de la compuerta, es decir, los pesos, medias y varianzas de los kernel Gaussianos, además de los parámetros de los expertos, es decir, la estructura y las probabilidades condicionales de cada red bayesiana. Para el caso de las redes bayesianas, se utiliza una versión modificada del algoritmo de Chow-Liu, que permite encontrar los parámetros relevantes de forma óptima para árboles [22]. En el caso del valor esperado de variables latentes (paso E) se obtiene:

$$p(z_i|x_n, b_n) =$$

$$\frac{p(z_i|x_n)p(b_n|z_i, x_n)}{p(b_n|x_n)} = \frac{p(z_i|x_n)p(b_n|z_i)}{\sum_{j=1}^{K} p(z_j|x_n)p(b_n|z_j)}$$

Para realizar la inferencia, seguimos el mismo proceso alternado de [16] para cada árbol y luego combinamos los scores usando los pesos de las compuertas.

En relación a los términos base de nuestro modelo, podemos ver que en general ellos dependen de datos de entrenamiento externos a la imagen de la detección actual, los que permiten aprender sobre distintos componentes que necesitamos para la aplicación de nuestro método. Por ejemplo, el término S_{ik} corresponde a la salida de un clasificador de objetos, entrenado con un conjunto de imágenes etiquetadas. Lo mismo sucede con el caso de los contextos a través de información global, el término que representa las confianzas de las clasificaciones C_{ik}, etc. Con esto podemos ver una vez más que la información de la imagen misma que está siendo analizada debe ser complementada con información externa, correspondiente a datos que entregan conocimiento acumulado sobre los objetos y escenas utilizadas, los que resultan esenciales para un buen rendimiento del método implementado.

> En este artículo hemos presentado dos trabajos realizados por nuestro Grupo de Investigación en Inteligencia de Máquina, GRIMA, en el área de reconocimiento visual. Estos muestran cómo el uso de técnicas de aprendizaje de máquina permite incorporar en el análisis de una imagen información que va más allá de ésta.

RESULTADOS

Para nuestro trabajo consideramos dos base de datos reales de objetos, las que denominamos OUTDOR y SUN09. OUTDOOR contiene 2.600 imágenes de ocho escenas distintas tales como costa, montaña, bosques, etc. En este set de datos consideramos 21 categorías de objetos. SUN09 contiene aproximadamente 8.500 imágenes. En este set de datos consideramos 111 categorías de objetos. Ambos sets de datos fueron divididos en partes iguales para el entrenamiento y el testeo. Como clasificador de objetos individuales utilizamos el detector de objetos de Felzenswalb et al.[3]. Se usa el promedio de la curva Precision-Recall (APR) [24] como métrica para la detección de objetos.

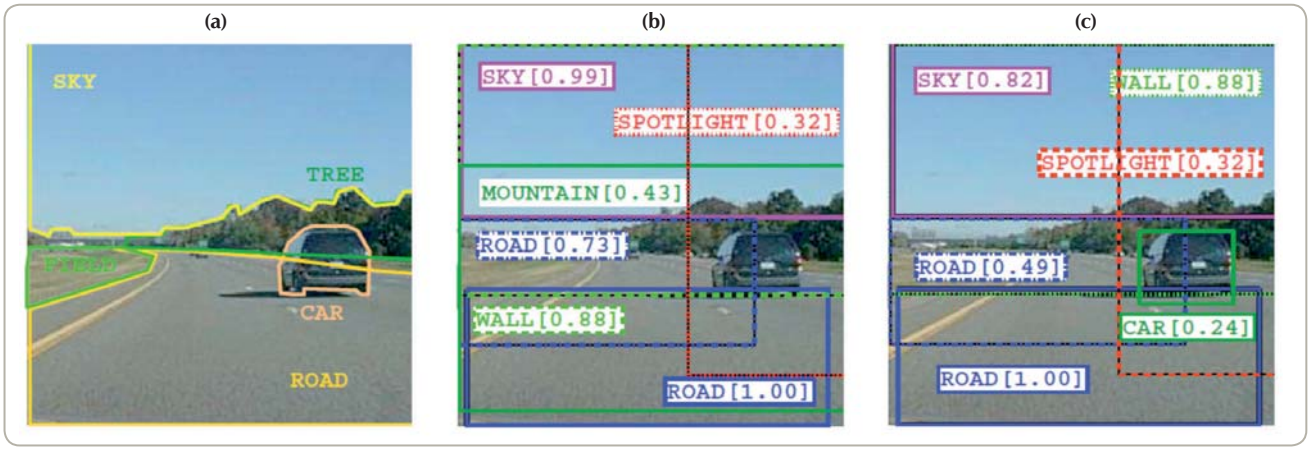
La Tabla 2 muestra una comparación entre los resultados utilizando el detector de objetos en [3] por sí sólo, el detector de un árbol único en [16] y nuestro método utilizando diversos números de árboles en la mezcla de expertos. Se puede apreciar que el mejor modelo resulta ser el de seis árboles para ambos sets de datos, con una mejora relativa de 5.5% y 5.7% respecto a un solo árbol para OUTDOOR y SUN09, respectivamente.

Tabla 2

Método	OUTDOOR	SUN09
Detector de Objetos en [3]	14.02 (6.5%)	6.82 (13.2%)
Método de un árbol en [16]	15.00 (0.0%)	7.87 (0.0%)
Mezcla de 2 árboles	15.07 (0.5%)	7.98 (1.5%)
Mezcla de 3 árboles	14.87 (-0.9%)	8.09 (2.9%)
Mezcla de 4 árboles	15.12 (0.8%)	8.06 (2.5%)
Mezcla de 5 árboles	15.25 (1.7%)	8.03 (2.2%)
Mezcla de 6 árboles	**15.83 (5.5%)**	**8.31 (5.7%)**
Mezcla de 7 árboles	14.84 (-1.1%)	7.88 (0.3%)

Resultados de APR sobre set de pruebas.

Figura 5



Comparación de resultados en una imagen de prueba. (a) Detecciones reales (ground-truth). (b) Método de un árbol en [16]. (c) Mezcla de árboles propuesta en nuestro trabajo.

La Figura 5 muestra un ejemplo de ejecución de nuestro método con seis árboles, en comparación con las detecciones reales (ground-truth) y con el método de un árbol en [16]. Se observa cómo el algoritmo es capaz de detectar un auto además de remover la falsa detección de montaña, mejorando lo entregado por el método de un árbol.

Como podemos ver, nuestro método presenta importantes ventajas en los resultados, las que de acuerdo a nuestra hipótesis, se deben al hecho de que incorporar contextos adaptivos ayuda a realizar una mejor detección al considerar la información más relevante en cada caso particular.

CONCLUSIONES

En este artículo hemos presentado dos trabajos realizados por nuestro Grupo de Investigación en Inteligencia de Máquina, GRIMA, en el área de reconocimiento visual. Éstos muestran cómo el uso de técnicas de aprendizaje de máquina permite incorporar en el análisis de una imagen información que va más allá de ésta, lo que se logra a través de aprender información relevante sobre objetos, escenas y las relaciones entre ellos desde miles de ejemplos de escenas cotidianas.

El método de reconocimiento de escenas a través de detección de objetos presentado muestra buenos resultados en el ambiente de oficinas probado, mostrando claras ventajas respecto a métodos que no utilizan objetos para el reconocimiento. Más aún, hemos probado este método en otro ambiente de interior, una casa, con distintos objetos y escenas, obteniendo resultados similares. Además, para aliviar la limitante de que una imagen particular de una escena no contiene todos los objetos presentes en ella, el método implementado fue probado sobre nuestro robot móvil utilizando una implementación secuencial que iba detectando objetos a medida que el robot se movía. Así, en el caso de la sala de conferencias, al entrar el robot detectó sólo el monitor y creyó que estaba en una oficina, sin embargo, gracias a la implementación probabilística utilizada, cuando posteriormente encontró la pantalla de proyector pudo actualizar sus creencias y definir que realmente estaba en una sala de conferencias.

Cabe destacar, que tal como se mencionó anteriormente, el método requiere de una gran capacidad computacional para ejecutar, lo que lo hace difícil de implementar para una operación en el mundo real. Dado esto, se implementó una aproximación a través de muestreo, utilizando el método de Monte Carlo, para poder realizar un eficiente cómputo del reconocimiento de escena.

En el caso del método de reconocimiento de objetos utilizando contextos jerárquicos adaptivos, se logró superar las limitaciones relevantes de un modelo de árbol de contexto fijo, a través de un modelo que utiliza una mezcla condicional de árboles para lograr resultados que se adapten a las distintas condiciones en las que pudo haber sido tomada la imagen. Nuestros experimentos usando distintos conjuntos de datos indican que el modelo propuesto mejora el rendimiento del reconocimiento de objetos con respecto a los métodos comparados, al considerar información de la escena subyacente que influye en las relaciones de objeto a objeto.

Cabe destacar que este método puede ser utilizado con distintos clasificadores de objetos, los cuales deberían mejorar su rendimiento debido a la inclusión de información adaptiva sobre el contexto. Así, por ejemplo, para alcanzar una buena escalabilidad al ejecutar el método con un gran conjunto de categorías de objetos, se puede incluir políticas adaptativas para

controlar la ejecución de los clasificadores de objetos, en forma similar al método propuesto en [25].

Más detalles de la implementación de cada uno de los métodos presentados pueden ser encontrados en [14] y [26].

AGRADECIMIENTOS

Este trabajo fue financiado en parte por Fondecyt, proyectos 1120720 y 1095140.BITS

REFERENCIAS

[1] P. Viola y M. Jones. "Rapid object detection using a boosted cascade of simple features". IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2001.

[2] J. Sivic y A. Zisserman. "Video Google: A Text Retrieval Approach to Object Matching in Videos". International Conference on Computer Vision (ICCV), 2003.

[3] P. Felzenszwalb, D. McAllester, y D. Ramanan. "A discriminatively trained, multiscale, deformable part model". IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008.

[4] Microsoft Corp. Redmond WA. Kinect for Xbox 360.

[5] Grupo de Inteligencia de Máquinas DCC PUC, GRIMA, http://grima.ing.puc.cl.

[6] C. Siagian y L. Itti. "Rapid biologically-inspired scene classification using features shared with visual attention". IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI), 29 (2), 300–312 , 2007.

[7] A. Oliva y A. Torralba. "Modeling the shape of the scene: a holistic representation of the spatial envelope". International Journal of Computer Vision (IJCV), 42, 145–175, 2001.

[8] L. Fei-Fei y P. Perona. "A bayesian hierarchical model for learning natural scene categories". IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2005.

[9] A. Bosch, A. Zisserman y X. Muñoz. "Scene classification via plsa". European Conference on Computer Vision (ECCV), 2006.

[10] J. Sivic, B. Russell, A. Efros, A. Zisserman, y W. Freeman. "Discovering objects and their localization in images". International Conference in Computer Vision (ICCV), 2005.

[11] S.Lazebnik, C. Schmid y J. Ponce. "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories". IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2006.

[12] A. Quattoni y A. Torralba. "Recognizing indoor scenes". IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009.

[13] L. Li, H. Su, E. Xing y L. Fei-Fei. "Object Bank: A High-Level Image Representation for Scene Classification and Semantic Feature Sparsification". Neural Information Processing Systems (NIPS), 2010.

[14] P. Espinace, T. Kollar, A. Soto y N. Roy. "Indoor Scene Recognition Through Object Detection". IEEE International Conference on Robotics and Automation (ICRA), 2010.

[15] N. Dalal y B. Triggs. "Histograms of oriented gradients for human detection". European Conference on Computer Vision (ECCV), 2005.

[16] M. Choi, J. Lim, A. Torralba, y A.Willsky. "Exploiting hierarchical context on a large database of object categories". IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010.

[17] C. Desai, D. Ramanan, and C. Fowlkes. "Discriminative models for multi-class object layout". International Journal of Computer Vision (IJCV), 95(1), 2011.

[18] A. Torralba, K. Murphy y W. Freeman. "Contextual models for object detection using boosted random fields". Advances in Neural Information Processing Systems (NIPS), 2005.

[19] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora y S. Belongie. "Objects in context". International Conference on Computer Vision (ICCV), 2007.

[20] R. Jacobs, M. Jordan, S. Nowlan y G. Hinton. "Adaptive mixtures of local experts". Neural Computation, 3, 79-87, 1991.

[21] L. Xu, M. Jordan y G. Hinton. "An alternative model for mixtures of experts". Advances in Neural Information Processing Systems (NIPS), 1994.

[22] M. Meila y M. Jordan. "Learning with mixtures of tres". Journal of Machine Learning., 1:1–48, 2001.

[23] A. Dempster, N. Laird y D. Rubin. "Maximum likelihood from incomplete data via the EM algorithm (with discussion)". Journal of the Royal Statistical Society, Series B, 39:1–38, 1977.

[24] J. Davis y M. Goadrich. "The relationship between precision-recall and roc curves". International Conference on Machine Learning, pages 233–240, ACM Press, 2006.

[25] P. Espinace, T. Kollar, A. Soto y N. Roy. "Indoor scene recognition through object detection using adaptive objects search". European Conference on Computer Vision (ECCV), Workshop on Robotics for Cognitive Tasks, 2010.

[26] B. Peralta, P. Espinace y A. Soto. "Adaptive hierarchical contexts for object recognition with conditional mixture of trees". British Machine Vision Conference (BMVC), 2012.