Modelado del Contexto Geométrico para el Reconocimiento de Objetos

Jose-Raul Ruiz-Sarmiento, Cipriano Galindo, Javier Gonzalez-Jimenez
Departamento de Ingeniería de Sistemas y Automática, Instituto de Investigación Biomédica de Málaga, Universidad de Málaga, Campus de Teatinos, 29071, Málaga
{*jotaraul,cgalindo,javiergonzalez*}*@uma.es*

Resumen

El reconocimiento de objetos es una tarea clave para dotar de cierta autonomía a un robot móvil. Los métodos de reconocimiento tradicionales han alcanzado un éxito aceptable empleando información sobre la apariencia y/o la geometría de los objetos, aunque pueden presentar resultados ambiguos. Persiguiendo mitigar esta desventaja, en este trabajo se estudia cómo modelar información sobre el contexto geométrico de los objetos, la cual resulta útil para inclinar la balanza en reconocimientos ambiguos, de tal manera que se alcance un reconocimiento tan exitoso como sea posible. Para ello hemos recurrido a los Campos Aleatorios Condicionales como herramienta de modelado, y a Robot@Home como conjunto de datos para la evaluación. Con estas premisas se han alcanzado conclusiones interesantes para cualquier sistema reconocedor empleando información contextual.

Palabras clave: Reconocimiento de objetos, contexto geométrico, campos aleatorios condicionales, robots de servicio.

1 INTRODUCCIÓN

Para que un robot móvil pueda prestar servicios con éxito en su lugar de trabajo necesita alcanzar un cierto grado de comprensión sobre su entorno. El reconocimiento de objetos es una tarea clave para ello, ya que permite al robot interactuar con los elementos detectados en su alrededor. Este reconocimiento ha de ser fiable, ya que una clasificación errónea puede comprometer la integridad del robot, de su entorno, o incluso de seres humanos. Para visualizar esto, supóngase un robot encargado de proveer medicación a una persona mayor, de regar las plantas, o de planchar la ropa.

Los métodos de reconocimiento tradicionales que reconocen individualmente cada objeto en el entorno han alcanzado un éxito notable [1, 2, 3]. No obstante, estos métodos pueden a menudo ofrecer resultados ambiguos que comprometen la operación del robot, *p.e.* un objeto cilíndrico de tamaño medio podría ser reconocido como una papelera, un jarrón, una botella, etc [4]. Una manera de eliminar estas ambigüedades es la utilización de información sobre el contexto geométrico de los objetos: si hay una flor encima, probablemente sea un jarrón [5]. Esta fuente de información permite analizar las relaciones entre los objetos, y es de gran valor a la hora de reconocerlos.

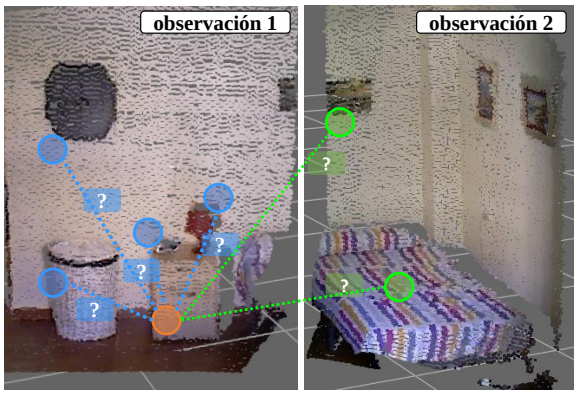

Figura 1: Ejemplo del problema del modelado contextual. En el caso de la mesa de noche (círculo naranja), sus relaciones de contexto (líneas azules y verdes) van a depender del rango de contexto empleado, y de la extensión de la información contextual a modelar.

Los *Campos Aleatorios Condicionales*, del inglés *Conditional Random Fields* (CRFs) [6], son una herramienta comúnmente usada para modelar y explotar información contextual. Estos modelos se basan en una representación en forma de grafo, donde los nodos son interpretados como objetos, y los arcos conectan nodos/objetos con una relación de contexto. Para crear dichos arcos se suele establecer un *rango de contexto*, que fija la distancia máxima a la que dos objetos pueden estar situados en el entorno para considerarse que existe una relación entre ellos. Por ejemplo, las relaciones que se muestran como líneas azules en la Fig. 1 se establecerán dependiendo de este rango de contexto. En la literatura también pueden encontrarse trabajos que consideran distintas fuentes de información contextual a modelar: proveniente de una observación del entorno (imagen de intensidad, RGB-D, etc.), donde esta información puede

ser escasa o pobre, o de una reconstrucción del mismo, lo que proporciona una mayor extensión de la información contextual (por ejemplo, las relaciones que aporta la imagen de la derecha en la Fig. 1, representada como líneas verdes). La elección del rango de contexto, o de la extensión de la información contextual, suelen hacerse de manera *ad-hoc* sin tener en cuenta sus posibles efectos en el reconocimiento.

Este trabajo persigue proveer indicaciones útiles y buenas prácticas sobre el modelado de información contextual, de tal manera que cualquier sistema reconocedor explotando esta fuente de información pueda alcanzar unos resultados tan exitosos como sea posible. Para ello se utilizan los CRFs como herramienta para el modelado y aprovechamiento del contexto, y se estudian principalmente los dos factores anteriormente citados, la elección i) del rango de contexto, y ii) de la extensión de la información a modelar. En dicho estudio se analiza la influencia de estos factores tanto en el éxito del reconocimiento, como en los tiempos de ejecución necesarios para los procesos de entrenamiento e inferencia sobre los CRFs, de tal manera que se pueda seleccionar la configuración que más se ajuste a las necesidades de cada aplicación. Para llevar a cabo los experimentos realizados durante el estudio se ha empleado el conjunto de datos *Robot@Home* [7], dada su complejidad y adecuación al problema: fue recogido por un robot móvil en entornos domésticos.

2 TRABAJOS RELACIONADOS

Los métodos de reconocimiento tradicionales han tenido un éxito notable en aplicaciones donde se especializan en detectar un cierto tipo de objeto (e.g. caras humanas [1]) o donde los objetos a reconocer aparecen aislados [2]. Ejemplos de estos métodos son los que emplean descriptores de la imagen como *Scale-Invariant Feature Transform* (SIFT) [8] o *Speeded-Up Robust Features* (SURF) [9], los cuales son explotados por clasificadores como las *Supported Vector Machines* (SVMs) [10] o las *Bag-of-Words* (BoW) [3]. No obstante, su rendimiento tiende a bajar en situaciones donde el número de posibles categorías a reconocer es elevado, o donde los objetos aparecen en escenas pobladas con múltiples objetos en diversas localizaciones y configuraciones, como es el caso de entornos humanos (oficinas, hogares, etc.) [4]. Uno de los principales motivos detrás de esta caída de rendimiento es la aparición de resultados ambiguos. No obstante, este fenómeno se puede paliar con la utilización de información sobre el contexto geométrico de los objetos [5].

Los *Modelos Gráficos Probabilísticos*, del inglés *Probabilistic Graphical Models* (PGMs) [6], son utilizados en multitud de trabajos para modelar y explotar eficientemente dicho contexto. Para el caso del reconocimiento de objetos, los *Campos Aleatorios Condicionales* (del inglés *Conditional Random Fields*, CRFs), un tipo particular de PGM, han resultad especialmente exitosos. Estos modelos fueron empleados, por ejemplo, por Xiong y Huber [11] para el reconocimiento de los componentes básicos de un edificio: pared, suelo, techo, etc. Estos autores relacionan cada objeto con los k objetos más cercanos, sin importar la distancia a la que se encuentren, enfoque que puede dar lugar a relaciones poco relevantes o inexistentes. Por su parte, el CRF diseñado por Rogers y Christensen [12] incluye relaciones entre los objetos y las habitaciones donde se encuentran, pero no entre los propios objetos, desaprovechando una valiosa fuente de información. Otro trabajo relevante es el de Lin, Fidler y Urtasun [13], donde los objetos son representados por sus cajas delimitadoras, y se considera que están relacionados si estas cajas se encuentran a una distancia menor de 50 centímetros. Los autores del presente estudio también presentaron trabajos previos donde se empleó dicho rango de contexto (*p.e.* [14, 15]).

Quizás el trabajo más relacionado con el nuestro es el de Anand *et al.* [16], donde se usa un Campo Aleatorio de Markov (variante discriminativa de los CRFs) para reconocer objetos en entornos de oficinas y domésticos. En él se realiza un estudio superficial de la influencia del rango de contexto y de la extensión de la información contextual. En este trabajo se realiza un análisis más profundo de ambos factores, y también se estudia su repercusión en los tiempos de ejecución de los procesos de entrenamiento e inferencia de los CRFs.

3 LA HERRAMIENTA: CRFs APLICADOS AL RECONOCIMIENTO

La tarea del reconocimiento de objetos en una escena consiste en asignar categorías de un conjunto $\mathcal{L}$ (*p.e.* mesa, maceta, cortina, cuadro, vaso, etc.) a las observaciones de los n objetos en la misma $\boldsymbol{x} = [x_1, \cdots, x_n]$. Si se considera $\boldsymbol{y} = [y_1, \cdots, y_n]$ como el vector de variables aleatorias que asignan a cada objeto en $\boldsymbol{x}$ una categoría de $\mathcal{L}$, el problema del reconocimiento desde un punto de vista probabilístico se define como la búsqueda de la asignación a $\boldsymbol{y}$ que maximiza la distribución de probabilidad condicionada $p(\boldsymbol{y}|\boldsymbol{x})$. Dada su complejidad, la definición exhaustiva de esta distribución no es factible. Es en este punto donde los CRFs nos ofrecen la posibilidad de representarla de tal manera que su computo pueda ser más eficiente.

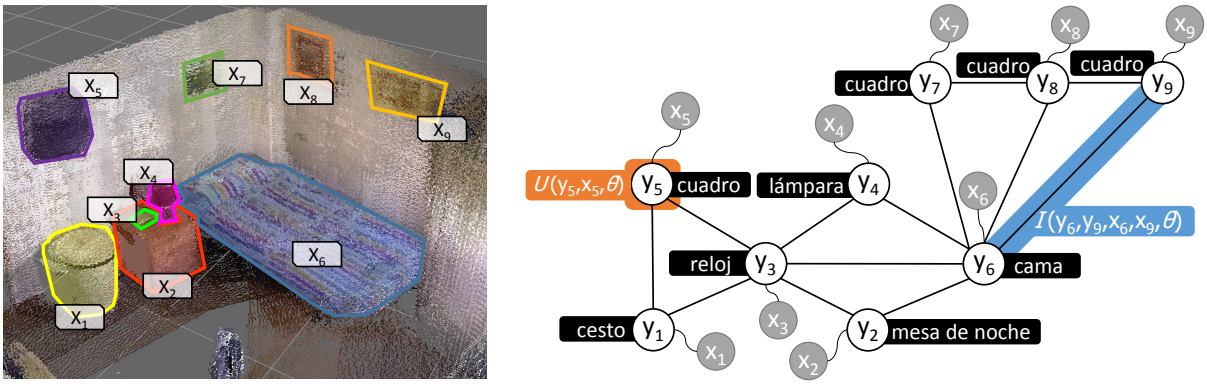

Figura 2: Izquierda, reconstrucción de una habitación con objetos segmentados $(x_1, \cdots, x_9)$. Derecha, representación mediante un CRF en forma de grafo de dicha escena, donde aparece una variable aleatoria/nodo por cada objeto, y los objetos relacionados se conectan con un arco. Las formas naranjas representan el ámbito de un factor local, las azules de un factor por pares, y las negras son el resultado de un proceso de inferencia sobre el CRF.

Para ello, los CRFs emplean una representación en forma de grafo $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, donde los nodos en $\mathcal{V}$ se asocian con las variables aleatorias en $\boldsymbol{y}$, y los arcos en $\mathcal{E}$ conectan nodos que guardan algún tipo de relación. En el caso del reconocimiento de objetos, los nodos se conectan acorde al contexto geométrico de sus objetos asociados. Así si dos objetos están situados *cerca* en la escena, se entiende que el reconocimiento de uno tiene influencia directa (y mutua) en la categorización del otro, conectándose sus nodos, mientras que si están alejados o en distintas habitaciones esta influencia no es tal. El cómo decidir si dos objetos están cerca es motivo de discusión en la Sec. 4.1.

Una vez construido el grafo $\mathcal{G}$ que representa los objetos en el entorno del robot, la probabilidad $p(\boldsymbol{y}|\boldsymbol{x})$ se codifica sobre el mismo empleando el concepto de *factor*. Un factor puede interpretarse como una función definida sobre parte del grafo que codifica un pedazo de dicha probabilidad, siendo típicamente de dos tipos: *locales* y *por pares*. Los factores locales se refieren a un nodo del grafo, y establecen como de probable es para una variable aleatoria y_i el pertenecer a una categoría de $\mathcal{L}$ de acuerdo a las características visuales y/o geométricas del objeto x_i. Por su parte, los factores por pares se definen sobre arcos, y determinan la compatibilidad de asignar dos categorías de $\mathcal{L}$ a dos variables relacionadas y_i y y_j teniendo en cuenta x_i y x_j. Estos factores suelen modelarse como clasificadores lineales de la siguiente forma:

$$\mathcal{U}(y_i, x_i, \boldsymbol{\theta}) = \sum_{l \in \mathcal{L}} \delta_{y_i = l} \boldsymbol{\theta}_l \boldsymbol{f}_{x_i} \quad (1)$$

$$\mathcal{I}(y_i, y_j, x_i, x_j, \boldsymbol{\theta}) = \sum_{l_1 \in \mathcal{L}} \sum_{l_2 \in \mathcal{L}} \delta_{y_i = l_1} \delta_{y_j = l_2} \boldsymbol{\theta} \boldsymbol{f}_{x_i x_j} \quad (2)$$

siendo $\mathcal{U}(\cdot)$ un factor local definido sobre el nodo asociado a y_i, y $\mathcal{I}(\cdot)$ un factor por pares sobre el arco que conecta y_i y y_j. En estas ecuaciones δ es la función delta de Kronecker que toma el valor 1 si $y_i = l$, y 0 si $y_i \neq l$, $\boldsymbol{\theta}$ es un vector de pesos o parámetros aprendido durante la fase de entrenamiento del CRF, y $\boldsymbol{f}_{x_i}$ y $\boldsymbol{f}_{x_i x_j}$ son vectores de características extraídas de los objetos (color, tamaño, forma, etc.) y de sus relaciones de contexto (distancia, diferencia en altura, ratio de tamaño, etc.) respectivamente.

Una vez definidas las piezas que componen un CRF, y deacuerdo con el teorema de Hammersley-Clifford, la función de probabilidad $P(\boldsymbol{y}|\boldsymbol{x})$ puede ser finalmente factorizada sobre el grafo $\mathcal{G}$ empleando modelos log-lineales como:

$$P(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{x}, \boldsymbol{\theta})} \prod_{i \in \mathcal{V}} \exp(\mathcal{U}(y_i, x_i, \boldsymbol{\theta})) \prod_{(i,j) \in \mathcal{V}} \exp(\mathcal{I}(y_i, y_j, x_i, x_j, \boldsymbol{\theta})) \quad (3)$$

La esencia de esta representación es que, al elevar al exponente los factores, el resultado es siempre un valor mayor que 0, requisito básico para que el problema pueda modelarse mediante un CRF. Por su parte, $Z(\cdot)$ (también llamada función de partición) normaliza los factores para que el resultado sea una distribución de probabilidad, esto es $\sum_{\xi(\boldsymbol{y})} P(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}) = 1$, siendo $\xi(\boldsymbol{y})$ una asignación posible a las variables en $\boldsymbol{y}$.

Para conseguir los resultados de reconocimiento hay que realizar un proceso de inferencia sobre el grafo $\mathcal{G}$, el cual nos permite obtener la asignación más probable $\hat{\boldsymbol{y}}$ a las variables en $\boldsymbol{y}$, esto es:

$$\hat{\boldsymbol{y}} = arg \max_{\boldsymbol{y}} P(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}) \quad (4)$$

Esta inferencia suele realizarse mediante métodos aproximados, ya que su cómputo requiere del cal-

culo de la función de partición $Z(\cdot)$, que suele resultar impracticable en problemas reales. En el estudio realizado en este trabajo se ha empleado el método *Loopy Belief Propagation* (LBP), dado su buen rendimiento [17].

4 MODELADO DEL CONTEXTO GEOMÉTRICO

Como se ha comentado, el modelado de la información contextual se basa en decidir qué objetos del entorno están relacionados entre sí, esto es, que nodos se conectan mediante un arco. Para tomar esta decisión se suele emplear un rango de contexto (distancia máxima a la que se considera que dos objetos están relacionados), cuya elección está estrechamente ligada al tipo de la información a modelar. Los trabajos más notorios en la literatura suelen usar observaciones de la escena proveyendo información de intensidad (imagen RGB) o de intensidad y profundidad (imágenes RGB-D).

En el caso de imágenes de intensidad, el rango de contexto puede establecerse en el plano de la imagen a nivel de pixel o super-pixel. Así, los pixeles o super-pixels que guardan relación contextual son los que aparecen colindantes en la imagen. Con este enfoque no se respeta la geometría de la escena, por lo que, por ejemplo, una región correspondiente con un objeto cercano a la cámara podría conectarse con otro lejano. También hay trabajos que realizan una reconstrucción tridimensional de la escena, bien buscando puntos de fuga, con imágenes estéreo, etc., la cual permite medir distancias geométricas entre los objetos para establecer su contexto.

Por su parte, las imágenes RGB-D ya proporcionan dicha información tridimensional, por lo que son aptas para realizar medidas geométricas. Este es el tipo de imágenes utilizadas en este trabajo, discutiéndose a continuación las distintas maneras de realizar mediciones en las mismas.

4.1 MEDICIÓN DE LA DISTANCIA ENTRE OBJETOS

Una vez contamos con una imagen RGB-D de una escena, y considerando una representación en forma de nube de puntos $\boldsymbol{pc} = [p_1, \cdots, p_m]$ donde $p_i = [x, y, z, r, g, b]$ (información geométrica y de color), el primer paso para establecer las relaciones de contexto es segmentar los objetos $\boldsymbol{o} = [o_1, \cdots, o_n]$ que aparecen en la misma. Una vez segmentados, cada objeto se corresponderá con una región de la nube $o_i = \boldsymbol{pc}_i$, $\boldsymbol{pc}_i \subseteq \boldsymbol{pc}$. A continuación se discuten las opciones más relevantes para calcular la distancias entre regiones.

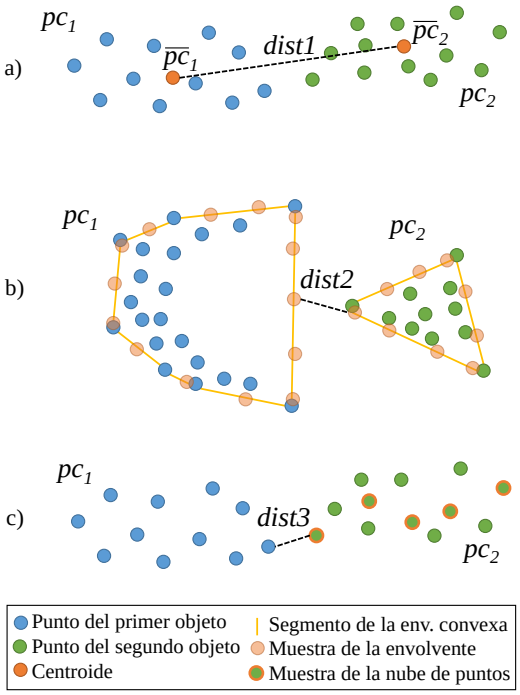

Figura 3: Ejemplos de nubes de puntos pertenecientes a dos objetos en 2 dimensiones, y 3 maneras de calcular la distancia entre ellos.

Entre los más populares, el primer enfoque para calcular la distancia entre $\boldsymbol{pc}_i$ y $\boldsymbol{pc}_j$ consiste en calcular sus centroides $\overline{\boldsymbol{pc}}_i$ y $\overline{\boldsymbol{pc}}_j$ y obtener la distancia euclidea entre ambos, esto es:

$$dist1(\boldsymbol{pc}_i, \boldsymbol{pc}_j) = \sqrt{\sum_{e=1}^{3} (\overline{\boldsymbol{pc}}_{ie} - \overline{\boldsymbol{pc}}_{je})} \qquad (5)$$

Aunque el cálculo de esta distancia es rápido, y el cómputo de los centroides puede aprovecharse para describir los objetos, tiene el problema de que pares de objetos grandes y próximos pueden presentar una distancia superior al rango de contexto y no ser conectados en el CRF (ver Fig. 3-a).

Una segunda opción es el cómputo de las envolventes convexas $C(\boldsymbol{pc}_i)$ y $C(\boldsymbol{pc}_j)$ que encierran a todos los puntos de cada objeto. Tomando muestras de dicha envolvente se pueden obtener una serie de puntos $\boldsymbol{c_i}$ y $\boldsymbol{c_j}$ sobre las que realizar mediciones de distancia euclídea, siendo la distancia entre dos objetos la menor de estas mediciones, es decir:

$$dist2(\boldsymbol{pc}_i, \boldsymbol{pc}_j) = \min_{(p_i \in \boldsymbol{c_i}, p_j \in \boldsymbol{c_j})} \sqrt{\sum_{e=1}^{3} (\boldsymbol{p}_{ie} - \boldsymbol{p}_{je})} \qquad (6)$$

El usar esta distancia tiene el inconveniente de que un objeto con partes cóncavas podría generar mediciones de distancia irreales, dando lugar a relaciones de contexto erróneas (ver Fig. 3-b).

Por último, un tercer enfoque consiste en usar la fuerza bruta para computar la distancia mínima entre cada par de puntos de dos objetos, lo cual soluciona los problemas presentados por las opciones anteriores. No obstante, este proceso es altamente ineficiente y haría impracticable el reconocimiento. Una alternativa viable es la de construir una representación de la nube de puntos que permita realizar búsquedas de puntos cercanos de manera eficiente, como es el caso de los árboles kd [18]. Este es el enfoque usado en este trabajo, concretamente, se construye el árbol kd de $\boldsymbol{pc}_i$, $kdtree_{\boldsymbol{pc}_i}$ (complejidad $O(n \log n)$, siendo n el número de puntos), y se muestrea $\boldsymbol{pc}_j$ para obtener una serie de puntos $\boldsymbol{m_j}$. De manera eficiente (complejidad $O(\log n)$) se busca el punto más cercano en el árbol a cada punto de $\boldsymbol{m_j}$, siendo la distancia entre los objetos el par más cercano (ver Fig. 3-c). Esto es:

$$dist3(\boldsymbol{pc}_i, \boldsymbol{pc}_j) = \min_{(p_i \in \boldsymbol{pc}_i, p_j \in \boldsymbol{m_j})} dist(kdtree_{\boldsymbol{pc}_i}, p_j) \tag{7}$$

Como se ha comentado, una vez calculada la distancia entre dos objetos, se usa el rango de contexto fijado (*p.e.* un metro, dos, etc.) para decidir si existe o no una relación de contexto geométrico entre ambos. La elección del rango de contexto es clave para poder sacar el máximo partido a estas relaciones, y es estudiado en profundidad en la Sec. 5. Con un rango corto se establecerían pocas relaciones y podría descartarse información contextual valiosa, pero los procesos de entrenamiento e inferencia sobre CRFs serían rápido. Por otro lado, un rango grande consideraría una mayor porción de dicha información, pero podría perjudicar a los tiempos de entrenamiento e inferencia y aumentar su complejidad. El rango de contexto se ve influenciado por un factor adicional: la extensión de la información contextual contenida en la imagen, tal y como se comenta en la siguiente sección.

4.2 EXTENSIÓN DE LA INFORMACIÓN CONTEXTUAL

Tanto cuando se usan imágenes de intensidad como RGB-D, para sacarle el máximo partido a la información contextual es necesario que en la observación aparezca la mayor porción de la escena posible. De no ser así, esta información puede resultar escasa e incompleta en algunos casos, siendo de poca utilidad. Por ejemplo, en la Fig. 4 se muestran a la izquierda dos observaciones con información contextual limitada, mientras que en las de la derecha la extensión de esta es mucho mayor.

Una manera de extender la información contextual a modelar es considerar una porción de la escena mayor que la proporcionada por una simple imagen. Para ello se hace necesario propagar en el tiempo y el espacio la información en cada observación mediante algún algoritmo de registro o reconstrucción. A pesar de lo interesante de su uso desde el punto de vista del aprovechamiento del contexto, el reconstruir una escena puede acarrear una serie de problemas adicionales fuente de errores en el sistema reconocedor. Por ejemplo, un mal registro de dos imágenes puede hacer que los objetos aparezcan dobles o deformes. Además, en aplicaciones donde se requiera que el reconocimiento de objetos funcione a una cierta frecuencia, el algoritmo de reconstrucción ha de ser suficientemente rápido para soportar dicha frecuencia. Aunque el análisis de distintos métodos de reconstrucción está fuera del alcance de este artículo, si es relevante el efecto de contar con distintas extensiones de la información contextual en el reconocedor, factor que se analiza en la siguiente sección.

Figura 4: A la izquierda, nubes de puntos de una cocina y un cuarto de baño con información contextual limitada. A la derecha, nubes de las mismas habitaciones donde la información contextual más extensa.

5 ESTUDIO Y RESULTADOS

En este apartado se introducen las herramientas y equipos empleados (Sec. 5.1) en el análisis del rango de contexto (Sec. 5.2) y la extensión de la información contextual (Sec. 5.3), así como los resultados que se desprenden del estudio realizado.

5.1 HERRAMIENTAS EMPLEADAS

Para el modelado, entrenamiento e inferencia de los CRFs en este trabajo se ha empleado la librería *Undirected Probabilistic Graphical Models in C++*

Tabla 1: Influencia de la utilización de distintos rangos de contexto sobre el número de relaciones contextuales establecidas, los tiempos necesarios para el entrenamiento y la inferencia de los CRFs usados, y el éxito en el reconocimiento de estos.

Rango	**# de relaciones**	**Tpo. entrenamiento**	**Tpo. inferencia**	**Éxito**
$0m$	0 (0%)	$2.02s$	$0.01ms$	64.92%
$0.5m$	631 (14%)	$17.48s$	$0.15ms$	70.54%
$1m$	1,379 (31%)	$29.10s$	$0.36ms$	71.17%
$1.5m$	2,107 (48%)	$35.76s$	$0.69ms$	72.39%
$2m$	2,917 (66%)	$37.14s$	$1.04ms$	73.51%
$3m$	3,805 (86%)	$32.50s$	$1.81ms$	70.29%
$4m$	4,248 (96%)	$27.40s$	$2.37ms$	69.26%
$5m$	4,387 (99%)	$19.57s$	$2.84ms$	68.18%
$6m$	4,410 (100%)	$18.09s$	$3.23ms$	67.86%

(UPGMpp) [19], un software libre especialmente desarrollado para facilitar la utilización de estos modelos en el reconocimiento de objetos.

Por otra parte, para el análisis de las distintas opciones de modelado contextual se ha contado con el conjunto de datos *Robot@Home* [7]. Este repositorio contiene más de 69,000 imágenes RGB-D capturadas por medio de un robot móvil en entornos domésticos reales, donde aparecen 157 categorías de objetos etiquetadas. De entre ellas, en este trabajo se han seleccionado para ser reconocidas las 19 más comunes, sumando un total de ∼600 instancias de objetos.

Para evaluar el éxito en el reconocimiento se ha empleado validación cruzada. En cada paso de este método se emplean las observaciones provenientes de una habitación elegida al azar para evaluar, y las 31 restantes para entrenar. Esto se repite 1,000 veces cambiando la habitación con la que evaluar, y los resultados son promediados.

Las pruebas se realizaron en un ordenador con un microprocesador Intel Core i7-3820 a 3.60GHz. y una memoria RAM de 4x4GB. DDR3 a 1,600MHz.

5.2 INFLUENCIA DEL RANGO DE CONTEXTO

Con el fin de medir la influencia del rango de contexto en el rendimiento del sistema de reconocimiento, se han usado las reconstrucciones de las 32 habitaciones comentadas (la Fig. 4 muestra a la derecha dos de ellas). Estas reconstrucciones tienen la forma de nubes de puntos con información geométrica y de apariencia (intensidad). La Tab. 1 muestra los resultados del estudio llevado a cabo, donde la primera fila se corresponde con un CRF que no emplea información contextual, mientras que el resto reportan el rendimiento de CRFs que usan esta información con distintos rangos de contexto. Se puede apreciar como el éxito en el reconocimiento siempre es mayor en las configuraciones que explotan relaciones sin importar el rango elegido.

En cuanto al número de relaciones consideradas por cada opción, empleando un rango de contexto de 0.5 metros se explotan el 14% de ellas (4,410 existentes), mientras que hay que irse hasta una distancia de 6 metros para que se incluyan todas. Desde los 0 hasta los 2 metros, el incremento del rango de contexto acarrea un aumento en el éxito del reconocedor, alcanzándose con el último un $\sim$ 73.5% (un $\sim$ 8.5% más que sin emplear contexto). Esto se debe a que, conforme aumenta el rango, entran en consideración relaciones que tienden a cumplirse aunque no siempre presenten distancias cortas. Por ejemplo, en una cocina pueden aparecer un grifo y una placa de inducción típicamente a una distancia superior a medio metro. No obstante, hay un punto a partir del cual el incremento de este rango tiene un efecto negativo en el éxito, dada la alta variabilidad de las relaciones y la aparición de otras que rara vez se cumplen. En el trabajo de Anand *et al.* [16] este punto se sitúa en 0.6 metros en entornos domésticos. Esta considerable diferencia se debe probablemente a la extensión de la información contextual usada, como veremos en el siguiente apartado.

El incremento del rango de contexto también conlleva un mayor tiempo de ejecución del algoritmo de inferencia, que va desde los 0.15ms. con medio metro, hasta los 3.23ms. con 6 metros, situándose en 1.04ms. para el rango con el que se alcanza el mayor éxito. Esto pone de manifiesto que una elección arbitraría del rango puede resultar en un rendimiento no óptimo del reconocedor.

Un hecho curioso a primera vista es la evolución del tiempo de entrenamiento. Al ser un proceso iterativo (*Stochastic Gradient Descent*, más información en [17]), el añadir más carga computa-

Tabla 2: Influencia que tiene la utilización de distintos rangos de contexto sobre el número de relaciones contextuales establecidas, los tiempos necesarios para el entrenamiento y la inferencia de CRFs, y el éxito en el reconocimiento.

Rango	**# relaciones**	**Éxito**
$0.5m$	235	70.63%
$1m$	446	70.21%
$1.5m$	650	69.71%
$2m$	757	69.69%
$3m$	854	69.49%
$4m$	872	69.44%
$5m$	908	69.32%
$6m$	918	69.07%

cional por iteración al considerar más relaciones contextuales hace más costosa la fase de entrenamiento. Así ocurre hasta los 2 metros, pero a partir de ahí el tiempo necesario para entrenar baja. Esto se debe a la aparición de relaciones espurias, que impiden al proceso converger a modelos más exactos, resultando en un tiempo de ejecución menor. En cualquier caso, los tiempos de entrenamiento son comedidos para un proceso que solo ha de ejecutarse una vez.

5.3 REPERCUSIÓN DE LA EXTENSIÓN DE LA INFORMACIÓN CONTEXTUAL

Para analizar como afectan distintas extensiones de la información contextual al éxito del reconocimiento se han entrenado y evaluado CRFs con imágenes RGB-D individuales de las 32 habitaciones (la Fig. 4 muestra a la izquierda dos nubes de puntos formadas a partir de estas imágenes). Estos CRFs se pueden comprar con los de la sección anterior, donde la extensión de esta información era más amplia. Para que la comparativa fuera lo más justa posible, de nuevo se ha empleado un método de validación cruzada, pero en esta ocasión una imagen de una habitación es escogida para evaluar, mientras que 31 imágenes del resto de habitaciones se usan para entrenar (estas selecciones se hacen todas al azar). El proceso se repite mil veces, y se promedian los resultados.

La Tab. 2 muestra los resultados obtenidos. Como se puede ver, el mayor éxito se alcanza para un rango de contexto de medio metro (∼70.5%), rango similar al óptimo alcanzado en [16], lo que hace pensar que la extensión de la información usada por Anand *et al.* era limitada. A partir de esa distancia, el éxito decrece paulatinamente hasta el ∼ 69% obtenido con un rango de 6 metros. Esto se debe a que al considerar mayores rangos manteniendo una extensión de la información contextual baja, aparecen relaciones en las imágenes que no se repiten a lo largo del conjunto de datos de entrenamiento, dificultando el ajuste de los CRFs. Así, la mejor configuración empleando imágenes individuales alcanza un éxito 3 puntos porcentuales menor que empleando una extensión más amplia (∼70.5% vs. ∼73.5%).

En lo referente al número de relaciones con las que se trabaja, este también es menor, tal y como muestra la segunda columna de la tabla. Por ejemplo, con un rango de 6 metros se incluyen 918, por las 4,410 de la sección anterior. Por otro lado, el tiempo de ejecución del algoritmo de inferencia se mantiene estable y por debajo de los 0.3ms., mientras que el de entrenamiento va desde los 8s. con un rango de 0.5m. hasta los 17s. con 3m., distancia a partir de la cual se mantiene constante.

6 CONCLUSIONES

En este trabajo se ha estudiado como influyen distintas opciones de modelado del contexto geométrico en el rendimiento de sistemas basados en Campos Aleatorios Condicionales (del inglés *Conditional Random Fields*, CRFs) para el reconocimiento de objetos por parte de un robot móvil. En concreto, se ha analizado como afecta la utilización de distintos rangos de contexto, esto es, distancias máximas a las que se considera que dos objetos están relacionados, y de distintas extensiones de la información contextual: información proveniente de una imagen individual de la escena, o de una reconstrucción de la misma.

En casos donde la información contextual es extensa (*p.e.* empleando reconstrucciones de la escena), el análisis realizado con el conjunto de datos *Robot@Home* reporta el beneficio de emplear un rango de información contextual de 2 metros, consiguiendo un éxito del ∼ 73.5%, un tiempo de inferencia de 1.04ms, y un tiempo de entrenamiento de 37.14s. Para rangos menores, aunque los tiempos de entrenamiento e inferencia decrecen, también lo hace el éxito alcanzado. En cambio, para rangos mayores, el tiempo de inferencia aumenta a la vez que desciende el éxito reportado. Este es un efecto poco deseable que pone de manifiesto la necesidad del estudio completado para fijar un rango óptimo.

Por otra parte, cuando la extensión de la información contextual no es extensa (*p.e.* trabajando con imágenes individuales), en el caso del conjunto de datos empleado los mejores resultados se consigue con un rango de 0.5 metros. En lo referente al éxito alcanzado, este es un 3% menor que empleando información contextual extensa,

aunque con tiempos de entrenamiento e inferencia también más bajos. Esto muestra la estrecha relación que existe entre el rango de contexto y la extensión de esta información, siendo necesario en cada aplicación particular adaptar el primero conforme a la amplitud del segundo.

En un futuro se plantea el estudio de como podría influir en el reconocimiento la utilización de rangos de contexto dinámicos, que se ajustaran automáticamente dependiendo de la información disponible sobre la escena.

Agradecimientos

Este trabajo se ha desarrollado en el marco de los proyectos TEP2012-530 y DPI2014-55826-R, financiados por la Junta de Andalucía y el Ministerio de Ciencia e Innovación respectivamente, ambos contando con fondos del Fondo Europeo de Desarrollo Regional (FEDER).

Referencias

[1] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001)*, volume 1, pages 511–518, 2001.

[2] Jianguo Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. In *2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06)*, pages 13–13, June 2006.

[3] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 2161–2168, 2006.

[4] M. Oliveira, L. Seabra Lopes, G. H. Lim, S. H. Kasaei, A. D. Sappa, and A. M. Tomé. Concurrent learning of visual codebooks and object categories in open-ended domains. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2488–2495, Sept 2015.

[5] Carolina Galleguillos and Serge Belongie. Context based object categorization: A critical survey. *Computer Vision and Image Understanding*, 114(6):712–722, June 2010.

[6] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

[7] J. R. Ruiz-Sarmiento, C. Galindo, and J. González-Jiménez. Robot@home, a robotic dataset for semantic mapping of home environments. *The International Journal of Robotics Research*, 36(2):131–141, 2017.

[8] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, November 2004.

[9] Jan Knopp, Mukta Prasad, Geert Willems, Radu Timofte, and Luc Van Gool. Hough transform and 3d surf for robust three dimensional classification. In *Proceedings of the 11th European Conference on Computer Vision: Part VI*, ECCV'10, pages 589–602, Berlin, Heidelberg, 2010. Springer-Verlag.

[10] M. Pontil and A. Verri. Support vector machines for 3d object recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(6):637–646, Jun 1998.

[11] Xuehan Xiong and Daniel Huber. Using context to create semantic 3d models of indoor environments. In *In Proceedings of the British Machine Vision Conference (BMVC 2010)*, pages 45.1–11, 2010.

[12] J. G. Rogers and H. I. Christensen. A conditional random field model for place and object classification. In *Robotics and Automation (ICRA), 2012 IEEE International Conference on*, pages 1766–1772, May 2012.

[13] Dahua Lin, Sanja Fidler, and Raquel Urtasun. Holistic scene understanding for 3d object detection with rgbd cameras. *IEEE International Conference on Computer Vision*, 0:1417–1424, 2013.

[14] J. R. Ruiz-Sarmiento, C. Galindo, and J. González-Jiménez. Building multiversal semantic maps for mobile robot operation. *Knowledge-Based Systems*, 119:257 – 272, 2017.

[15] J. R. Ruiz-Sarmiento, M. Günther, C. Galindo, J. González-Jiménez, and J. Hertzberg. Online context-based object recognition for mobile robots. In *17th International Conference on Autonomous Robot Systems and Competitions (ICARSC)*. IEEE, April 2017.

[16] Abhishek Anand, Hema Swetha Koppula, Thorsten Joachims, and Ashutosh Saxena. Contextually guided semantic labeling and search for three-dimensional point clouds. *In The International Journal of Robotics Research*, 32(1):19–34, January 2013.

[17] J. R. Ruiz-Sarmiento, C. Galindo, and J. González-Jiménez. A survey on learning approaches for probabilistic graphical models. application to scene object recognition. *International Journal of Approximate Reasoning*, 83(C):434–451, April 2017.

[18] Jerome H. Friedman, Jon Louis Bentley, and Raphael Ari Finkel. An algorithm for finding best matches in logarithmic expected time. *ACM Trans. Math. Softw.*, 3(3):209–226, September 1977.

[19] J.R. Ruiz-Sarmiento, C. Galindo, and J. González-Jiménez. UPGMpp: a Software Library for Contextual Object Recognition. In *3rd. Workshop on Recognition and Action for Scene Understanding*, 2015.