





RGB-D Scene Recognition with Object-to-Object Relation

Xinhang Song^{1,2}, Chengpeng Chen^{1,2}, Shuqiang Jiang^{1,2} ¹Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, CAS, China ²University of Chinese Academy of Sciences, China xinhang.song@vipl.ict.ac.cn

Introduction

> Observations

- Objects are helpful to recognize scenes
- Object co-occurrences may confuse the scene recognition
- **□** RGB-D data is helpful to capture the spatial information

>Motivation

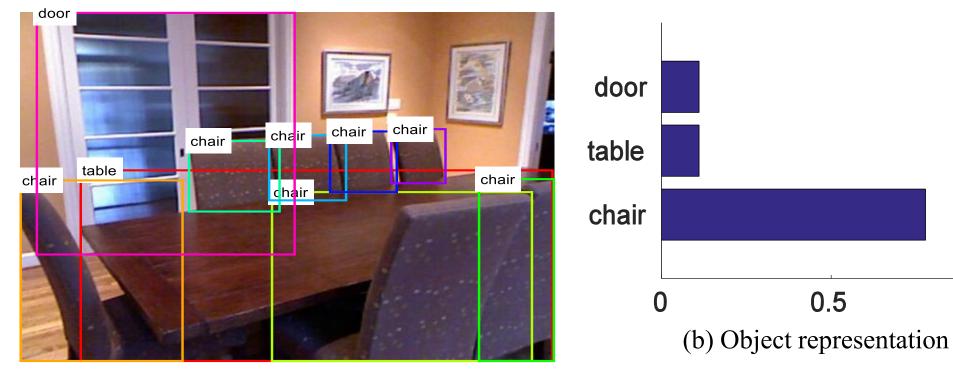
□ Improving scene recognition with spatial information, i.e., object-to-object relations (OOR)

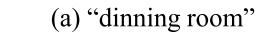
Contributions

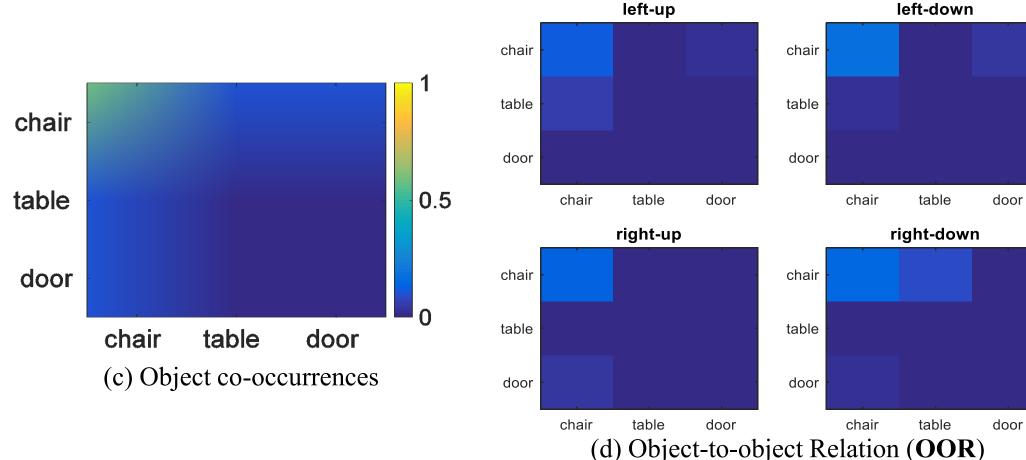
- □ Propose to detect OOR for image representation
- Propose to combine RGB-D representations with multi-modal fusion of object proposals



Representation of Object-to-Object Relation





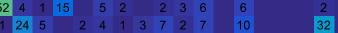


Framework of RGB-D Scene Recognition

Comparisons of representations



conferer



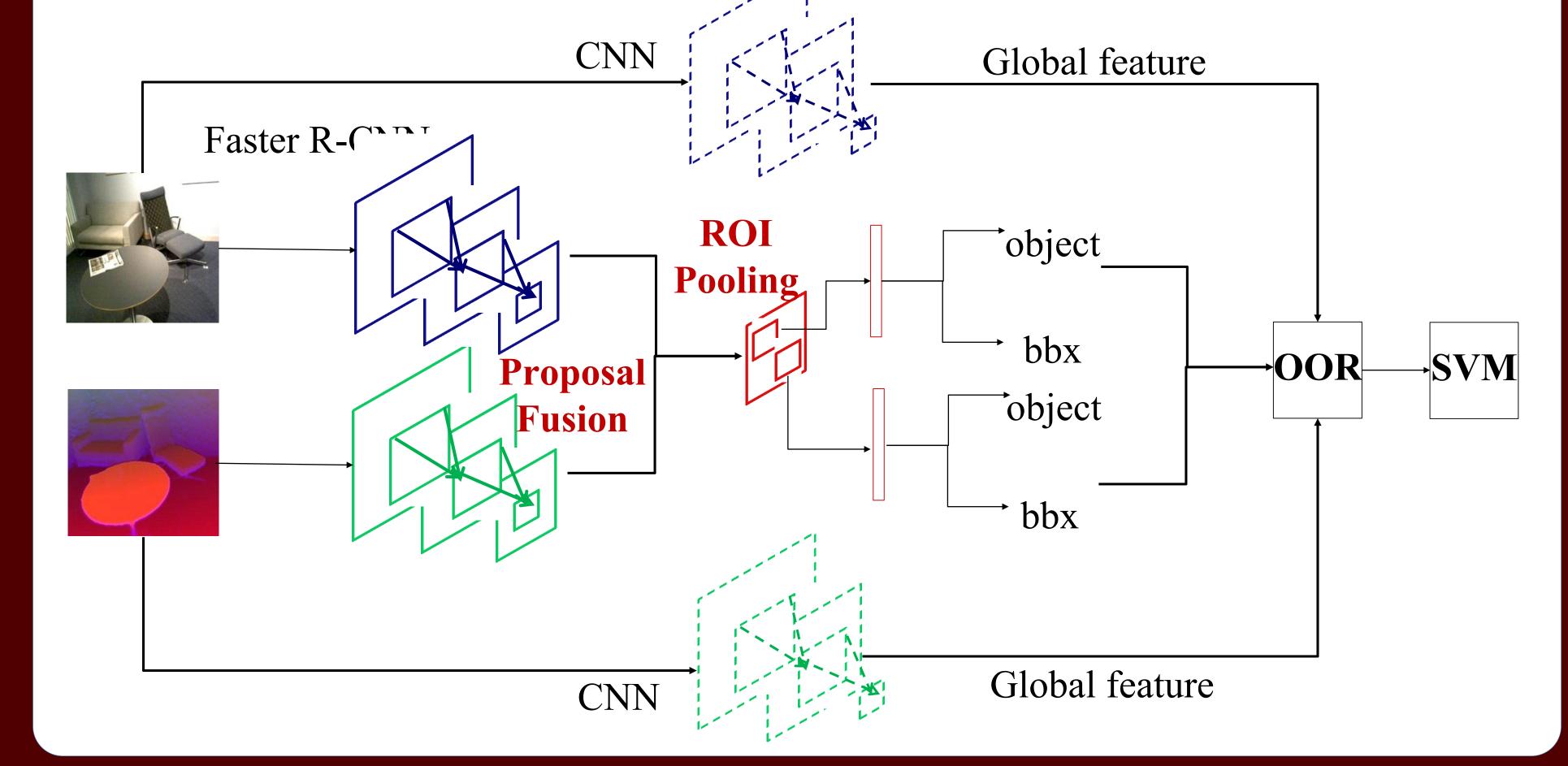




Table 1: Object detection	AP (%)	of SUN	RGB-D
---------------------------	--------	--------	-------

Model	bathtub	bed	bookshelf	box	chair	counter	desk	door	dresser	garbage_bin
FRCN-RGB	34.4	63.2	39.8	12.5	43.9	42.2	20.3	30.7	30.0	40.0

office 40 46 8 1 1 2 classroom 5 92 2 1 1 1 furniture_store 2 1 21 75 1 1 1 rest_space 1 2 49 43 1 1 2 2 bathroom 1 9 87 3 3 3 3 living_room 5 4 22 60 1 2 1 4 dining_area 1 94 3 7 71 2 4 dining_area 1 94 3 8 1 46 onference_room 1 95 2 1 1 46 onference_room 1 95 2 1 1 2 discussion_area 1 3 84 10 1 2 1 1 study_space 3 84 4 3 6 2 1 1 lecture_theatre <t< th=""><th>1</th><th>corridor 1 nference_room 3 lab iscussion_area dining_room 4 study_space 1 omputer_room lecture_theatre home_office 8</th><th>4 2 14 3 32 1 1 5 20 17 4 2 2 8 20 17 4 58 1 58 1 1 2 24 18 4 2 3 7 58 2 1 3 7 58 2 1 4 1 7 1 1 2 3 7 58 2 1 1 4 1 54 1 1 1 1 26 23 1 1 1 2 7 74 2 3 3 75 2 5 1</th><th>95 95 2 1 80 1 6 24 1 6 24 2 5 10 24 1 3 25 24 11 26 24 11 27 24 11 28 5 22 10 24 10 29 4 22 20 4 22 21 5 5 22 5 5</th><th>7 3 8 3 3 3 7 2 8 20 3 50 10 3 13 3 5 15 3 3 5 15 3 3 5 4 5 5</th><th>10 1 2 2 1 3 2 2 4 1 1 1 3 4 (CL) 1</th><th>32 3 1 2 2 10 2 1 7 1 1 3 54 9 22</th></t<>	1	corridor 1 nference_room 3 lab iscussion_area dining_room 4 study_space 1 omputer_room lecture_theatre home_office 8	4 2 14 3 32 1 1 5 20 17 4 2 2 8 20 17 4 58 1 58 1 1 2 24 18 4 2 3 7 58 2 1 3 7 58 2 1 4 1 7 1 1 2 3 7 58 2 1 1 4 1 54 1 1 1 1 26 23 1 1 1 2 7 74 2 3 3 75 2 5 1	95 95 2 1 80 1 6 24 1 6 24 2 5 10 24 1 3 25 24 11 26 24 11 27 24 11 28 5 22 10 24 10 29 4 22 20 4 22 21 5 5 22 5 5	7 3 8 3 3 3 7 2 8 20 3 50 10 3 13 3 5 15 3 3 5 15 3 3 5 4 5 5	10 1 2 2 1 3 2 2 4 1 1 1 3 4 (CL) 1	32 3 1 2 2 10 2 1 7 1 1 3 54 9 22
(a) Object representation (transf Accuracy: 22.8%	er matrix)			t repres			M)
office 23 4 2 12 1 2 7 2 4 1 6 classroom 30 32 6 13 1 4 1 2 1 furniture_store 14 2 9 41 1 2 18 1 7 2 - 1 1 furniture_store 14 2 9 41 1 2 18 1 7 2 - 1 1 bathroom 5 2 4 3 11 16 41 3 3 4 1 1 1 bathroom 5 2 4 3 11 16 41 3 3 4 1 1 1 kitchen 2 4 3 12 37 1 4 3 2 1 2 corridor 1 1 39 2 6 1 1 1 2 1 2 1 2 1 2	28 6 5 3 1 3 1 4 1 4 1 4 1 1 1 1 1 1 1 3 1 1 1 1 1 1 1 1 5 1 1 1 2 1 3 13 57 7 7 5 15 21	office classroom furniture_store rest_space bathroom living_room kitchen dining_area library corridor conference_room lab discussion_area dining_room study_space computer_room lecture_theatre home_office		37 1 1 50 3 1 1 33 1 1 3 2 1 2 7 4 1 3 1 2 1 2 3 6 1 19 3 2 4 8 1 8 1 8 3 9 1 1 1	20 1 27 1 5 5 10 3 8 12 7 6 12 7 6 9 17 3 2 5 5 1	2 2 26 5 4 5 1 5 5 3 1 3 2	7 26 2 11
(c) Object co-occurrences (Accuracy: 28.8%	SVM)	(d)) Object-t A	o-Obje ccuracy		```	SVM

With OOR, the classifier obtains less confusion and better accuracy

								S	UN	RO	GB-	D							
bedroom	52	6	3	7	2	3	10	1			2		5		1				9
office	2	43	8	3	5	3	1	2		1	5	1	9		1	3	8		6
classroom	2	3	47	2	4	2	1		5	2	3	7	2	2		7	5	8	
furniture_store	1	3	1	83	2		1		1	1					5	1		1	
rest_space	4	3	7	2	54	1	4		3	5	6	1		6		3	1		
bathroom	1		1	2	1	91		1		1	1		1						
living_room	18	3	2	3	6	1	46	1			3		3		5				6
kitchen	4	3	3	3	3	6	1	64					6		5				1
dining_area	1	3	8	10	10	1			31	11	2	1	1	3	7	6			
library	1	5	9	1	23				5	32	1	5		3		9	4	1	
corridor	2	5	4	5	7	3	1			1	64	1	5	1					

FRCN-Depth	54.5	71.6	25.5	5.0	45.4	39.5	22.2	10.5	18.0	34.2
FRCN-RGBD	57.5	75.6	44.2	17.7	49.6	48.9	25.4	33.6	40.2	49.2
Model	lamp	monitor	night_stand	pillow	sink	sofa	table	tv	toilet	mAP
FRCN-RGB	38.5	34.3	39.2	33.0	46.9	39.5	34.6	23.2	74.5	37.9
FRCN-Depth	40.0	18.8	34.8	40.2	49.2	44.9	41.2	14.3	70.0	35.8
FRCN-RGBD	53.0	44.0	47.6	48.6	61.1	50.3	43.2	35.2	81.7	47.7

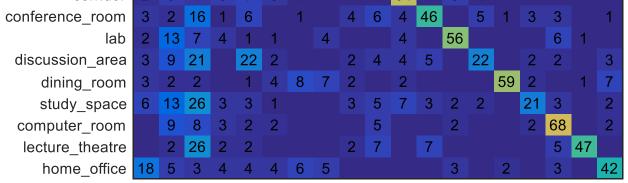
Table 2: Scene recognition accuracy (%) with intermediate Table 4: Comparisons on SUN RGB-D in accuracy(%) Table 5: Comparisons on NYUD2 in accuracy(%) representation

Intermediate	RGB	Depth	RGB-D
representations			
P_S^I	16.8	13.9	17.8
P_O^I	31.4	26.5	31.9
P_{OO}^{I}	32.7	28.7	33.4
P_{OOR}^{I}	33.5	30.0	36.3
	.1 1		

 P_S^I : inference with object representation P_{O}^{I} : SVM classification with object representation P_{OO}^{I} :SVM classification with object co-occurrence P_{OOR}^{I} :SVM classification with OOR

	Method	RGB-D				
Proposed	Local-OOR	50.3				
	Global+Local	52.6				
	Global+Local-OOR	54.0				
	Song <i>et al.</i> [33]	39.0				
State-of-the-art	Zhu <i>et al.</i> [43]	41.5				
State-of-the-art	Wang <i>et al.</i> [39]	48.1				
	Song <i>et al.</i> [34]	52.4				
Global: CNN	Global: CNN features of images					
Local: CNN features of bounding boxes						

Method	RGB	Depth	RGB-D						
Proposed methods									
Local	51.2	46.4	56.4						
OOR	45.1	40.9	48.6						
Global	57.3	54.1	64.0						
Local-OOR	-	-	60.1						
Global+Local-OOR	-	-	66.9						
State-of-the-art									
Gupta <i>et al.</i> [17]			45.4						
Wang <i>et al.</i> [39]			63.9						
Song <i>et al.</i> [34]			65.8						



Confusion matrix of Global+Local-OOR.

Conclusion

► Introduce some analysis and insights between objects and scenes

► Propose a framework to extract object-toobject relation (OOR) for scene recognition

The propose method achieves the state-ofthe-art on public RGB-D databases