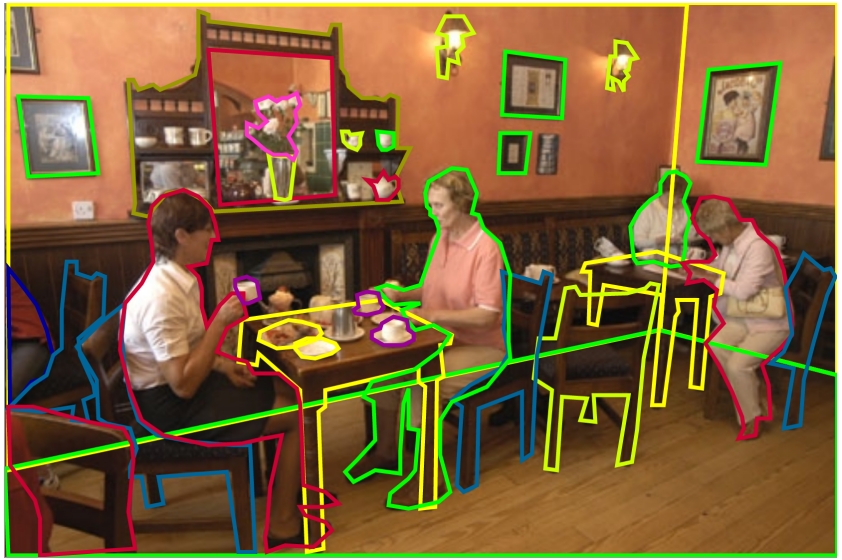
Scenes – Structure + Content

Yang Cai, David Fouhey

What is a scene?

What is a scene?



Scene is a collection of objects organized in certain structure.

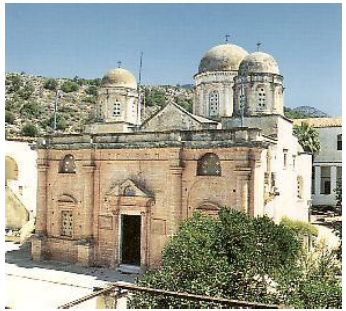
How should we represent scenes?

How should we represent scenes?

Semantic/linguistic categories?

Very popular answer in vision.

Beer Garden



Monastery

Bedroom

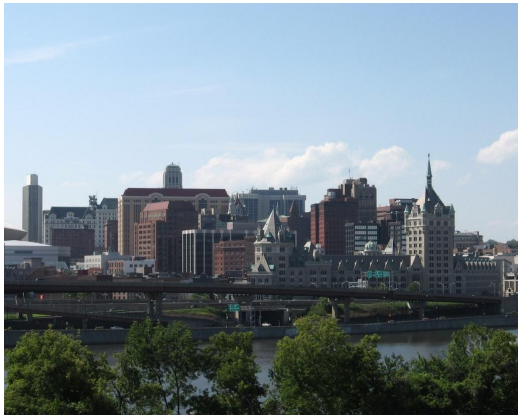
Issues with Semantic Categories

Intra-class variation

New classes?

Issues with Semantic Categories

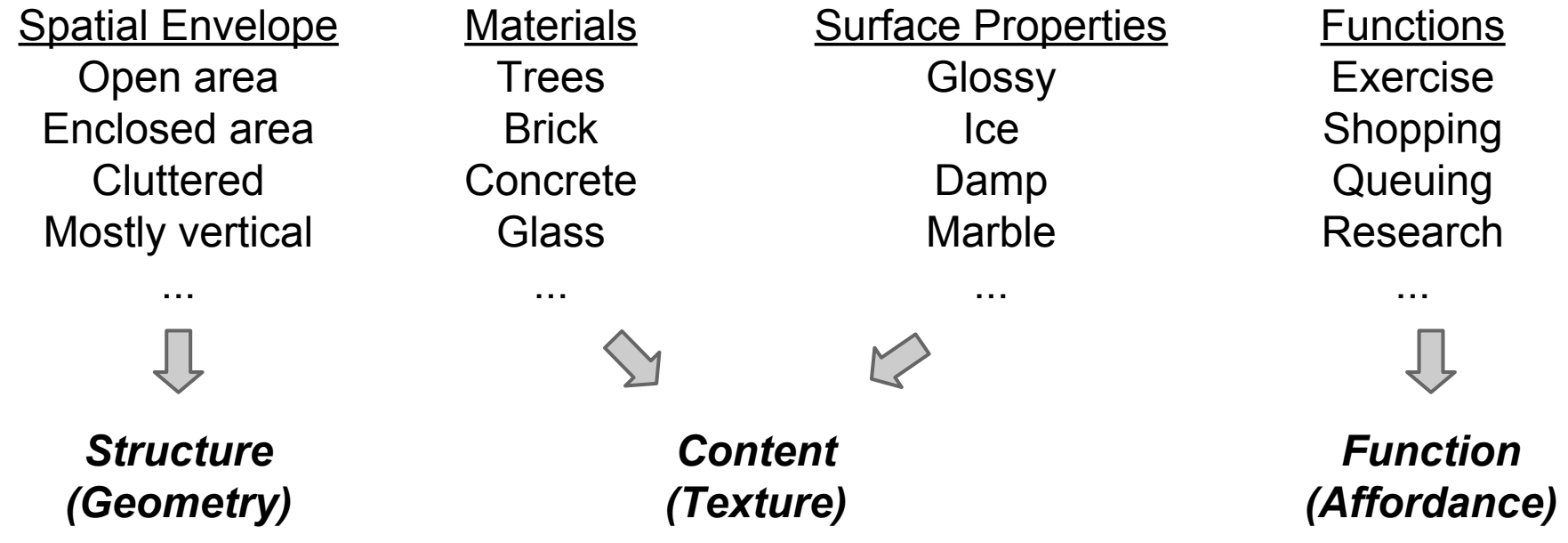




Multiple membership in unrelated categories,
Smooth transitions between categories

Patterson and Hayes

Core idea: represent scenes via attributes



Example

Exercise
Walking
...

Shopping
Residence
...

Structure (Geometry)

Content (Texture)

Function (Affordance)

Patterson and Hayes

Dataset: attribute labels for images

Function:sailing/boating, sunbathing, swimming,
Content: waves surf, moist/damp,
Structure: open area, far-away horizon,
….

Patterson and Hayes

Task 1: Kitchen sink of features (HOG, SSIM,...), predicts attribute label

Open Area

Patterson and Hayes

Task 2: Ground-truth attributes predict scene

Function:sailing/boating, sunbathing, swimming,
Content: waves surf, moist/damp,
Structure: open area, far-away horizon,
….

Beach

Patterson and Hayes

Task 3 (IJCV edition): predicted attributes predict scene

Function:sailing/boating, ...
Content: grass,surf, ...
Structure: open area, ...

Beach

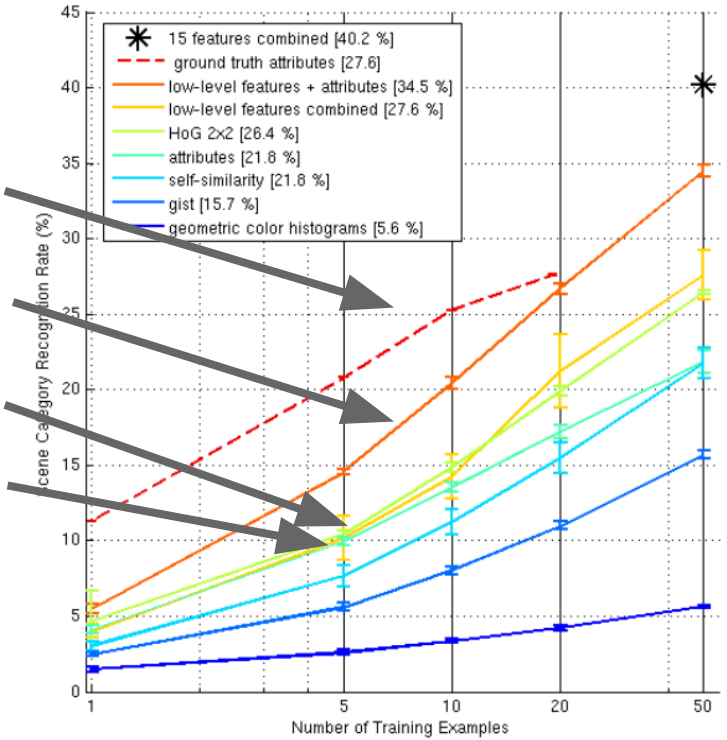
Patterson and Hayes

Ground-truth attributes alone

Attributes + Appearance

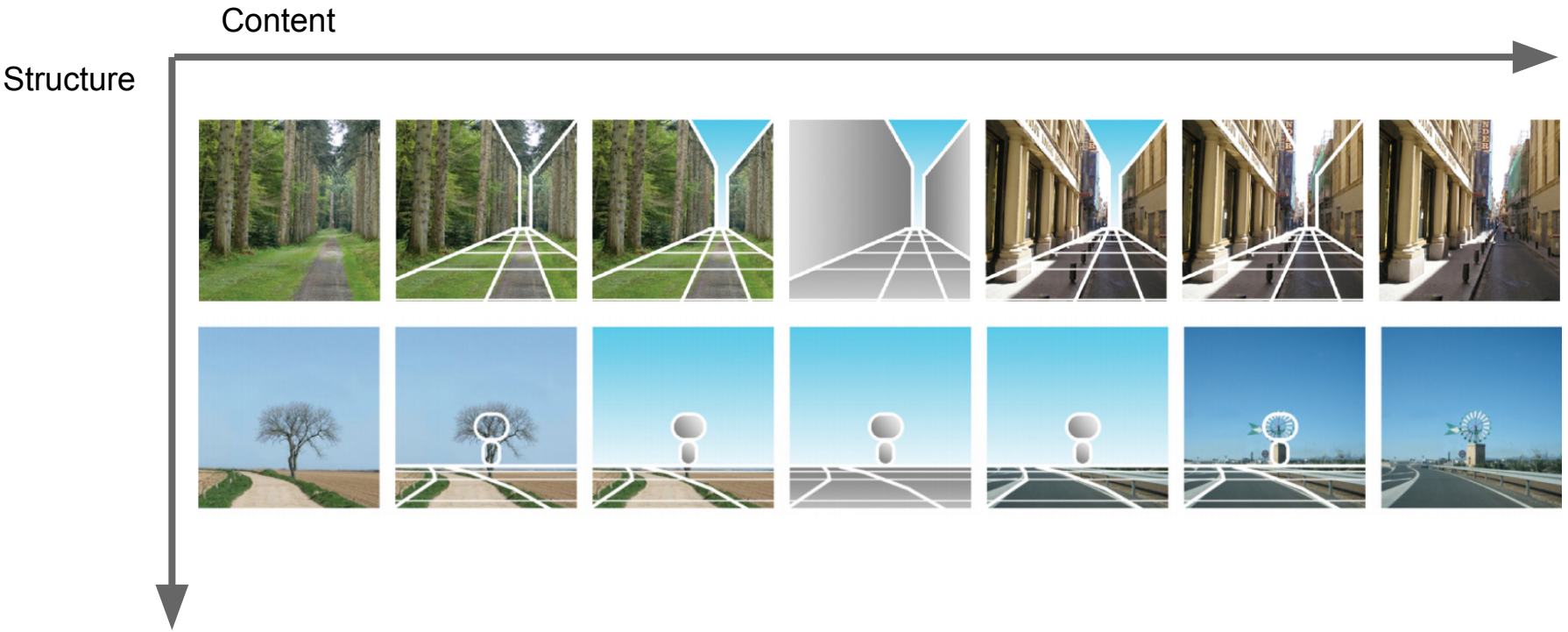
Appearance Alone

Attributes Alone



Patterson et al., IJCV to appear

Are All Variations the Same?



Are All Variations the Same?

Content

Kitchen, indoors, marble, has a chandelier

Structure

Closed, enclosed, wide view, has floors visible

Question

Can we really handle all variations (e.g., structure, content) the same?

Is structure just some attribute?

How do we represent scenes?

Goal

(a) What do the representations of the parahippocampal place area (PPA), lateral occipital complex (LOC), and early visual cortex (EVC) encode?

(b) Is this complementary or overlapping?

Kravitz et al.

Content:
Man-made/Natural


Per class
(beaches are natural)

Relative Distance:
Near/Far


Per instance
(a close up city)

Expanse:
Closed/Open


Per class
(highways are open)

Images from Oliva and Torralba, 2011

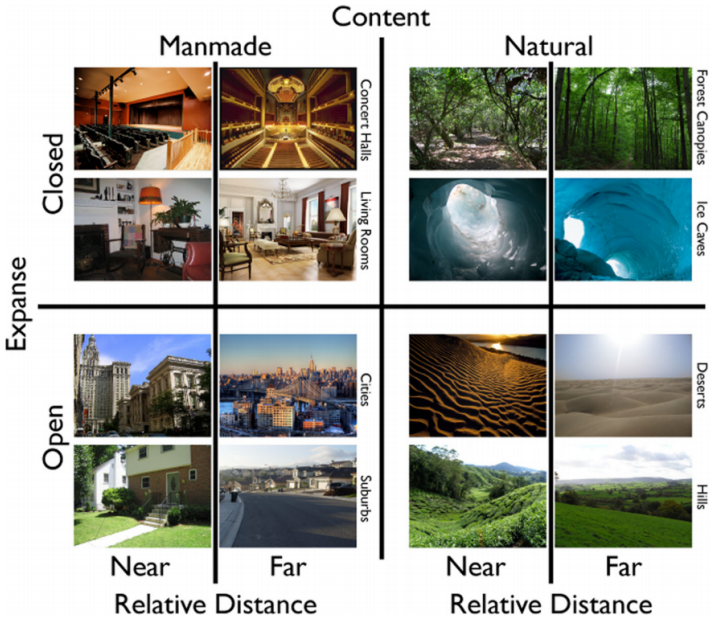
Method

Participants do a fixation cross task, judging which arm grew longer

Participants viewed a scene for 500ms.

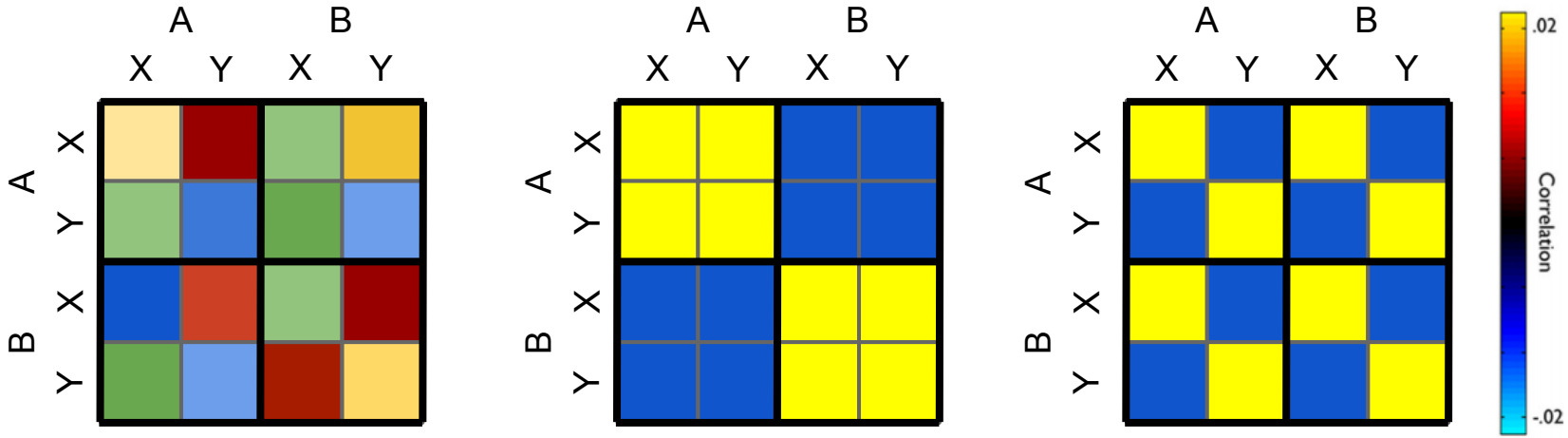
Method

Separate participants judge a pair of scenes giving labels

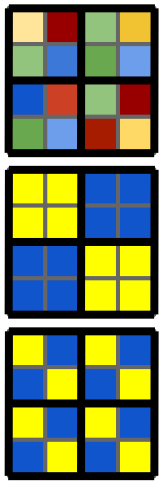


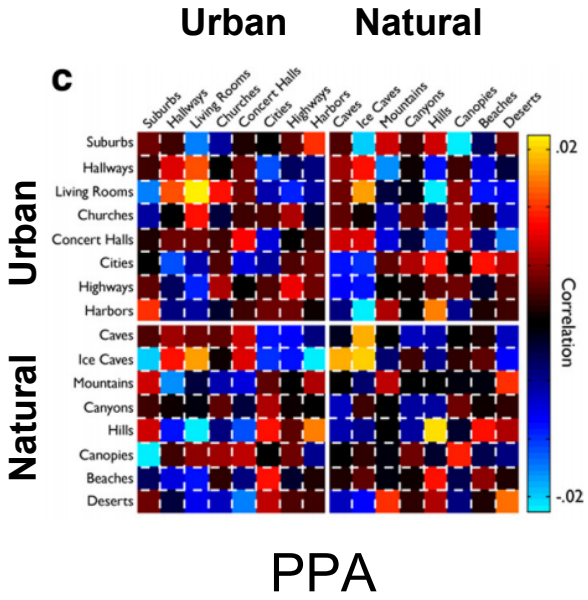
Results

Paradigm: compare representational similarity

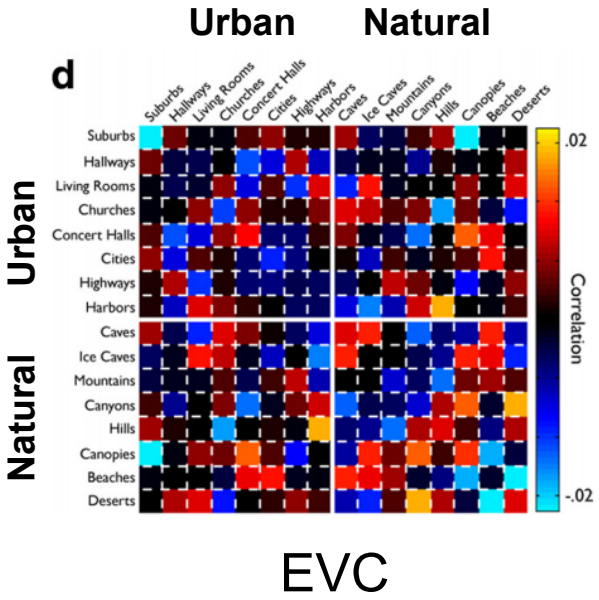


Results - Content



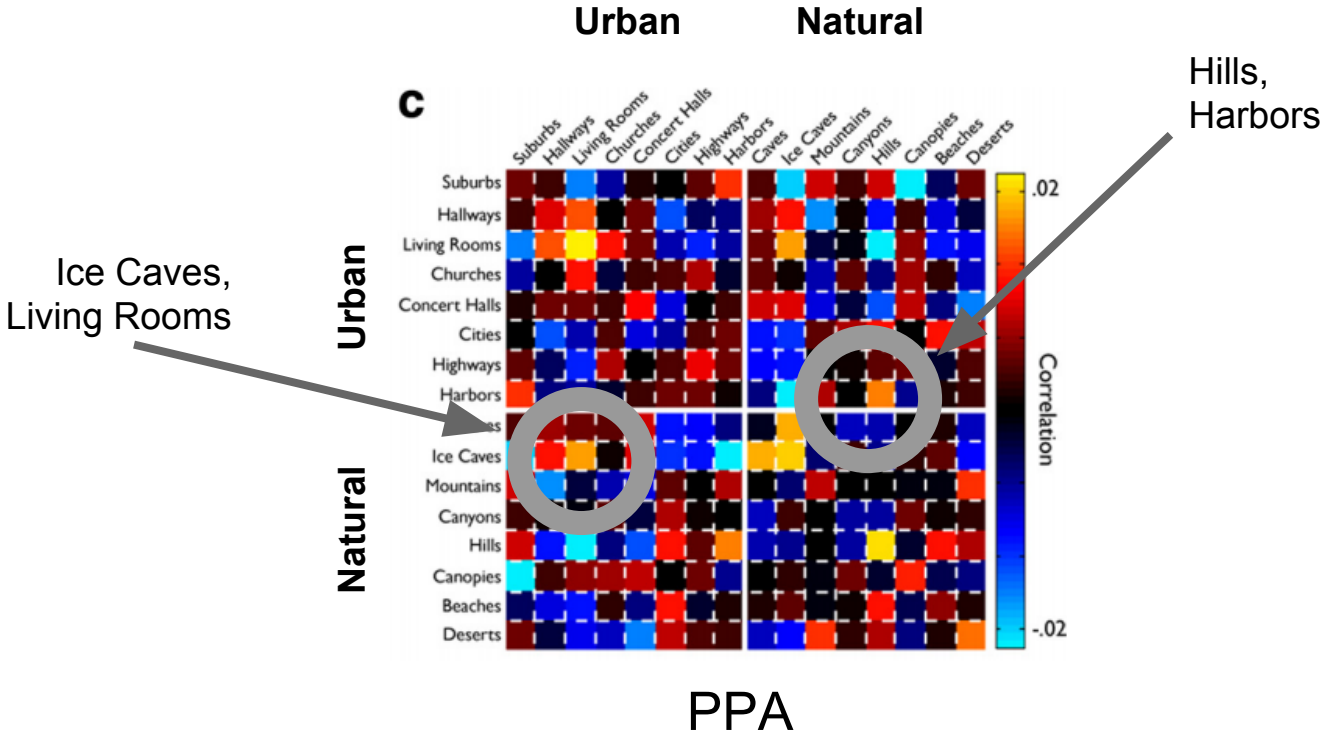


PPA

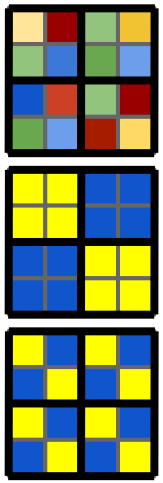


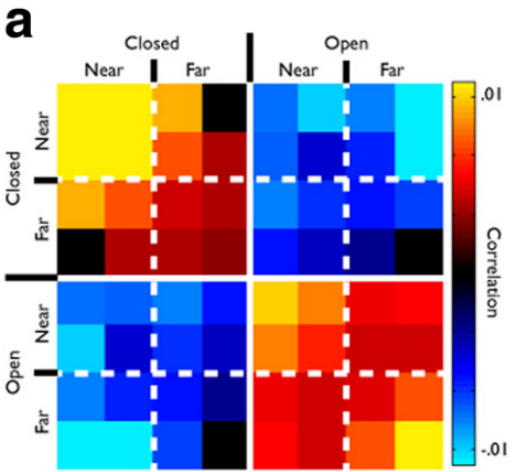
EVC

Results - Content

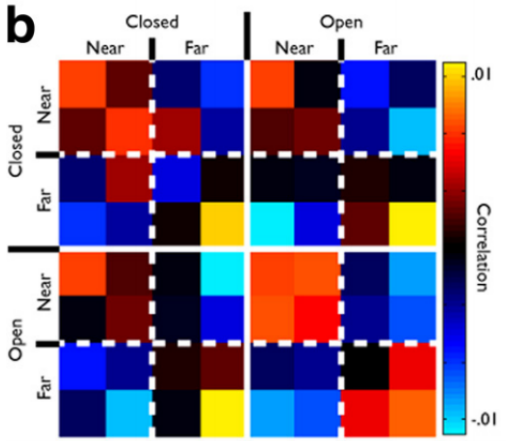


Results - Expanse and Distance





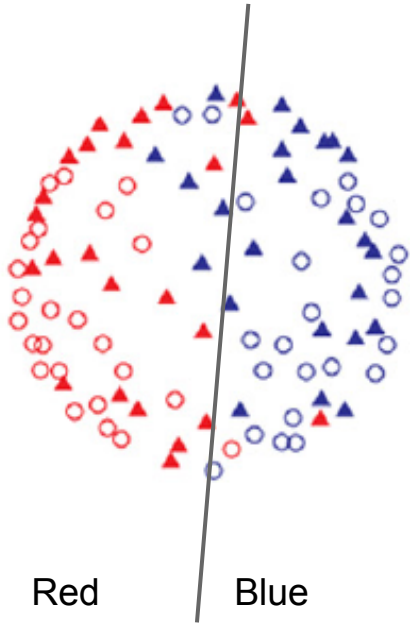
PPA

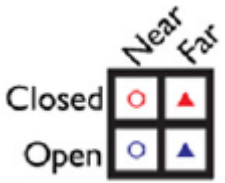


EVC

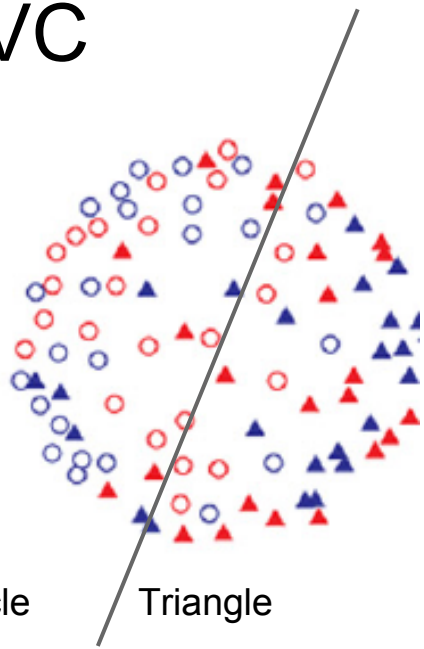
Results - Multidimensional Scaling

PPA





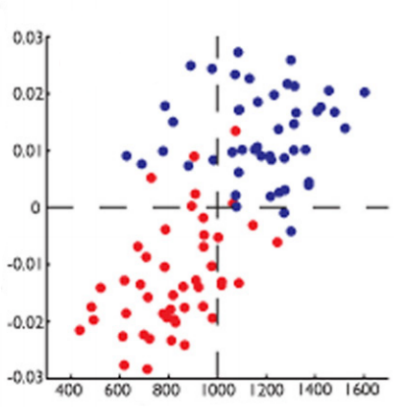
EVC



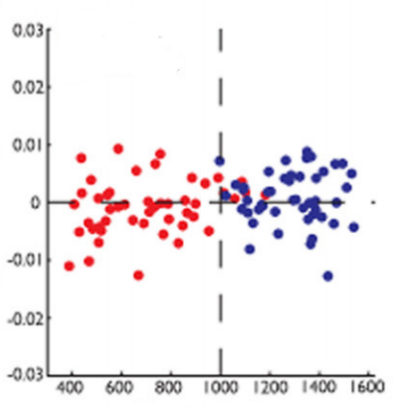
Results

Continuous results: compare “fMRI Scores”, compare with ranking of users.

FMRI Scores



Ranking from users



Ranking from users

Results

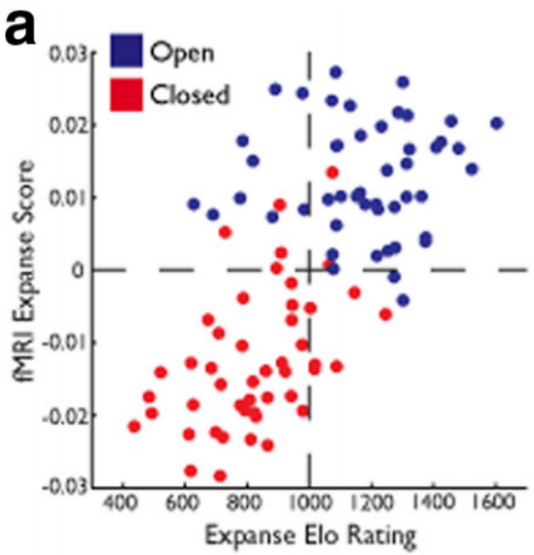
FMRI Score =

correlation with positives -

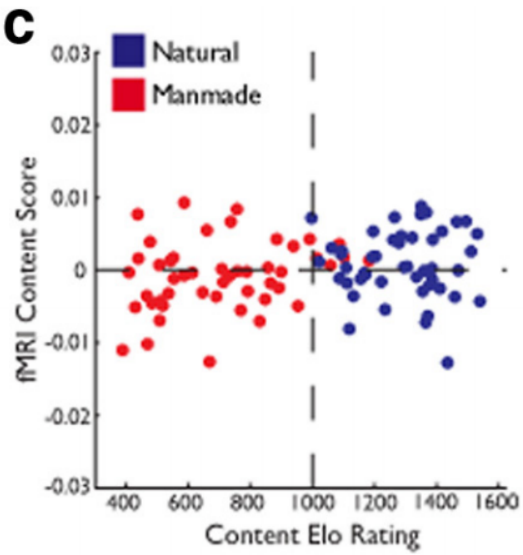
correlation with negatives

Results - PPA

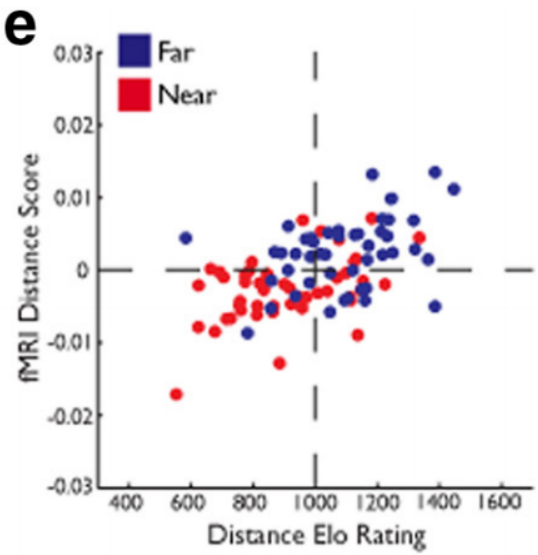
Expanse



Content

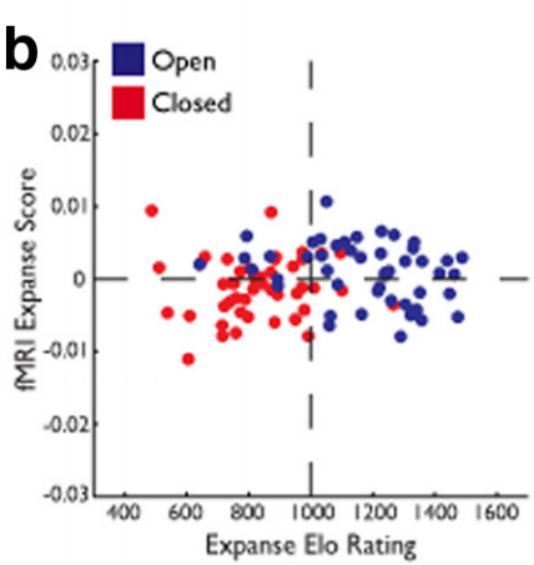


Distance

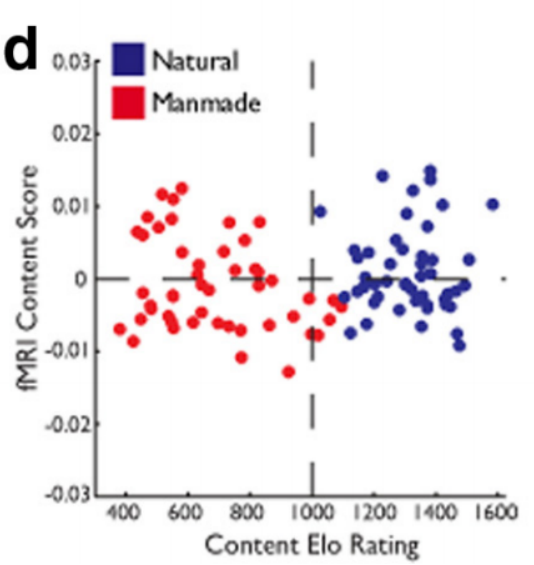


Results - EVC

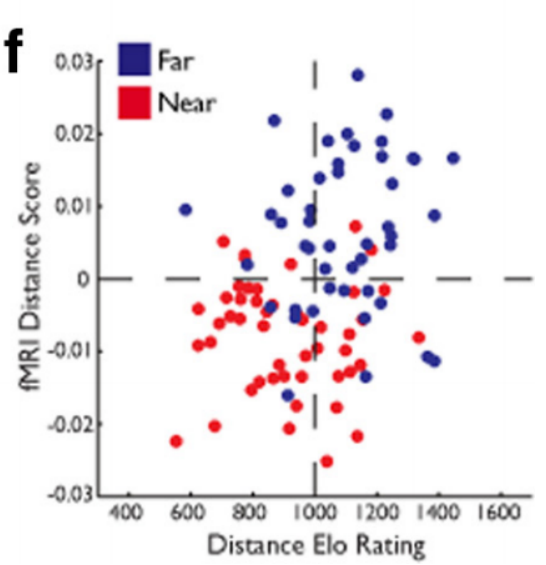
Expanse



Content



Distance

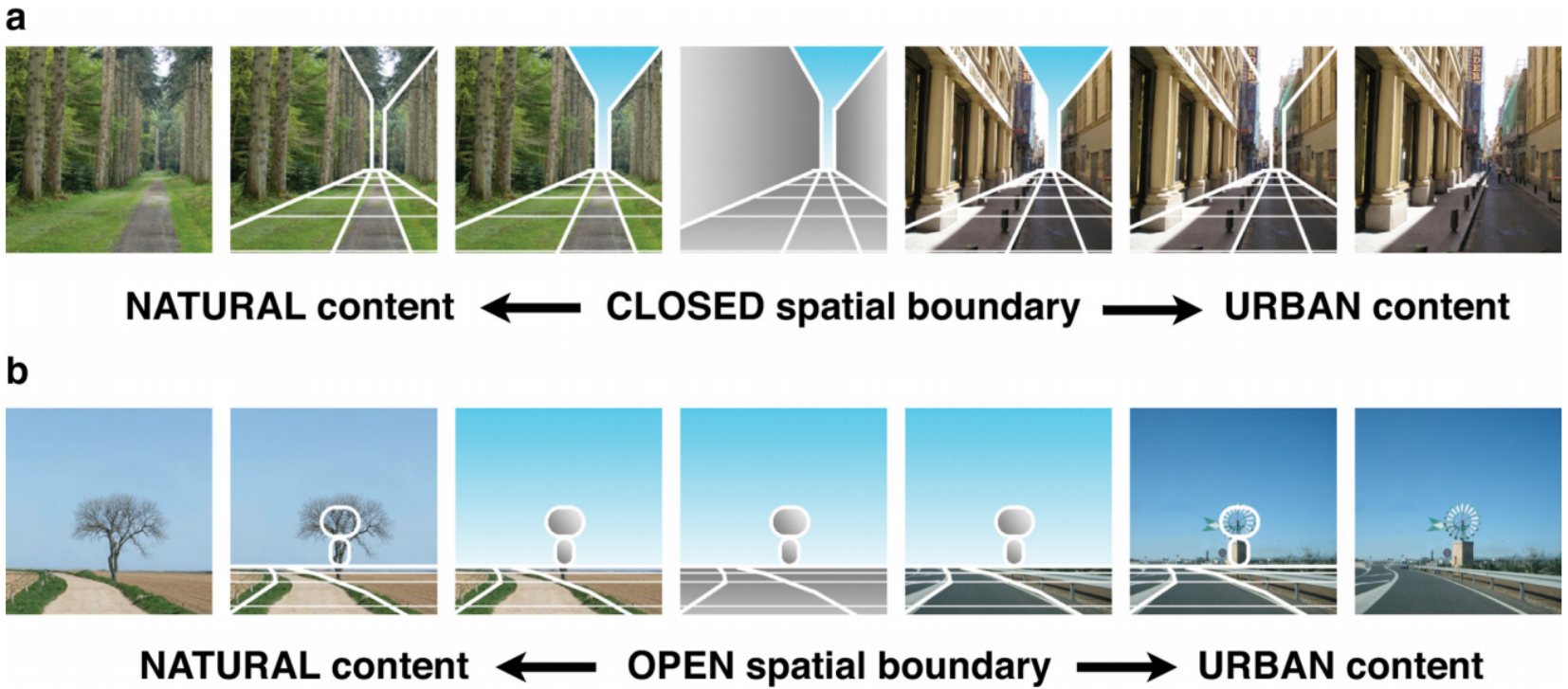


Kravitz: PPA does not encode content!

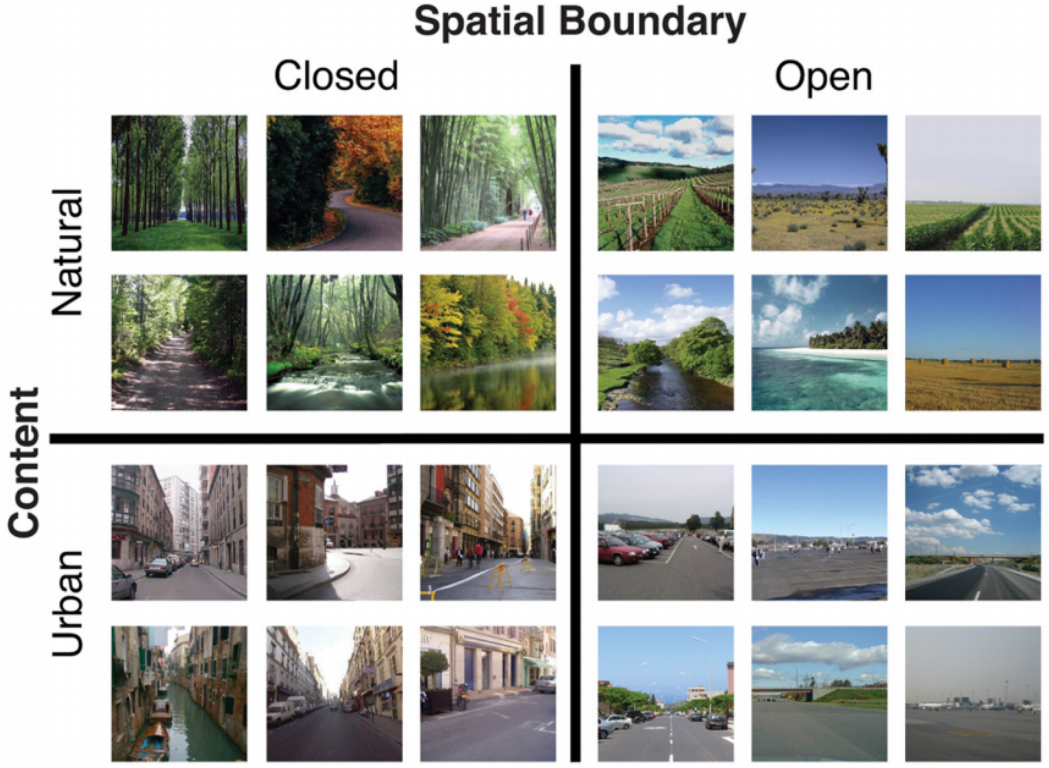
Me: But... where the content is encoded?

Park: Read my paper.

Park et al.

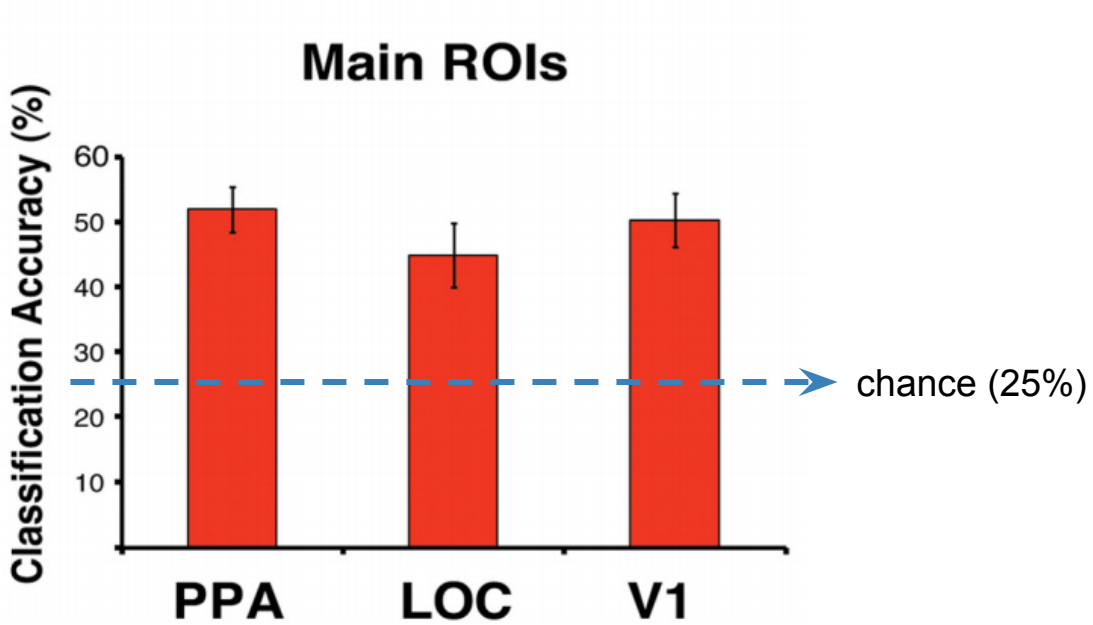

a
NATURAL content
CLOSED spatial boundary
URBAN content
b
NATURAL content
OPEN spatial boundary
URBAN content

Method



Visual stimuli: examples scenes of four conditions

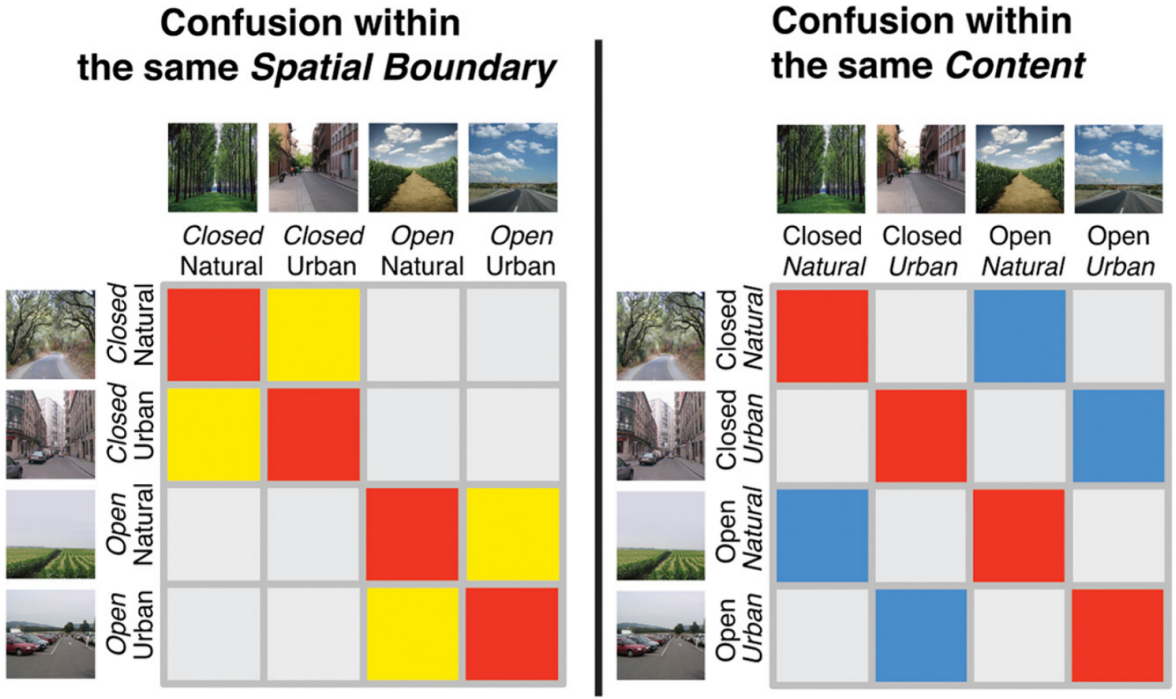
Results



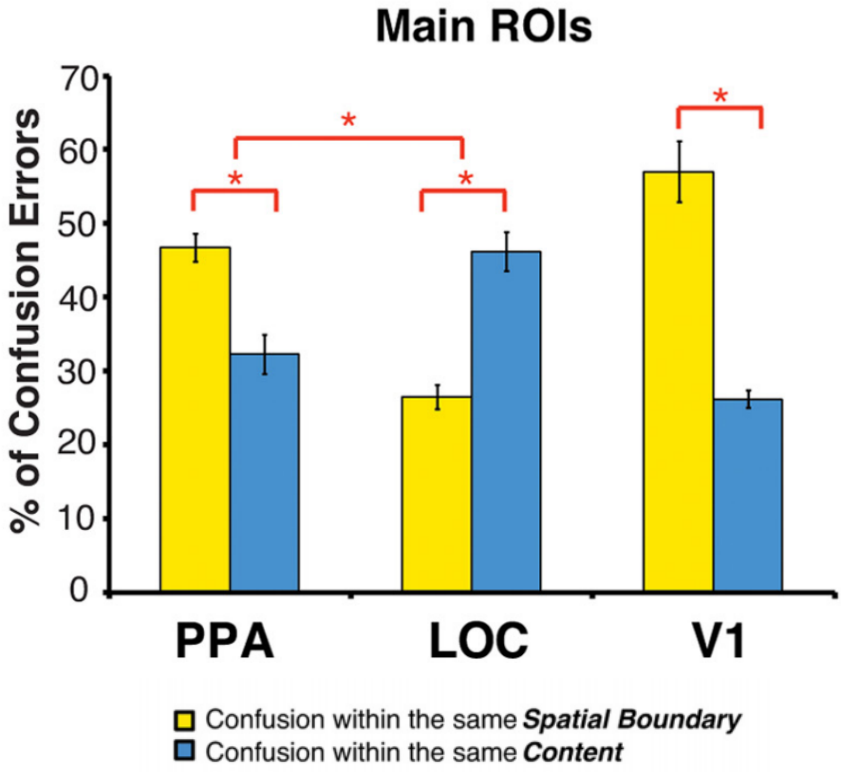
Average classification accuracy

The classification accuracy did NOT inform about the nature of scene representation in each region.

Method


Confusion within the same *Spatial Boundary*
Closed Natural
Closed Urban
Open Natural
Open Urban
Confusion within the same *Content*
Closed Natural
Closed Urban
Open Natural
Open Urban
Confusion within the same *Spatial Boundary*
Confusion within the same *Content*

Results


Main ROIs
% of Confusion Errors
70
60
50
40
30
20
10
0
*
*
*
*
PPA
LOC
V1
Confusion within the same Spatial Boundary
Confusion within the same Content

Results

By merging the voxels of PPA and LOC, the classification accuracy increased to 56% (PPA only: 50.4%, LOC only: 45%).

Conclusions:

(1) *structure* was primarily represented in PPA and *content* was mainly represented in LOC.

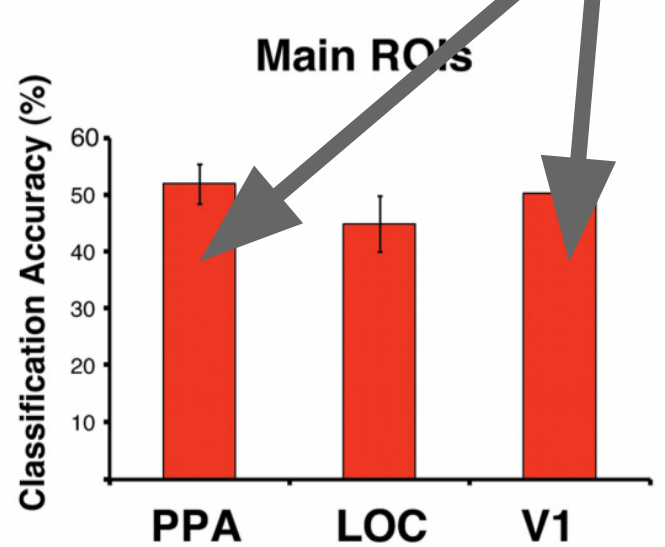
(2) scene representation within these regions is *complementary*.

Concerns?

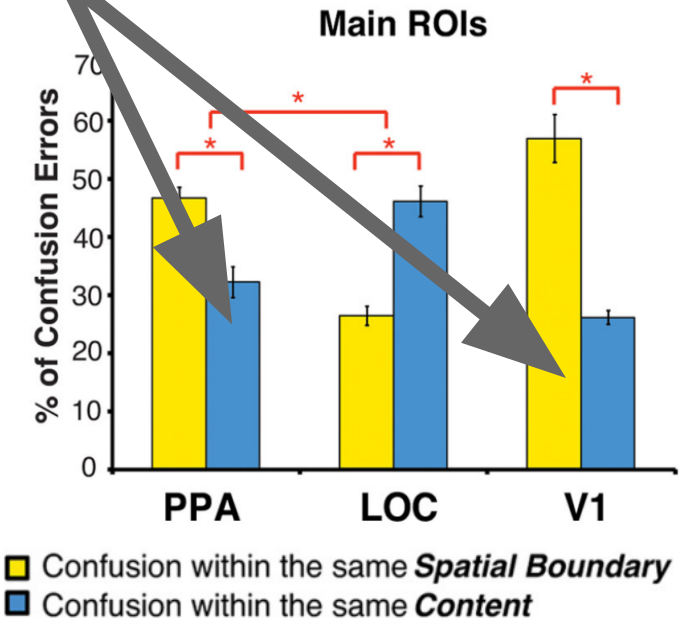
- 1. Performance correlation of PPA & V1: is it just low-level image statistics?
- 2. Walther et al. 2009 shows categories in the PPA. Are categories encoded in the PPA?

Is it just low-level image statistics?

PPA and V1 have very similar confusion patterns

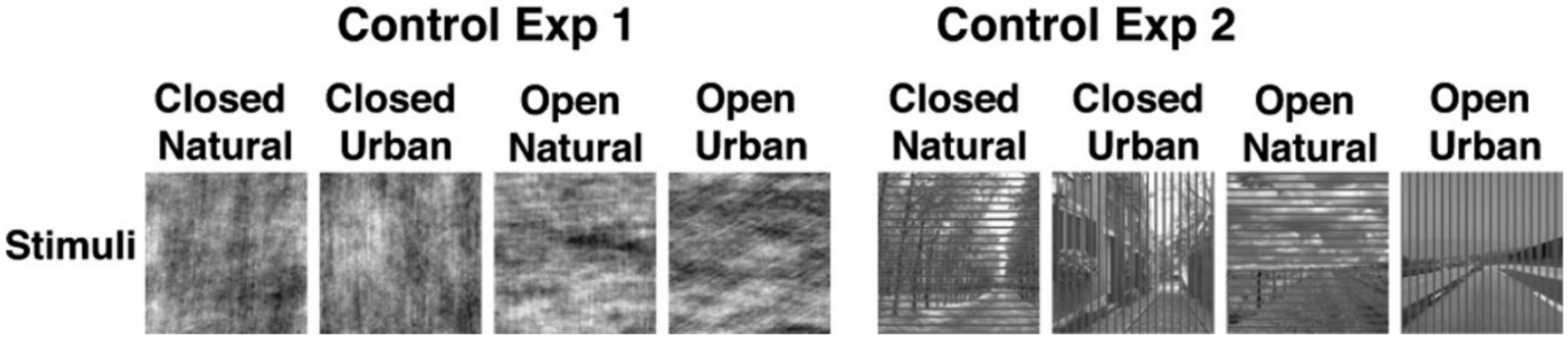


Average classification accuracy



Is it just low-level image statistics?

How much can be explained by low-level statistics alone?

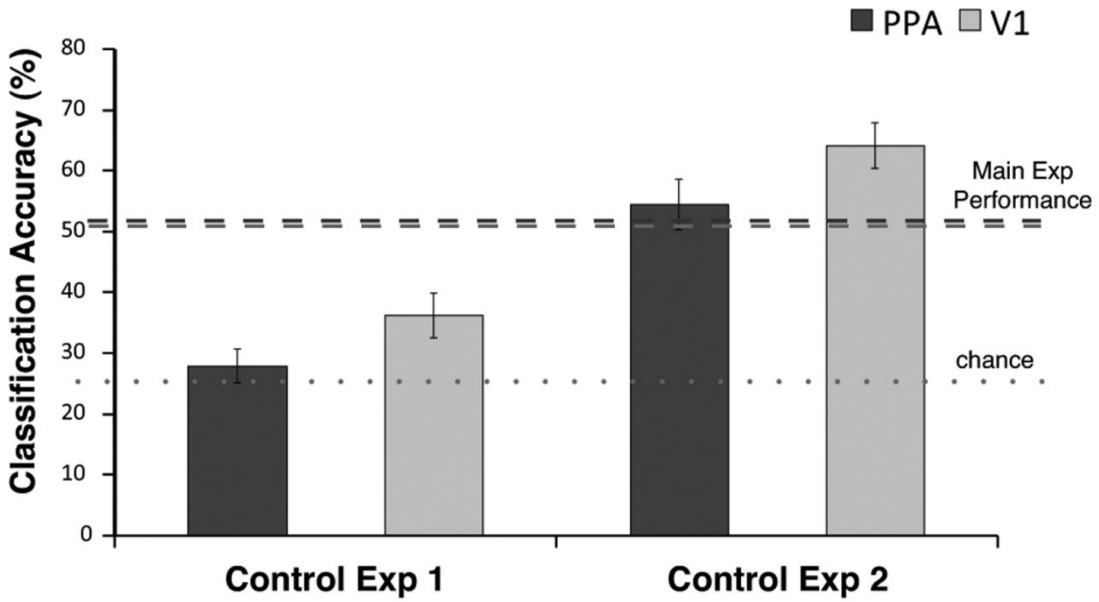


phase-scrambled images: to remove the high-order information.

original scenes with artificially added vertical and horizontal lines: increase the low-level visual discriminability

Is it just low-level image statistics?

Result: not much!



High-level categories in PPA?

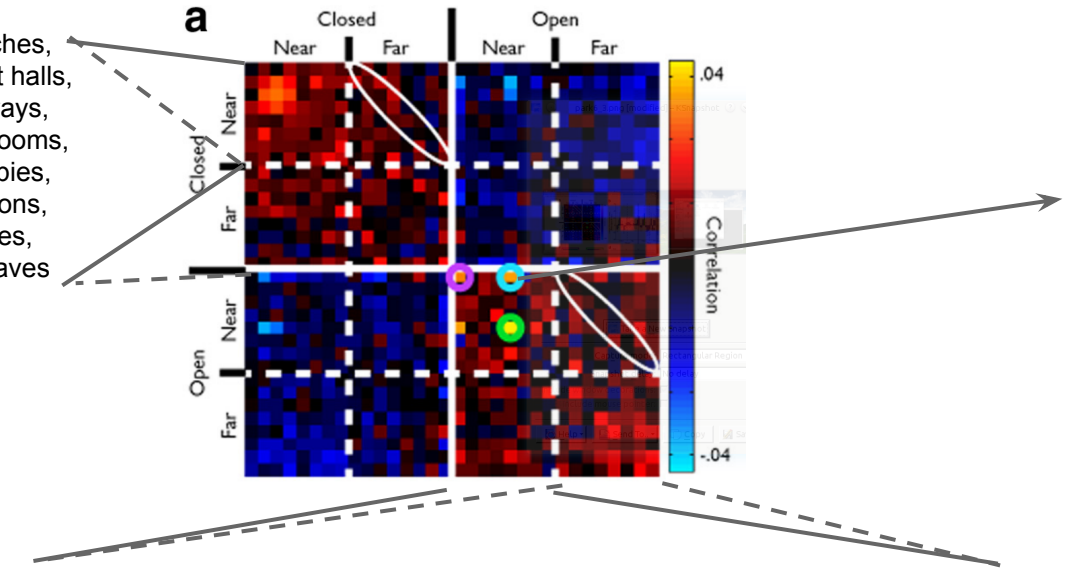
Walther et al. could decode category in PPA

	Decoder prediction (PPA)					
	beaches	buildings	forests	highways	industry	mountains
beaches	**0.28**	0.08	0.10	0.20	0.10	0.23
buildings	0.12	**0.27**	0.07	0.27	0.15	0.13
forests	0.15	0.12	**0.33**	0.13	0.15	0.12
highways	0.18	0.17	0.10	**0.40**	0.10	0.05
industry	0.08	0.30	0.10	0.12	**0.25**	0.15
mountains	0.18	0.08	0.23	0.07	0.08	**0.35**

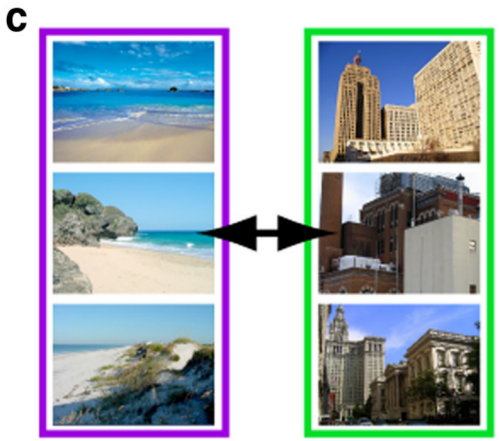
High-level categories in PPA?

Average the similarity by category with constant spatial factor

churches,
concert halls,
hallways,
living rooms,
canopies,
canyons,
caves,
ice caves



cities, harbors, highways,suburbs, beaches, deserts, hills, and mountains



Near beach

Near cities

Unanswered questions

- Processed separately does not tell us how it is processed. Next presentations address this
- What does this mean for computer vision?
- Where is scene function (affordances) encoded in brain?

Questions raised on blog

- Is open/closed just outdoor/indoor?
- Are per-class expanse labels problematic?
- Is the fixation cross design / passive viewing ideal?