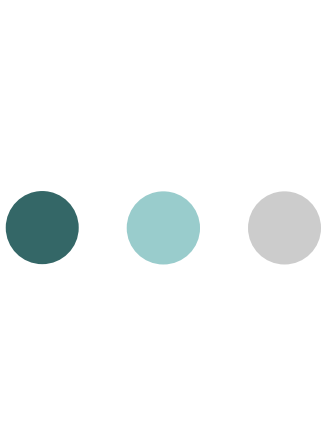


Putting Context into Vision

Derek Hoiem

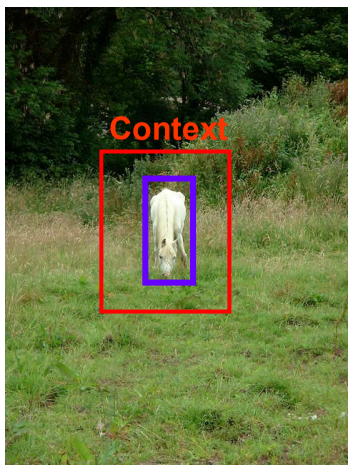
September 15, 2004

Questions to Answer

- What is context?
- How is context used in human vision?
- How is context currently used in computer vision?
- Conclusions

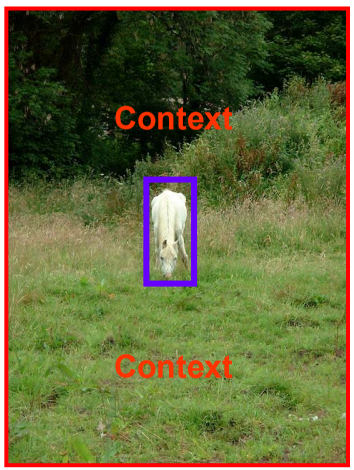
What is context?

- Any data or meta-data not directly produced by the presence of an object
  - Nearby image data



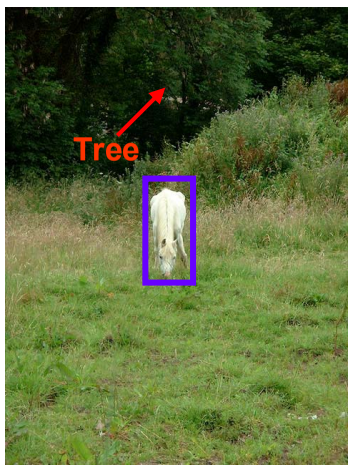
What is context?

- Any data or meta-data not directly produced by the presence of an object
  - Nearby image data
  - Scene information



What is context?

- Any data or meta-data not directly produced by the presence of an object
  - Nearby image data
  - Scene information
  - Presence, locations of other objects



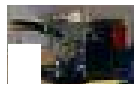
How do we use context?

Attention

- Are there any live fish in this picture?

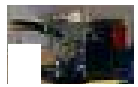
Clues for Function

- What is this?



Clues for Function

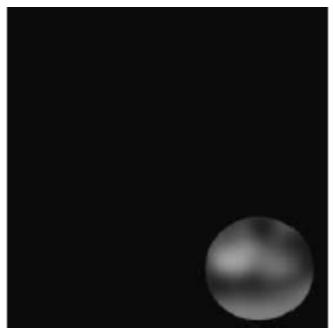
- What is this?



- Now can you tell?

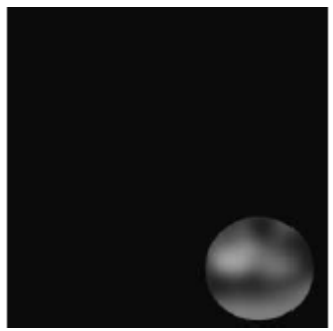
Low-Res Scenes

- What is this?



Low-Res Scenes

- What is this?



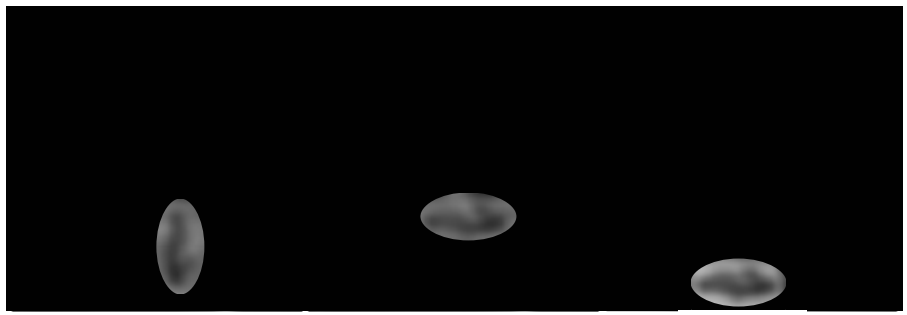
- Now can you tell?

More Low-Res

- What are these blobs?

More Low-Res

- The same pixels! (a car)



Why is context useful?

- Objects defined at least partially by function
  - Trees grow in ground
  - Birds can fly (usually)
  - Door knobs help open doors

Why is context useful?

- Objects defined at least partially by function
  - Context gives clues about function
    - Not rooted into the ground ➔ not tree
    - Object in sky ➔ {cloud, bird, UFO, plane, superman}
    - Door knobs always on doors

Why is context useful?

- Objects defined at least partially by function
  - Context gives clues about function
- Objects like some scenes better than others
  - Toilets like bathrooms
  - Fish like water

Why is context useful?

- Objects defined at least partially by function
  - Context gives clues about function
- Objects like some scenes better than others
- Many objects are used together and, thus, often appear together
  - Kettle and stove
  - Keyboard and monitor

How is context used in computer vision?

Neighbor-based Context

- Markov Random Field (MRF) incorporates contextual constraints

Blobs and Words – Carbonetto 2004

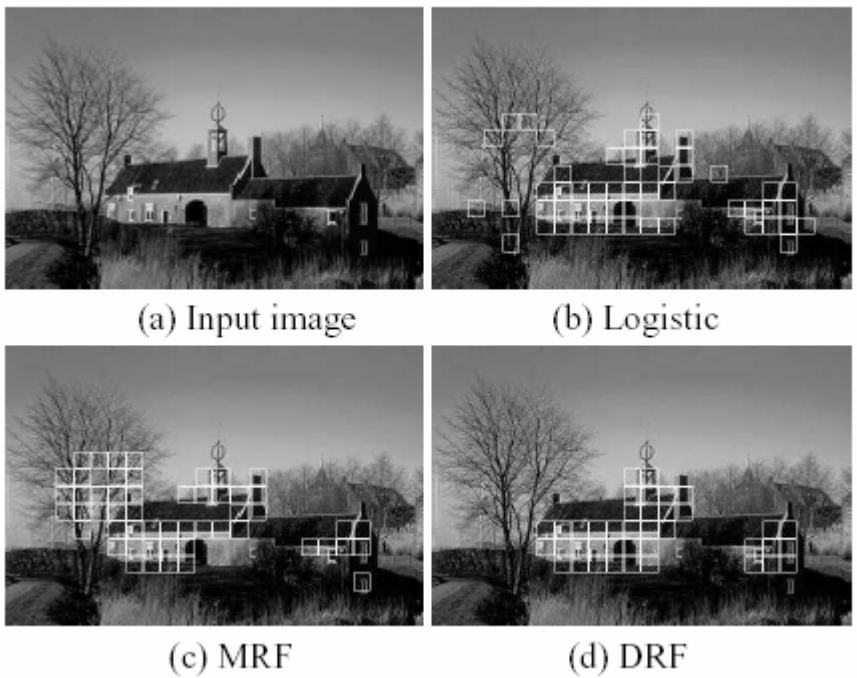
- Neighbor-based context (MRF) useful even when training data is not fully supervised

- Learns models of objects given captioned images

WORD	LABEL % TRAIN	LABEL % TEST†	ANNOTATION %† TRAIN	ANNOTATION %† TEST	*dInd* PR. TRAIN	*dInd* PR. TEST	*dMRF* PR. TRAIN	*dMRF* PR. TEST
airplane	0.060	0.055	0.036	0.028	0.135	0.102	0.290	0.187
astronaut	0.003	0.003	0.001	0.002	0.794	0.087	0.000	0.135
atm	n/a	0.003	n/a	0.006	n/a	0.000	n/a	0.000
bear	0.031	0.017	0.021	0.013	0.192	0.092	0.452	0.272
beluga	n/a	0.003	n/a	0.005	n/a	0.000	n/a	0.000
bill	0.019	0.017	0.046	0.031	0.269	0.175	0.335	0.146
⋮								
Totals	1.000	1.000	1.000	1.000	0.294	0.199	0.486	0.301

Discriminative Random Fields – Kumar 2003

- Using data surrounding the label site (not just at the label site) improves results

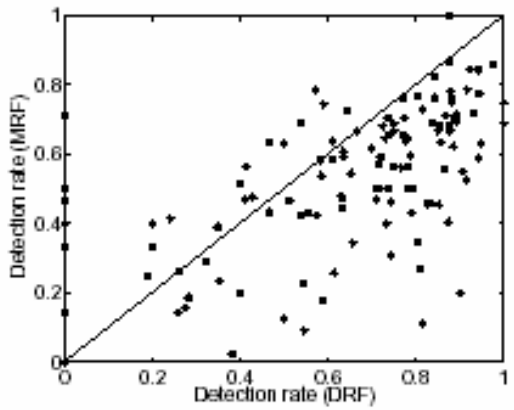


(a) Input image (b) Logistic

(c) MRF (d) DRF

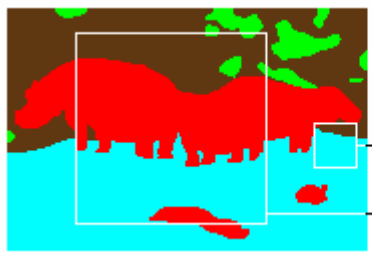
Buildings vs.
Non-Buildings

Method	FP (per image)	DR (%)
→ MRF	2.36	57.2
Logistic$^-$	2.24	45.5
→ DRF$^-$	2.24	60.9
Logistic	1.37	55.4
DRF ($K = 0$)	1.21	68.6
→ DRF	1.37	70.5

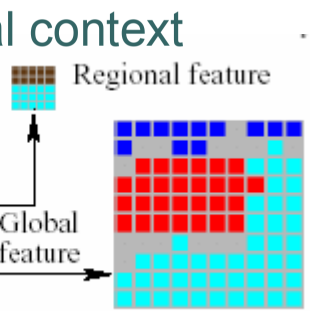


Multi-scale Conditional Random Field (mCRF) – He 2004

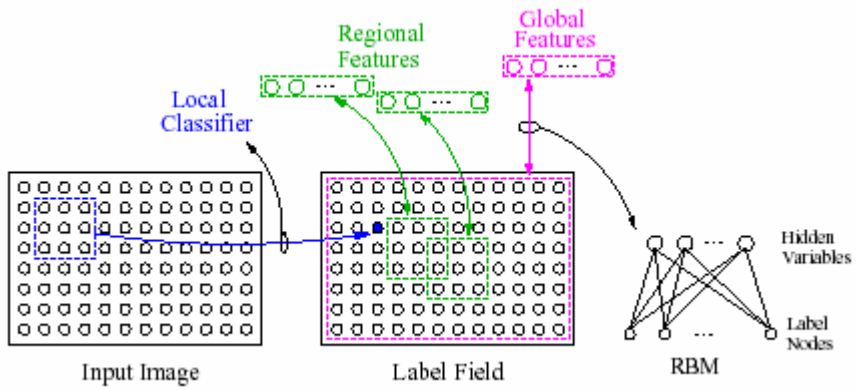
Raw image



Independent data-based labels



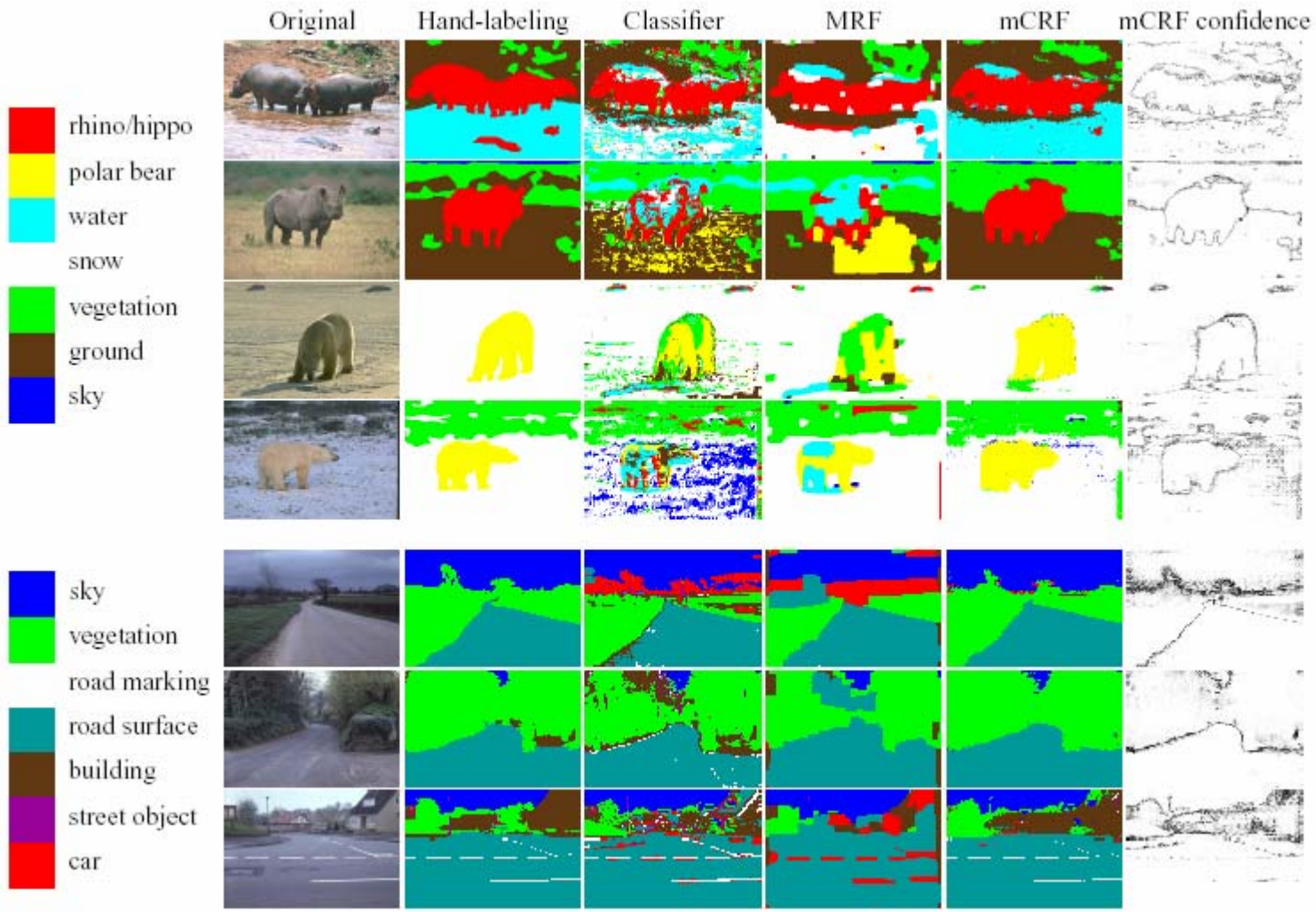
Scene context



mCRF

- Final decision based on
  - Classification (local data-based)
  - Local labels (what relation nearby objects have to each other
  - Image-wide labels (captures coarse scene context)

mCRF Results



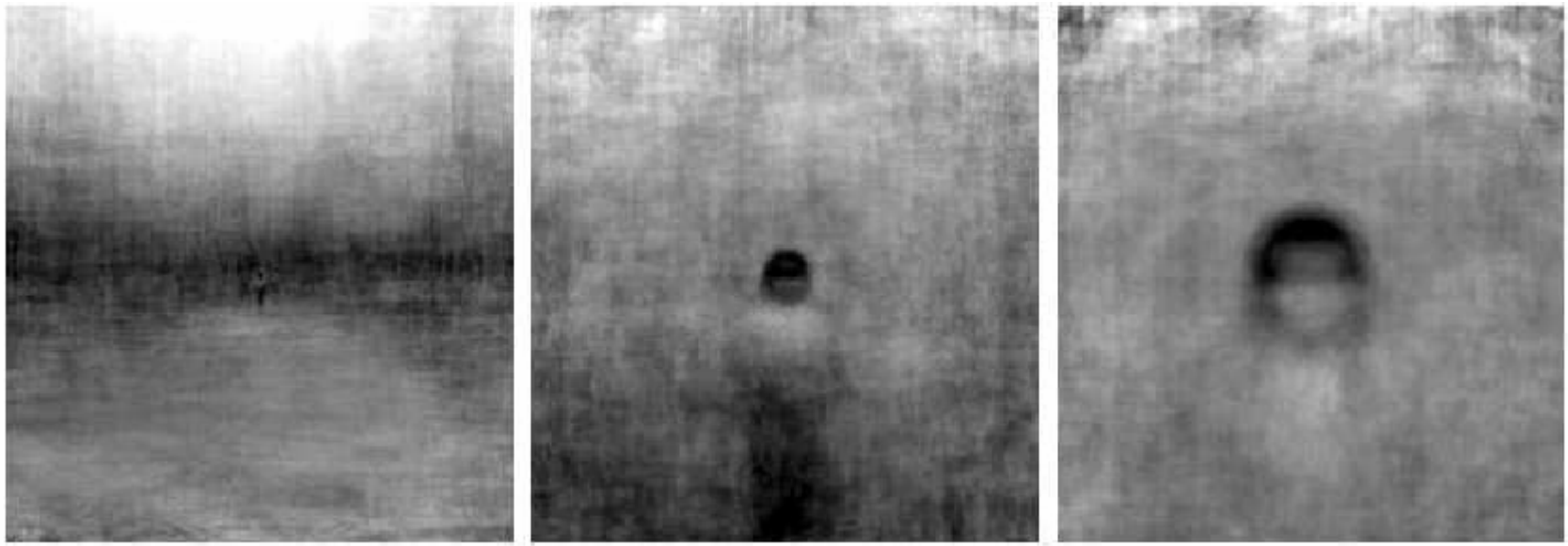
Neighbor-based Context References

- P. Carbonetto, N. Freitas and K. Barnard. “A Statistical Model for General Contextual Object Recognition,” *ECCV*, 2004

- S. Kumar and M. Hebert, “Discriminative Random Fields: A Discriminative Framework for Contextual Interaction in Classification,” *ICCV*, 2003

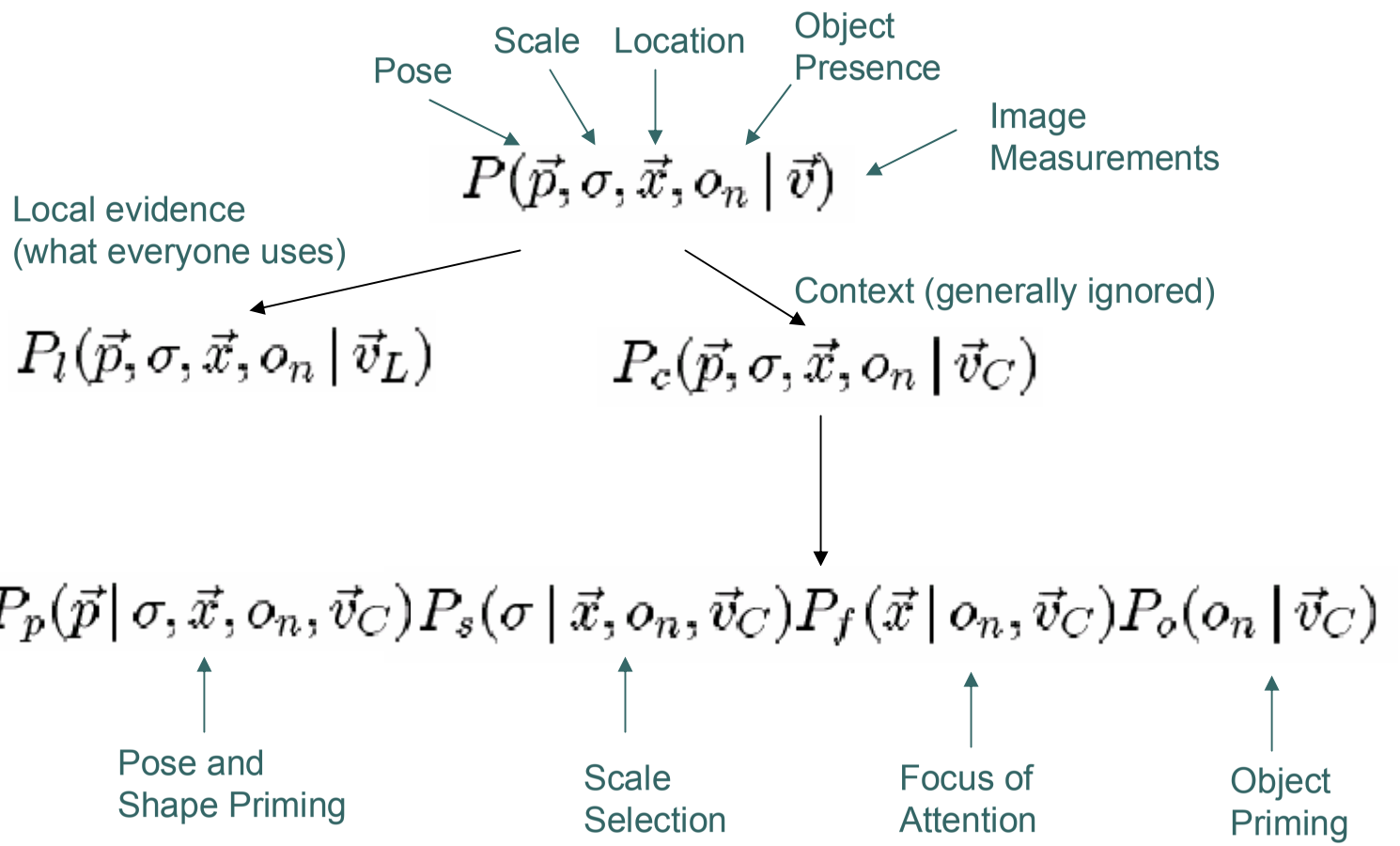
- X. He, R. Zemel and M. Carreira-Perpiñán, “Multiscale Conditional Random Fields for Image Labeling,” *CVPR*, 2004

Scene-based Context



Average pictures containing heads at three scales

Context Priming – Torralba 2001/2003



Getting the Gist of a Scene

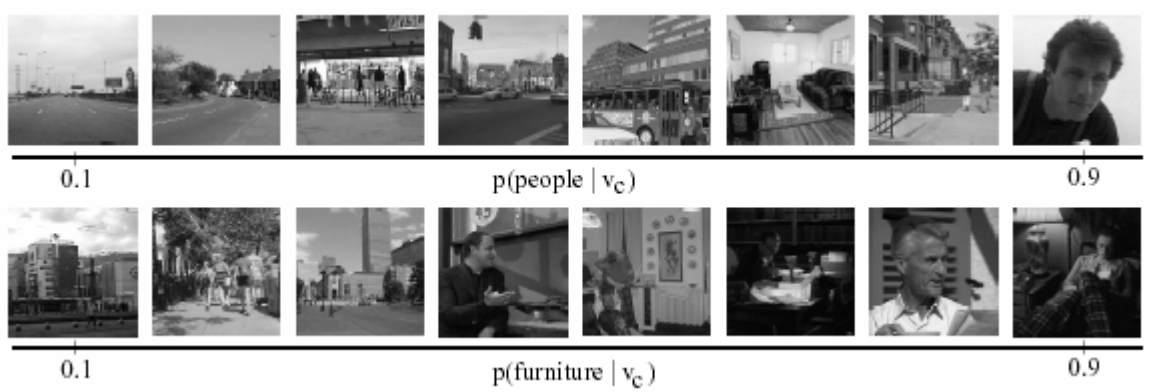
$$P_s(\sigma \mid \vec{x}, o_n, \vec{v}_C) P_f(\vec{x} \mid o_n, \vec{v}_C) P_o(o_n \mid \vec{v}_C)$$

- Simple representation
  - Spectral characteristics (e.g., Gabor filters) with coarse description of spatial arrangement
  - PCA reduction
  - Probabilities modeled with mixture of Gaussians (2003) or logistic regression (Murphy 2003)

Context Priming Results

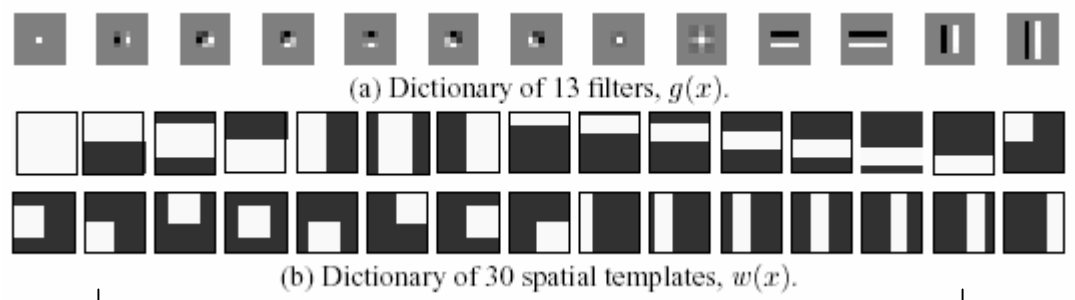
Focus of Attention

Object Presence



Scale Selection

Using the Forrest to See the Trees – Murphy (2003)



Adaboost Patch-Based Object Detector

Gist of Scene (boosted regression)

Detector Confidence at Location/Scale

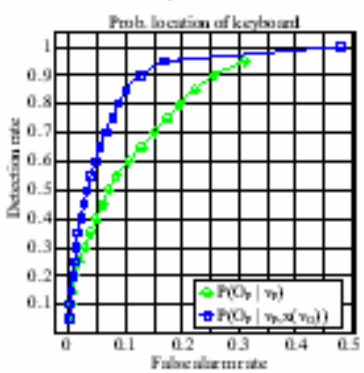
Expected Location/Scale of Object

Combine (logistic regression)

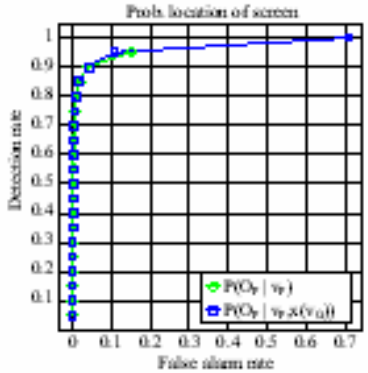
Object Probability at Location/Scale

Object Detection + Scene Context Results

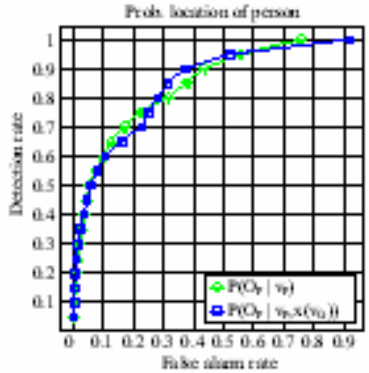
Keyboard



Screen



People



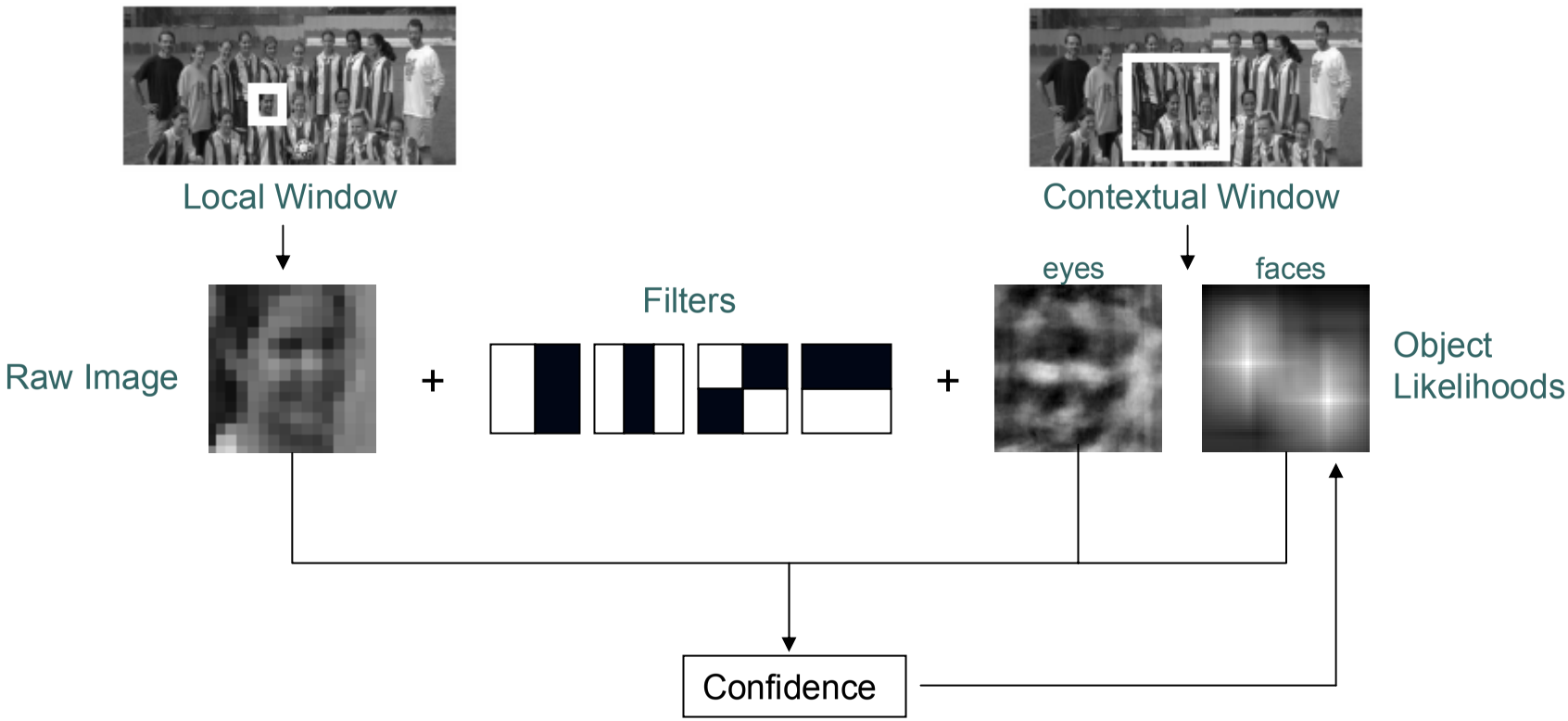
- Often doesn't help that much
- May be due to poor use of context
  - Assumes independence of context and local evidence
  - Only uses expected location/scale from context

Scene-based Context References

- E. Adelson, “On Seeing Stuff: The Perception of Materials by Humans and Machines,” *SPIE*, 2001
- B. Bose and E. Grimson, “Improving Object Classification in Far-Field Video,” *ECCV*, 2004
- K. Murphy, A. Torralba and W. Freeman, “Using the Forrest to See the Trees: A Graphical Model Relating Features, Object, and Scenes,” *NIPS*, 2003
- U. Rutishauser, D. Walther, C. Koch, and P. Perona, “Is bottom-up attention useful for object recognition?,” *CVPR*, 2004
- A. Torralba, “Contextual Priming for Object Detection,” *IJCV*, 2003
- A. Torralba and P. Sinha, “Statistical Context Priming for Object Detection,” *ICCV*, 2001
- A. Torralba, K. Murphy, W. Freeman, and M. Rubin, “Context-Based Vision System for Place and Object Recognition,” *ICCV*, 2003

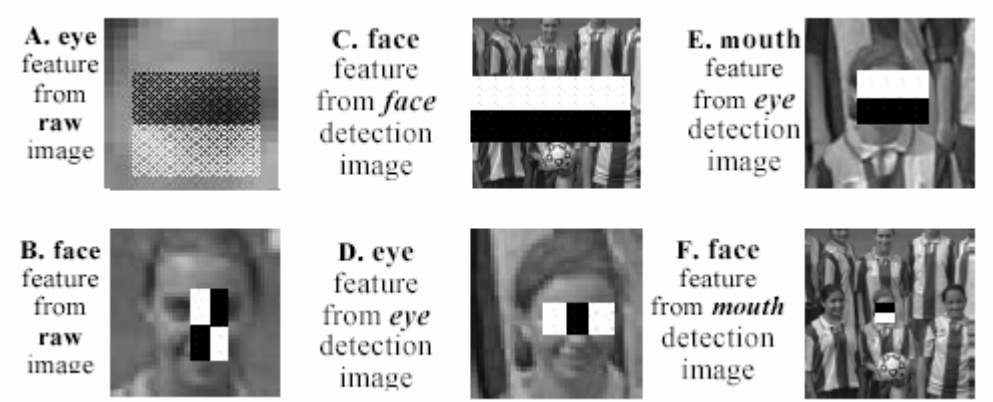
Object-based Context

Mutual Boosting – Fink (2003)

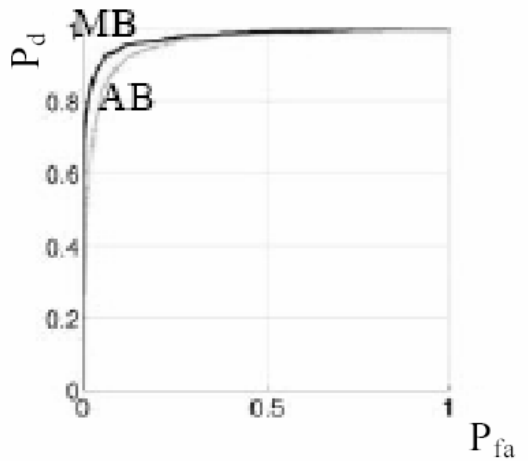


Mutual Boosting Results

Learned Features



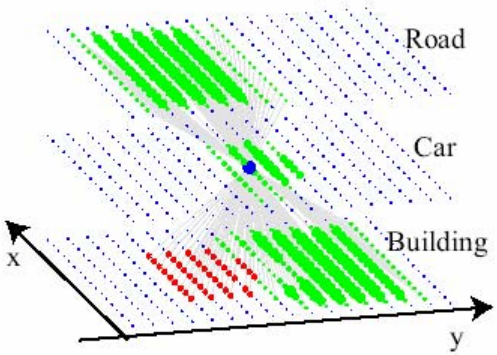
First-Stage Classifier (MIT+CMU)



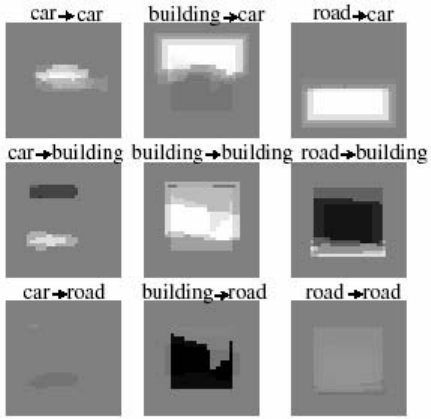
Contextual Models using BRFs – Torralba 2004

- Template features
- Build structure of CRF using boosting
- Other objects' locations' likelihoods propagate through network

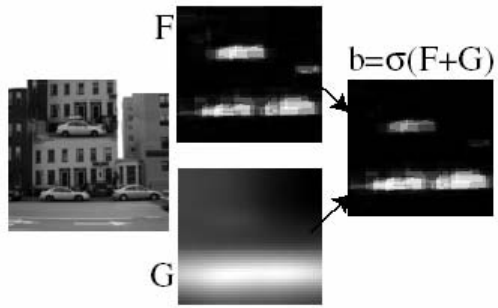
Labeling a Street Scene



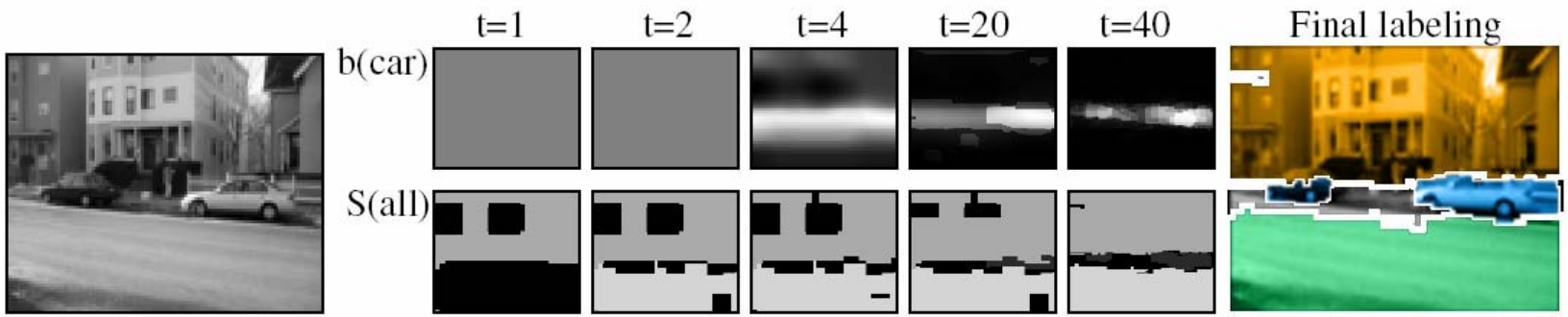
a) Incoming messages to a car node.



b) Compatibilities (W').



c) A car out of context (outside 3rd floor windows) is less of a car.

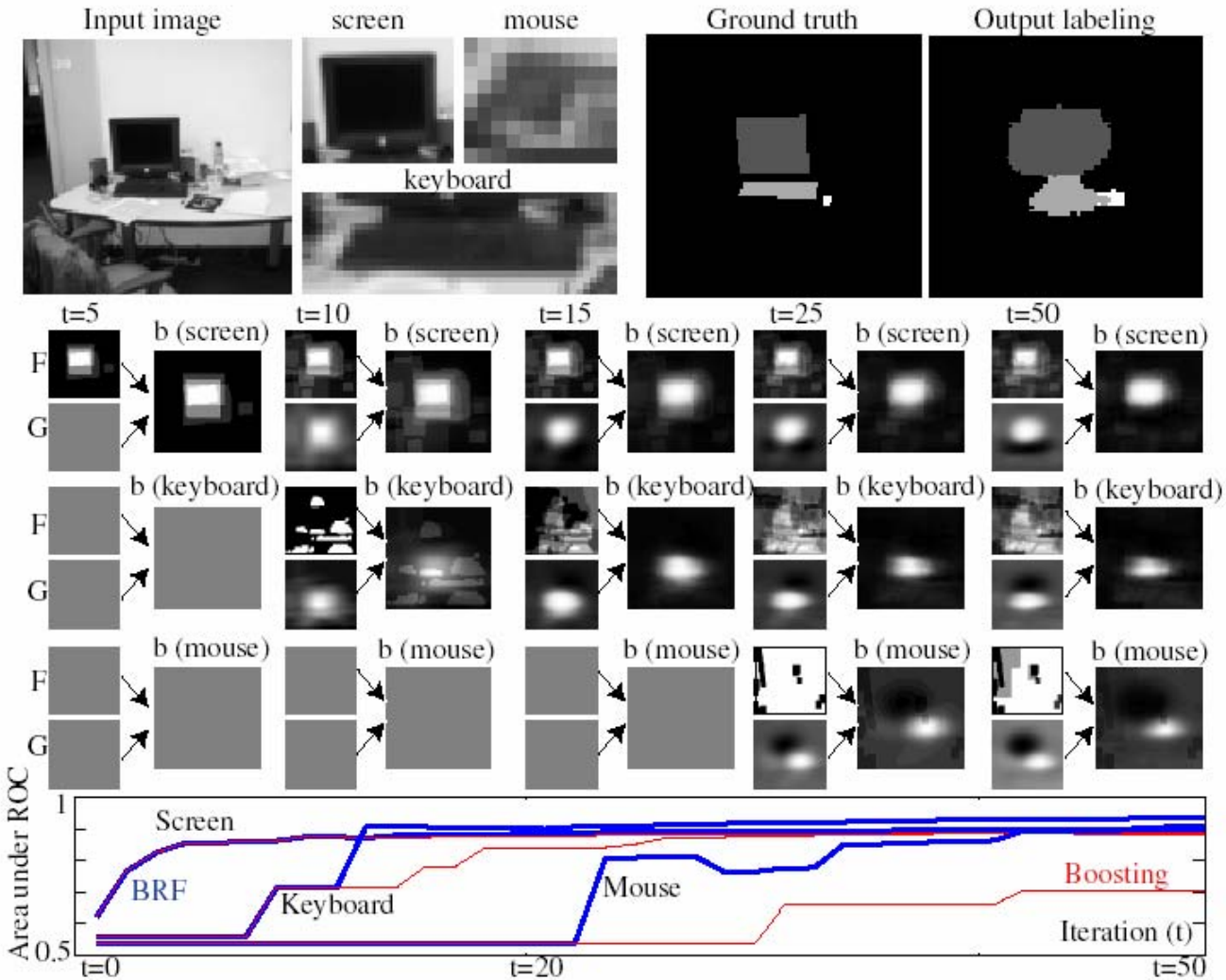


d) Evolution of the beliefs for the car nodes ($b(car)$) and labeling (S) for (road, building, car).

Labeling an Office Scene

F = local

G = compatibility



Object-based Context References

- M. Fink and P. Perona, “Mutual Boosting for Contextual Inference,” *NIPS*, 2003

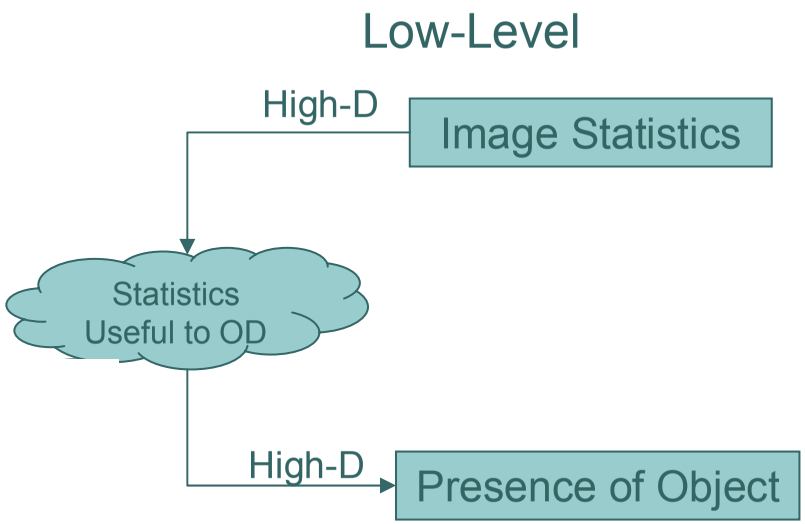
- A. Torralba, K. Murphy, and W. Freeman, “Contextual Models for Object Detection using Boosted Random Fields,” *AI Memo 2004-013*, 2004

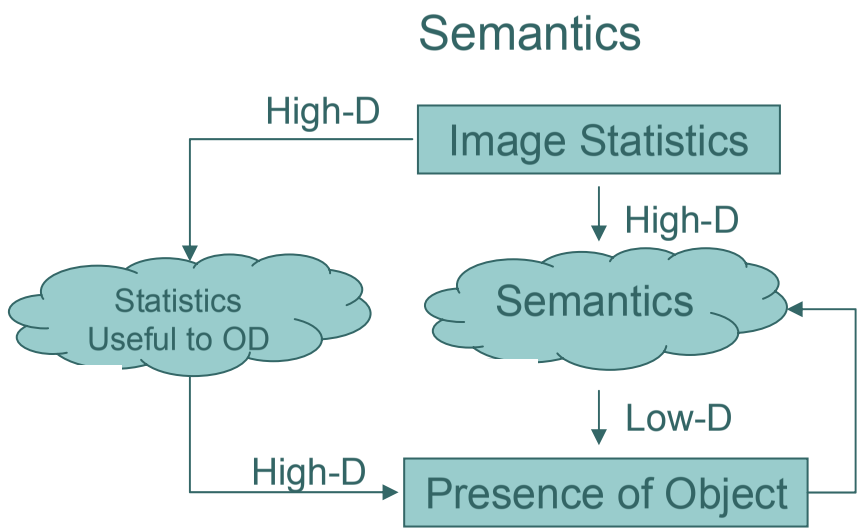
What else can be done?

Scene Structure

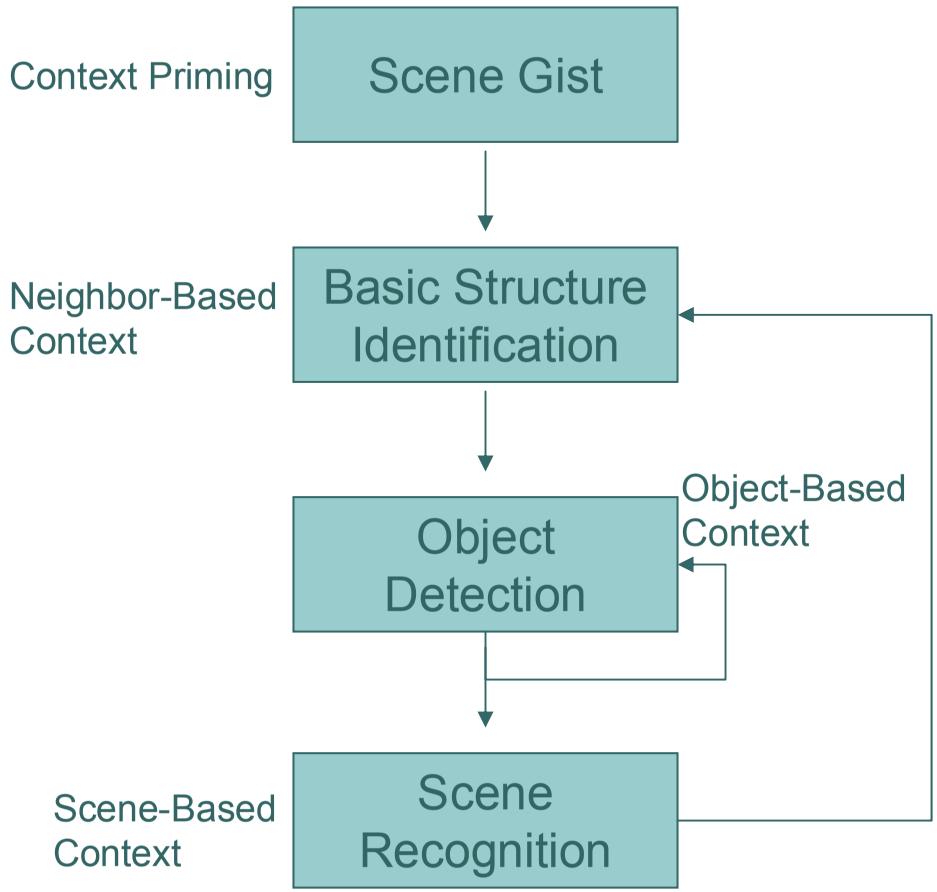
- Improve understanding of scene structure
  - Floor, walls, ceiling
  - Sky, ground, roads, buildings

Semantics vs. Low-level





Putting it all together



Summary

- Neighbor-based context
  - Using nearby labels essential for “complete labeling” tasks
  - Using nearby labels useful even without completely supervised training data
  - Using nearby labels *and* nearby data is better than just using nearby labels
  - Labels can be used to extract local and scene context

Summary

- Scene-based context
  - “Gist” representation suitable for focusing attention or determining likelihood of object presence
  - Scene structure would provide additional useful information (but difficult to extract)
  - Scene label would provide additional useful information

Summary

- Object-based context
  - Even simple methods of using other objects' locations improve results (Fink)
  - Using BRFs, systems can automatically learn to find easier objects first and to use those objects as context for other objects

Conclusions

- General
  - Few object detection researchers use context
  - Context, when used effectively, can improve results dramatically
  - A more integrated approach to use of context and data could improve image understanding

References

- E. Adelson, “On Seeing Stuff: The Perception of Materials by Humans and Machines,” *SPIE*, 2001
- B. Bose and E. Grimson, “Improving Object Classification in Far-Field Video,” *ECCV*, 2004
- P. Carbonetto, N. Freitas and K. Barnard. “A Statistical Model for General Contextual Object Recognition,” *ECCV*, 2004
- M. Fink and P. Perona, “Mutual Boosting for Contextual Inference,” *NIPS*, 2003
- X. He, R. Zemel and M. Carreira-Perpiñán, “Multiscale Conditional Random Fields for Image Labeling,” *CVPR*, 2004
- S. Kumar and M. Hebert, “Discriminative Random Fields: A Discriminative Framework for Contextual Interaction in Classification,” *ICCV*, 2003
- J. Lafferty, A. McCallum and F. Pereira, “Conditional random fields: Probabilistic models for segmenting and labeling sequence data,” *ICML*, 2001
- K. Murphy, A. Torralba and W. Freeman, “Using the Forrest to See the Trees: A Graphical Model Relating Features, Object, and Scenes,” *NIPS*, 2003
- U. Rutishauser, D. Walther, C. Koch, and P. Perona, “Is bottom-up attention useful for object recognition?,” *CVPR*, 2004
- A. Torralba, “Contextual Priming for Object Detection,” *IJCV*, 2003
- A. Torralba and P. Sinha, “Statistical Context Priming for Object Detection,” *ICCV*, 2001
- A. Torralba, K. Murphy, and W. Freeman, “Contextual Models for Object Detection using Boosted Random Fields,” *AI Memo 2004-013*, 2004
- A. Torralba, K. Murphy, W. Freeman, and M. Rubin, “Context-Based Vision System for Place and Object Recognition,” *ICCV*, 2003