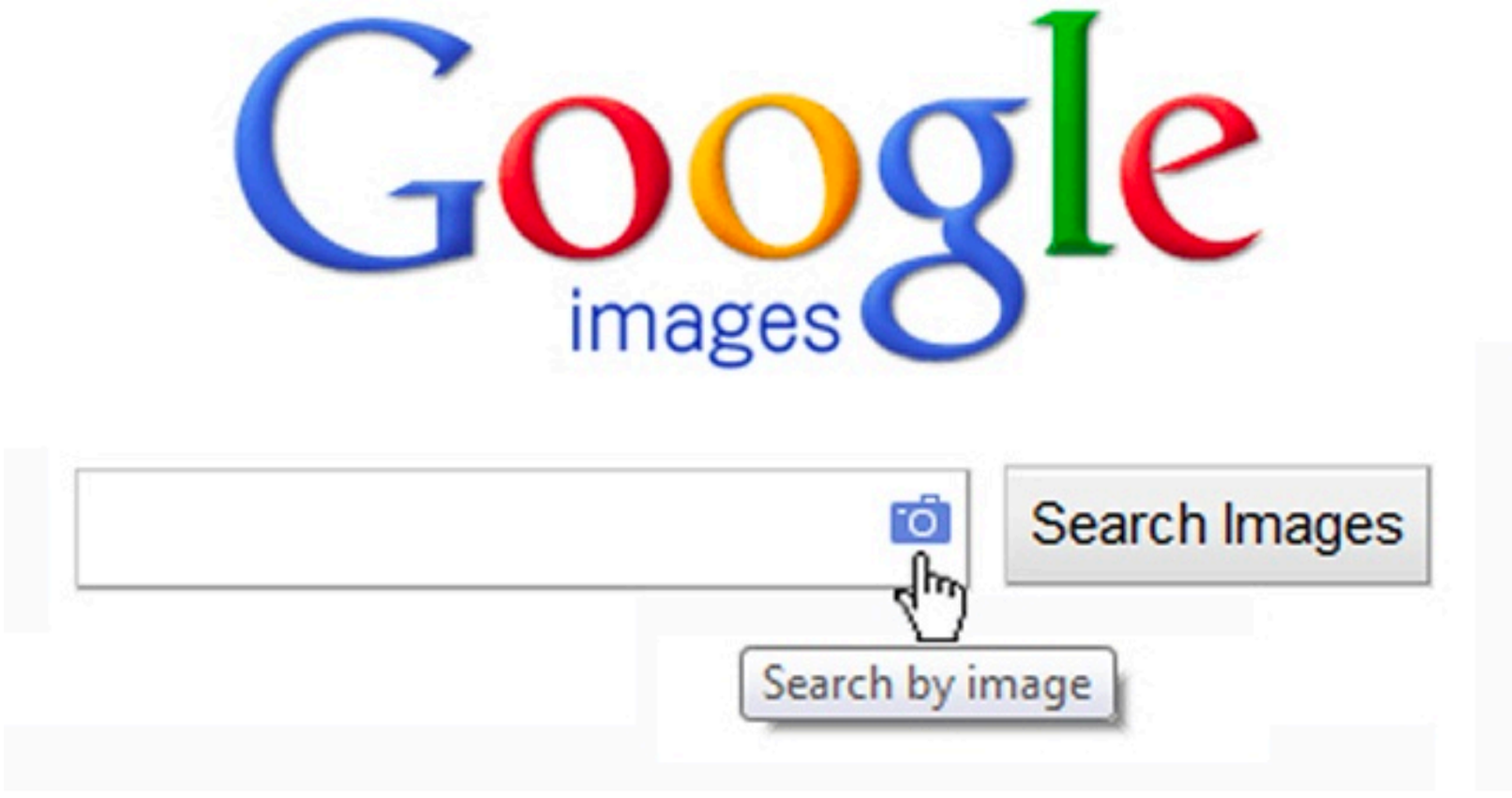


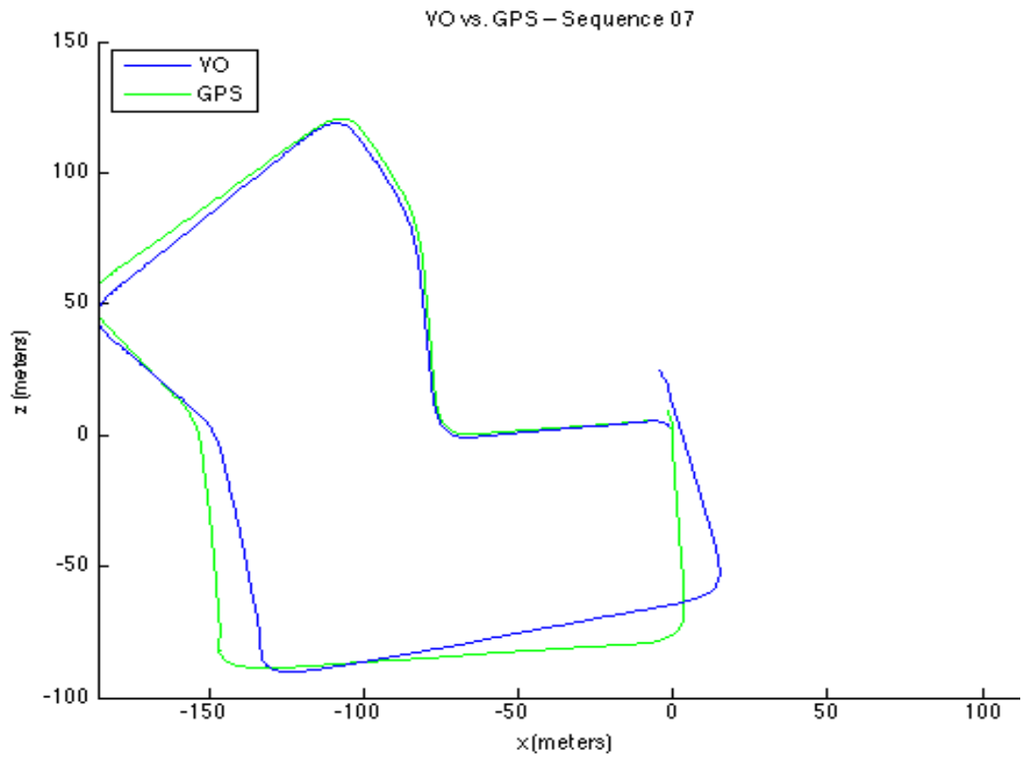
16.485: VNAV - Visual Navigation for Autonomous Vehicles

Luca Carlone

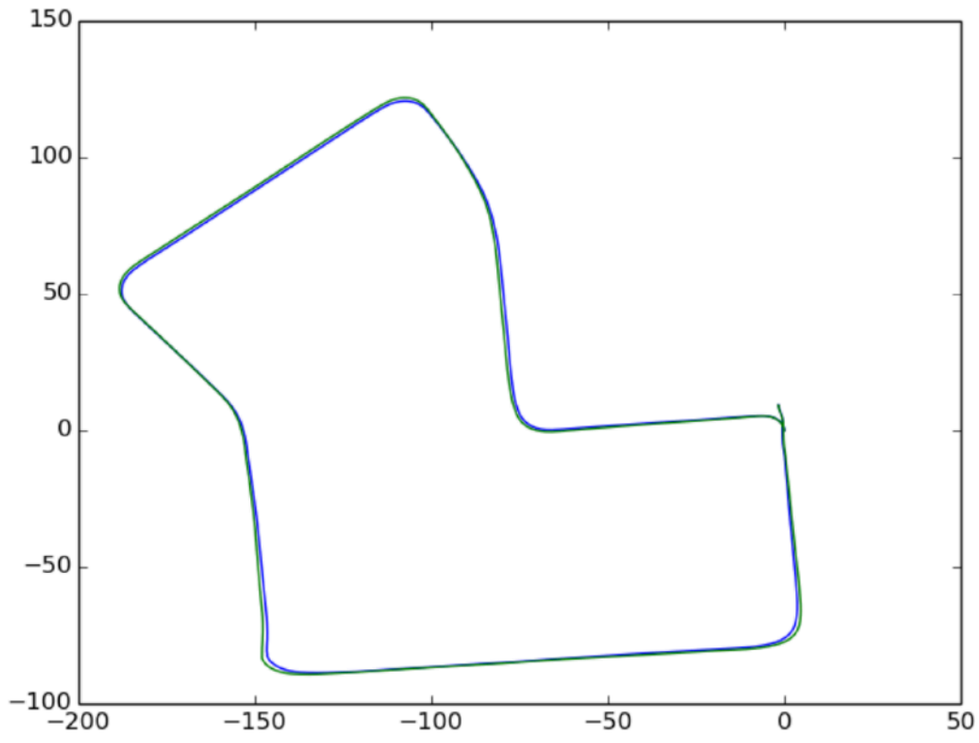
Lecture 21: Place Recognition

Recap

Visual odometry



SLAM



SLAM (Simultaneous Localization and Mapping) requires:

- place recognition => loop closure detection

and / or

- Object detection => landmark detection

Need for loop closure

Today + Next Lecture

- **Place recognition**
- **Object detection / recognition**

Visual Place Recognition: A Survey

Stephanie Lowry, Niko Sünderhauf, Paul Newman, *Fellow, IEEE*, John J. Leonard, *Fellow, IEEE*, David Cox, Peter Corke, *Fellow, IEEE*, and Michael J. Milford, *Member, IEEE*

***Abstract*—Visual place recognition is a challenging problem due to the vast range of ways in which the appearance of real-world places can vary. In recent years, improvements in visual sensing capabilities, an ever-increasing focus on long-term mobile robot autonomy, and the ability to draw on state-of-the-art research in other disciplines—particularly recognition in computer vision and animal navigation in neuroscience—have all contributed to significant advances in visual place recognition systems. This paper presents a survey of the visual place recognition research landscape. We start by introducing the concepts behind place recognition—the role of place recognition in the animal kingdom, how a "place" is defined in a robotics context, and the major components of a place recognition system. Long-term robot operations have revealed that changing appearance can be a significant factor in visual place recognition failure; therefore, we discuss how place recognition solutions can implicitly or explicitly account for appearance change within the environment. Finally, we close with a discussion on the future of**

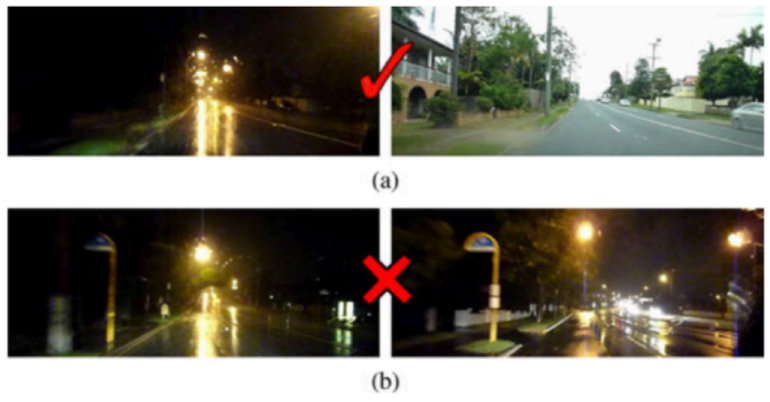
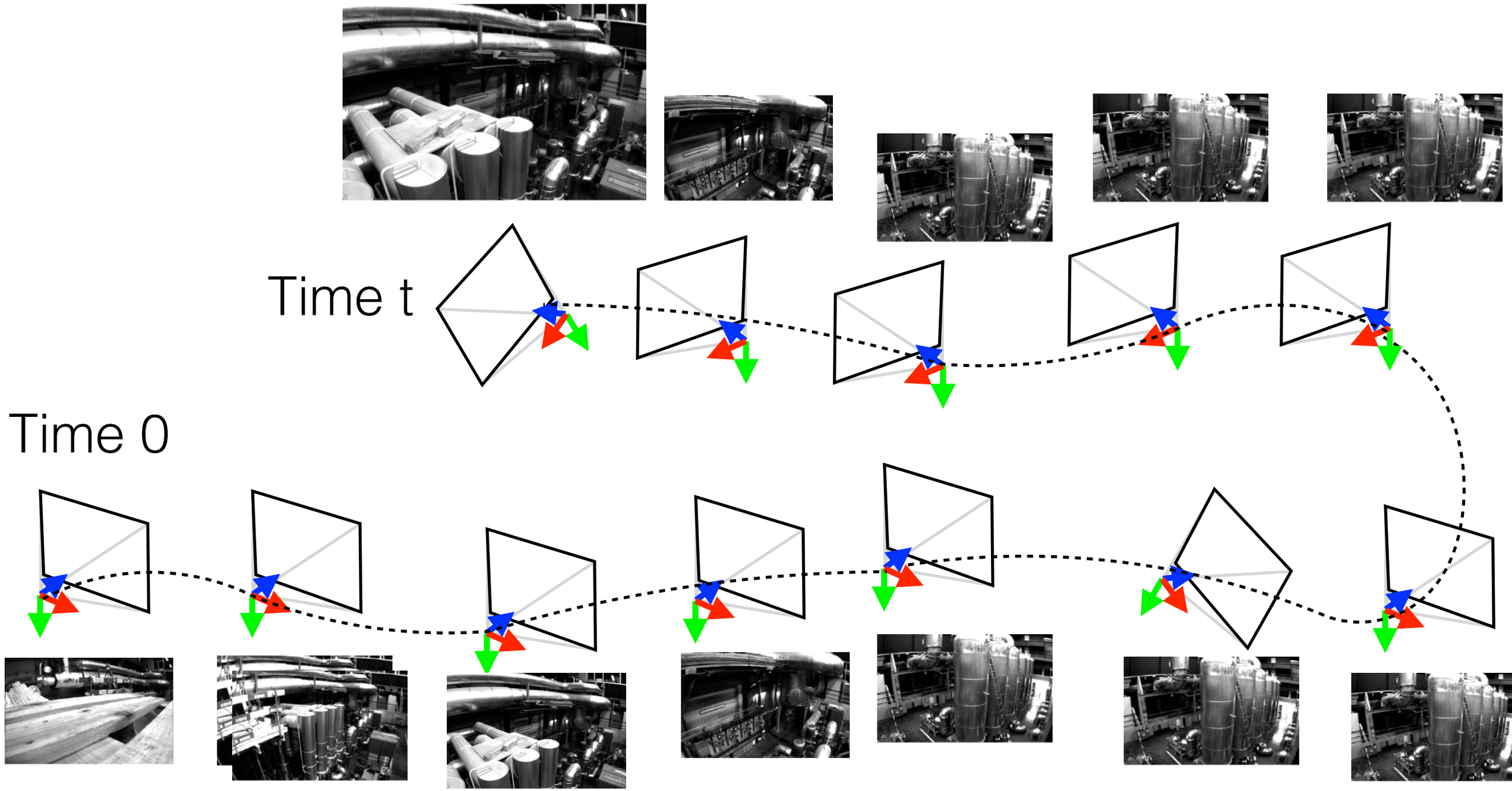


Fig. 1. Visual place recognition systems must be able to (a) successfully match very perceptually different images while (b) also rejecting incorrect matches between aliased image pairs of different places.

+ a few more recent papers

Place Recognition & Image Retrieval

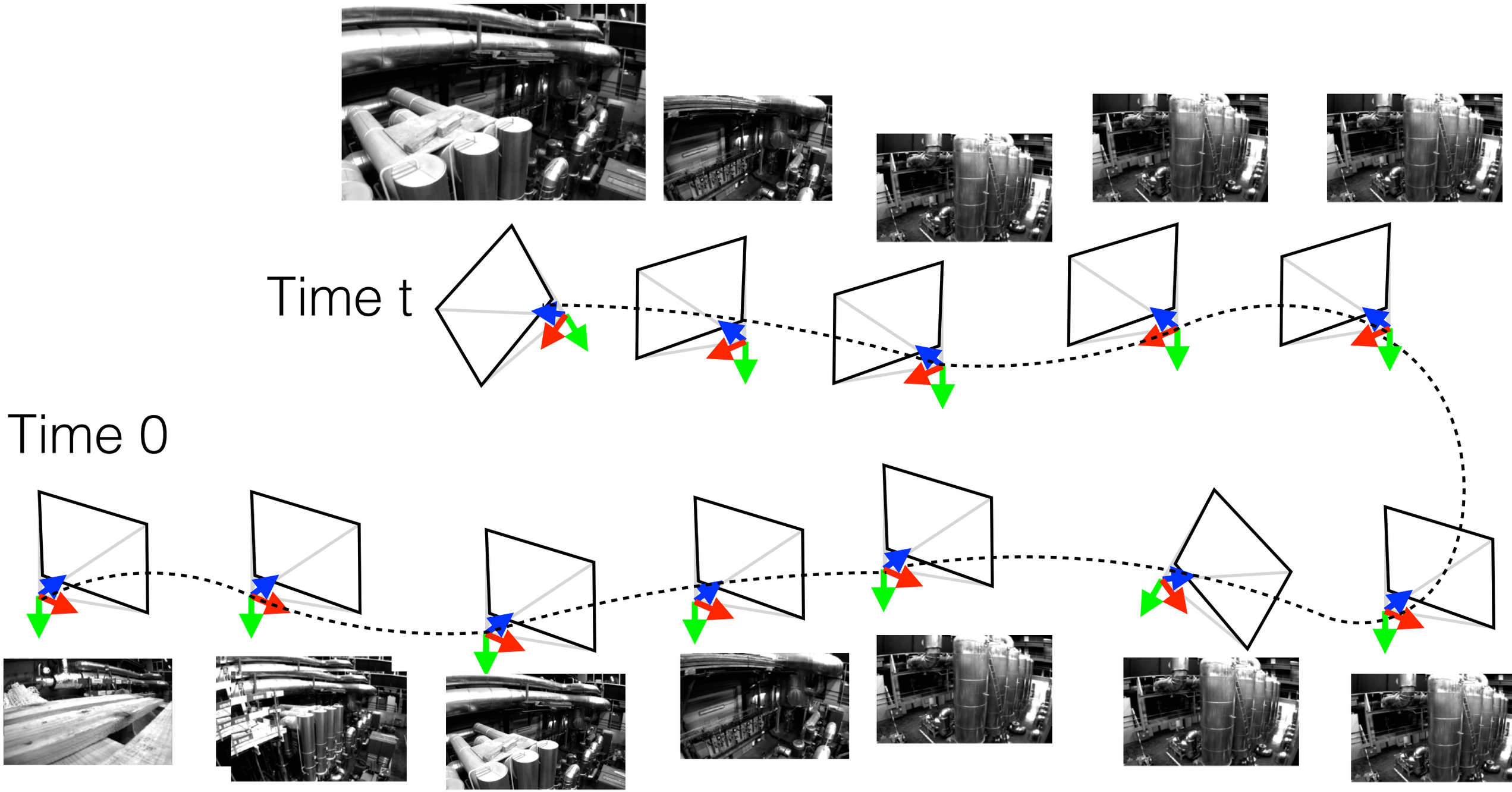


Does the image at time “t” picture a place seen in previous images?

Place Recognition: Challenges

- **Appearance changes**:
  - Illumination
  - Weather conditions
  - Dynamic objects (people, cars,…)
  - Viewpoint changes
- **Perceptual aliasing**: two different places may look similar (building, roads, …)

A brute force approach



Scalability is crucial..

Image retrieval/place recognition vs. pose estimation

Image Retrieval

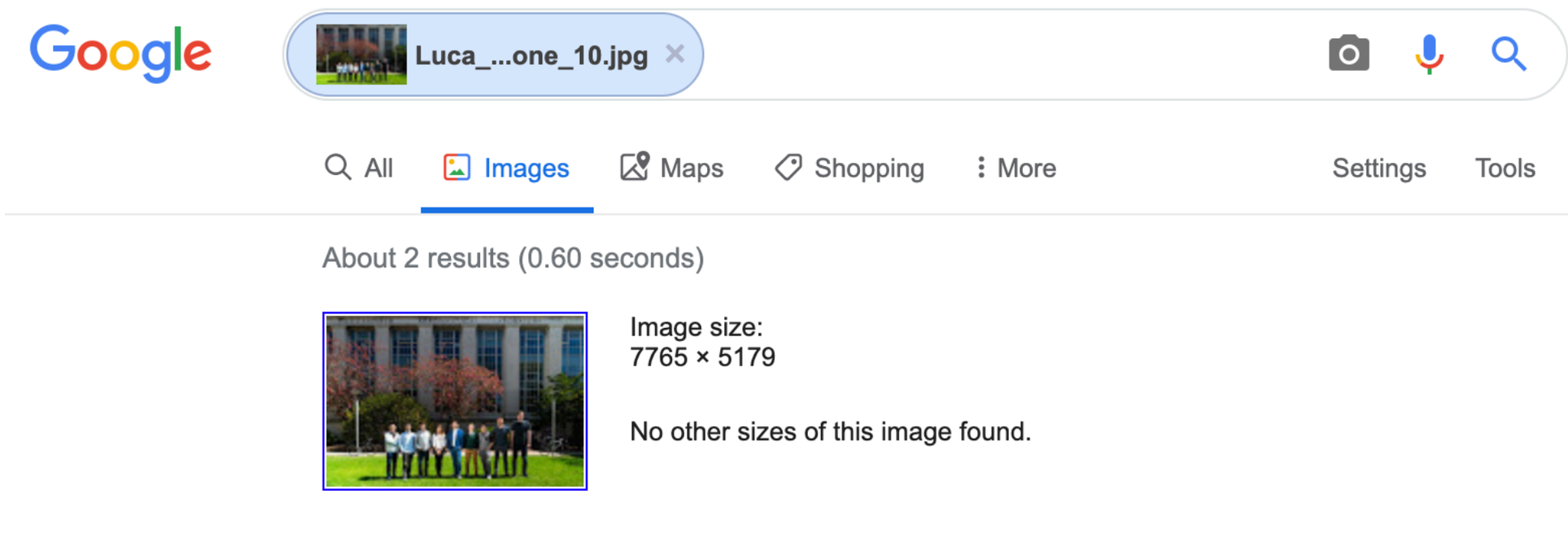
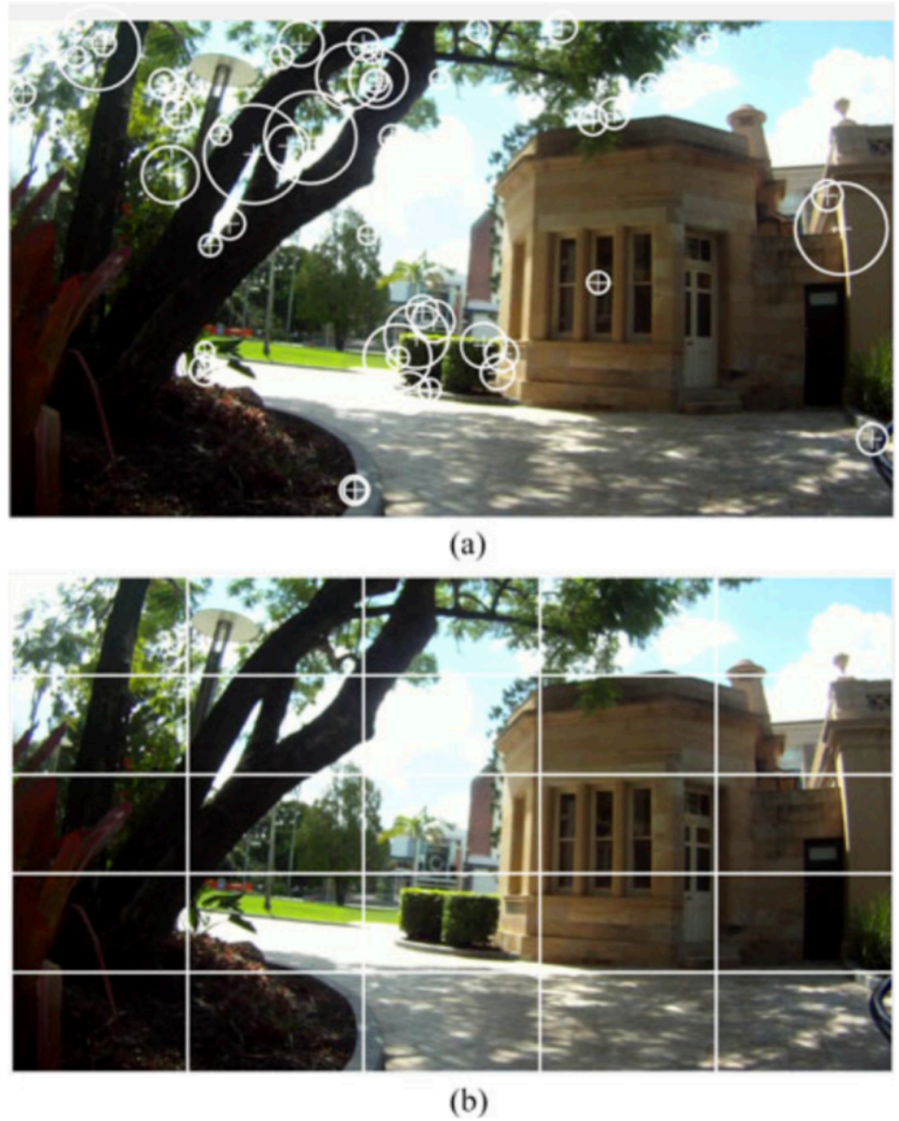


Image Retrieval: Approaches

- Local descriptors

- Global descriptors

- Learning-based methods



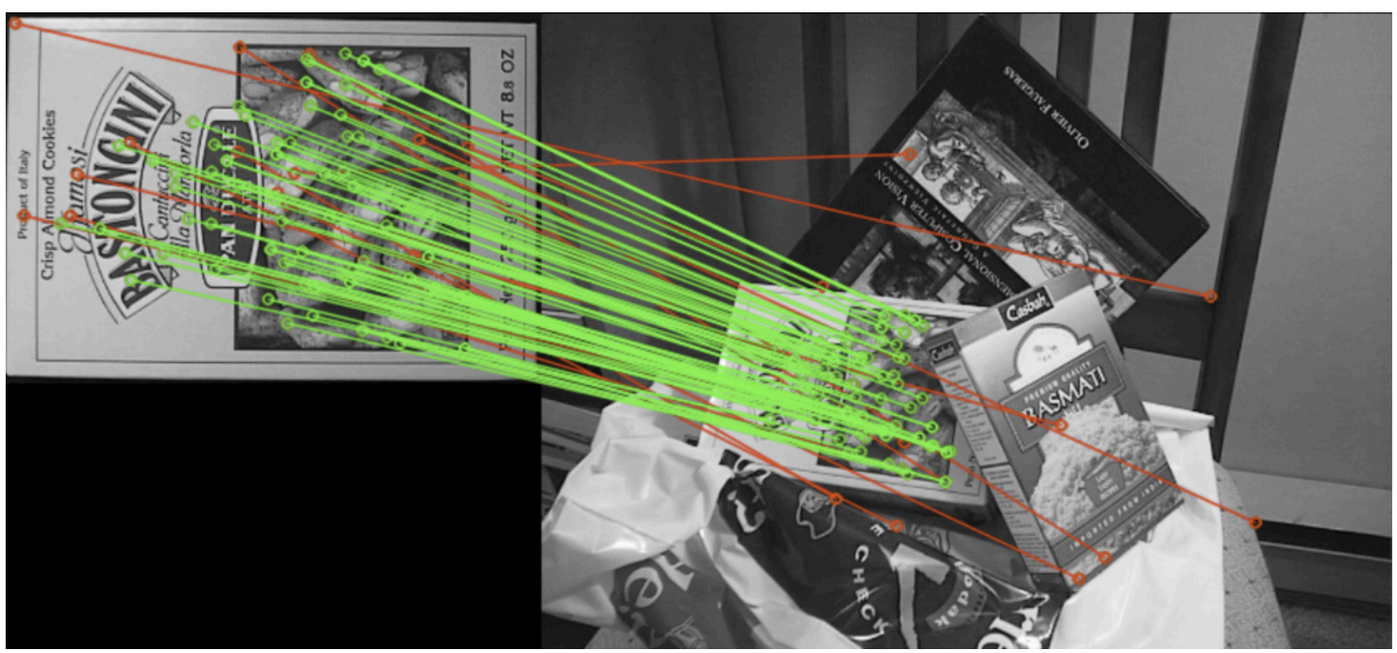
(a)

(b)

[courtesy of Lowry'16]

Local descriptors

SIFT, SURF, ORB, Brief, …



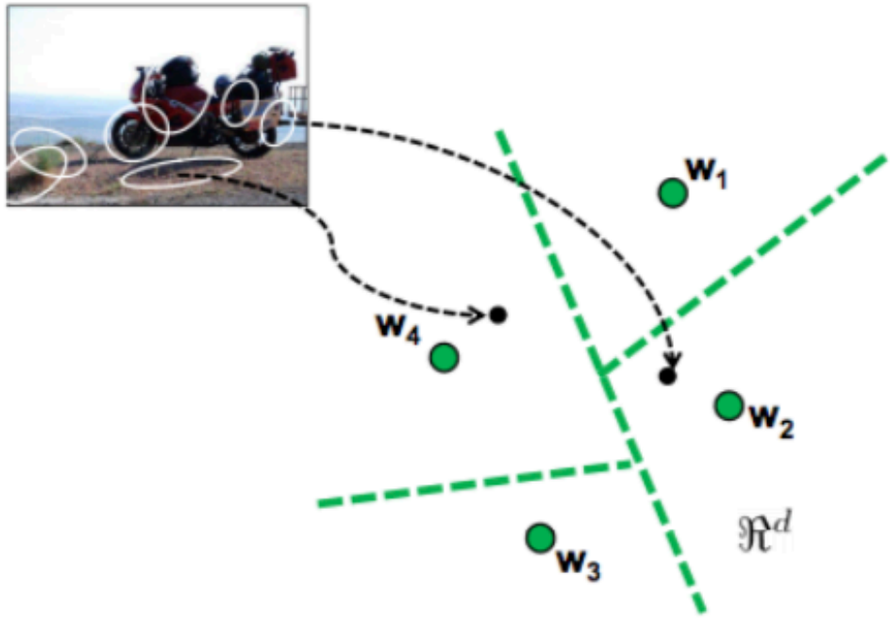
Naive approach: stack all descriptors in a vector

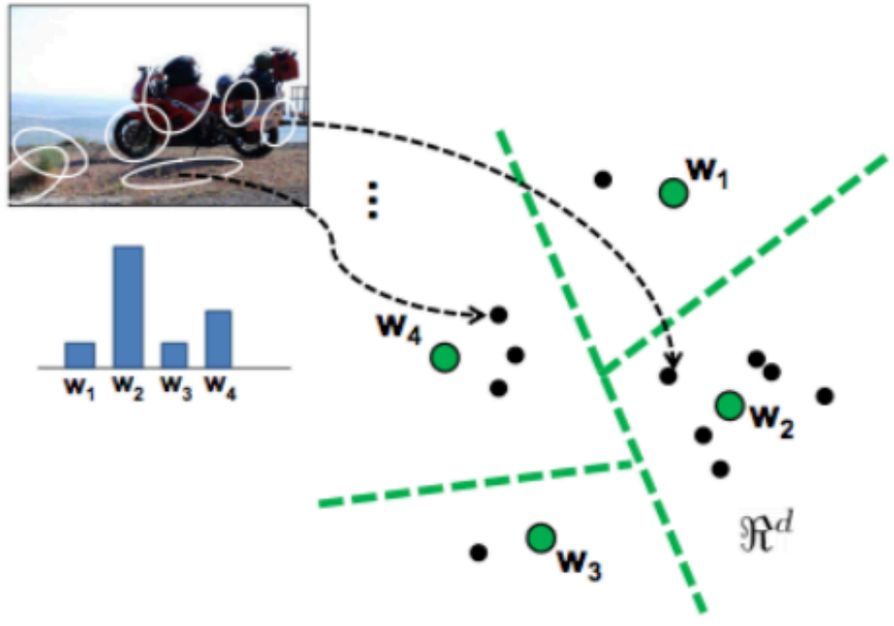
Is this a good image descriptor?

Local descriptors: Bag of Words

Based on text retrieval and summarization methods

1) Extract features and descriptors in image
2) Discretize feature space (clustering)
3) Store the frequency of the features for each image

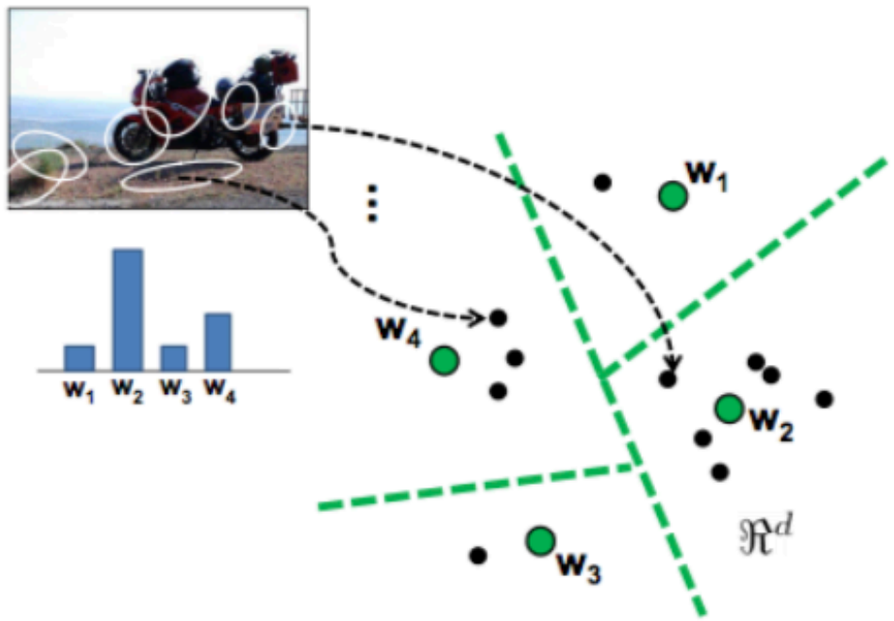


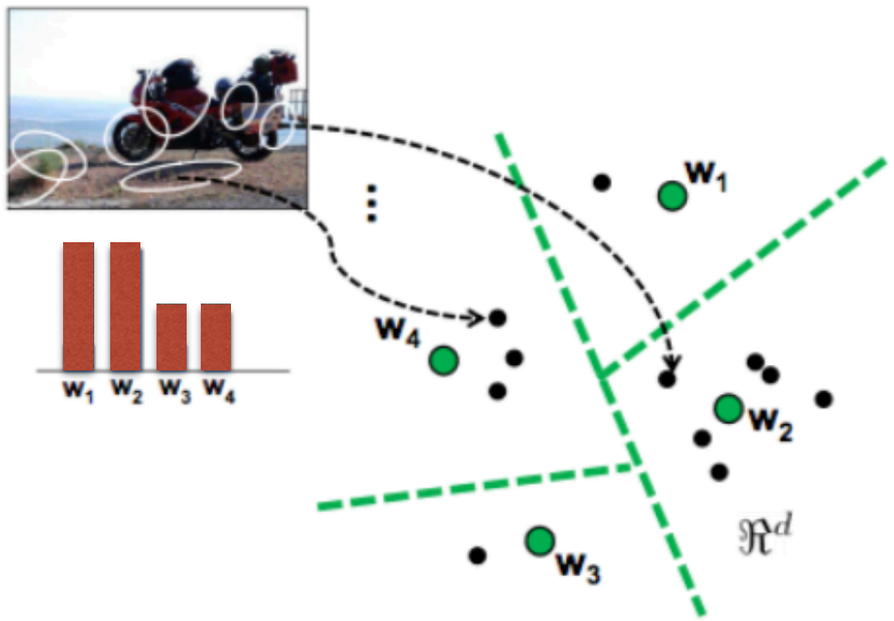


Each cluster is a “visual word”
Typically 5k-10k (up to 100k) visual words

J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *ICCV*, 2003.

Local descriptors: Bag of Words

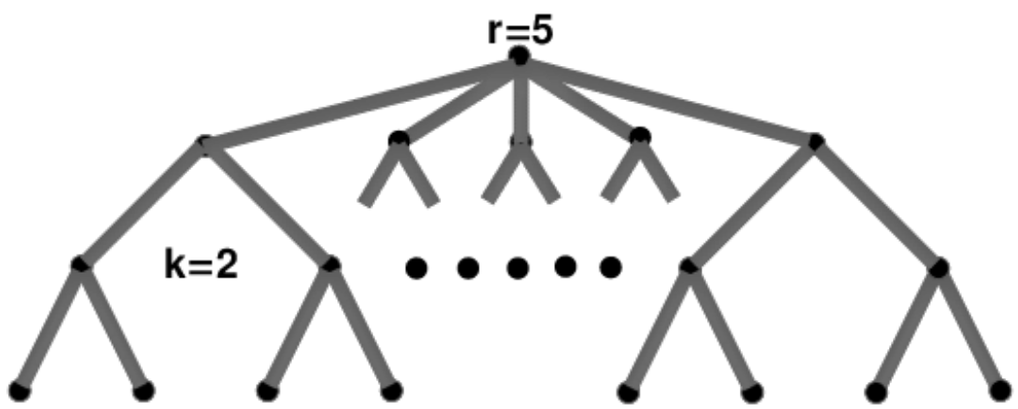




Two images are compared based on the corresponding histogram (Hamming distances, other metrics, …)

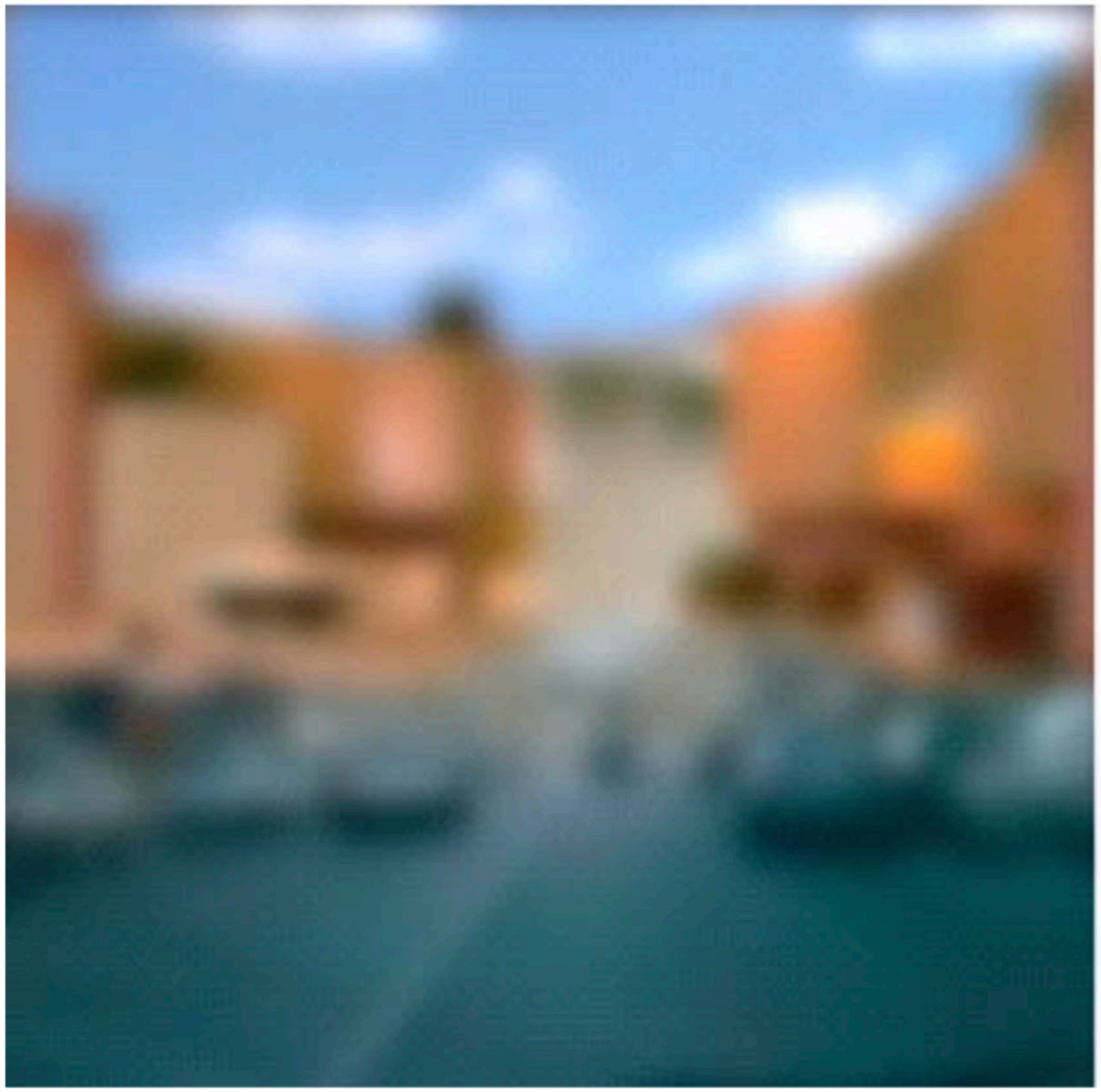
Faster version: **vocabulary tree**

Alternatives**: VLAD** (Vector of Locally Aggregated Descriptors), **Fisher vectors**



Global descriptors

From collection of features/objects to global properties:

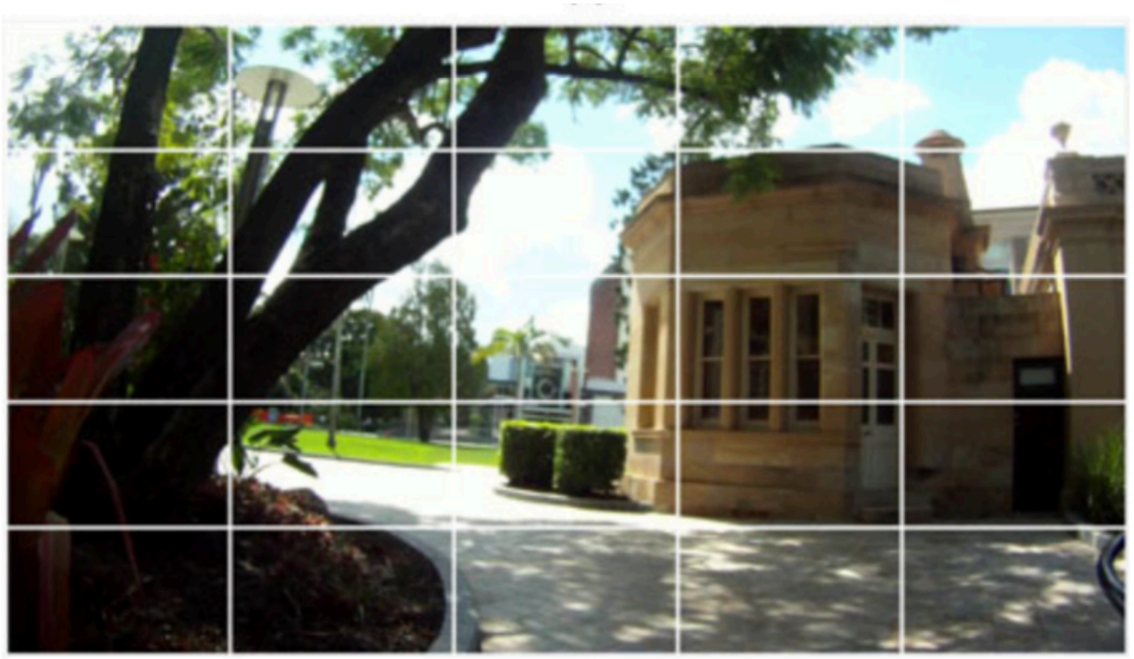


A. Oliva and A. Torralba, Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope, IJCV, 2001.

Global descriptors

Early approaches:

- color histograms
- principal component analysis
- other statistics on edges, corners, and color patches



Early 2000:

- **GIST descriptor**:
  - image is filtered at different orientations and different frequencies to extract information from the image
  - results are averaged to generate a compact vector that represents the “gist” of a scene

A. Oliva and A. Torralba, Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope, IJCV, 2001.

Visual Experiment

Visual Experiment

Visual Experiment

Visual Experiment

What was the content of the image?

- A: building
- B: beach
- C: dog
- D: car

Global descriptors

[Oliva, Torralba'01]:

- **20 ms**: observers used the low spatial frequency part of hybrids (street in Fig. B)
- **150 ms**: observers categorized the image on the basis of the high spatial frequencies (e.g., beach in Fig. B)

A. Oliva and A. Torralba, Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope, IJCV, 2001.

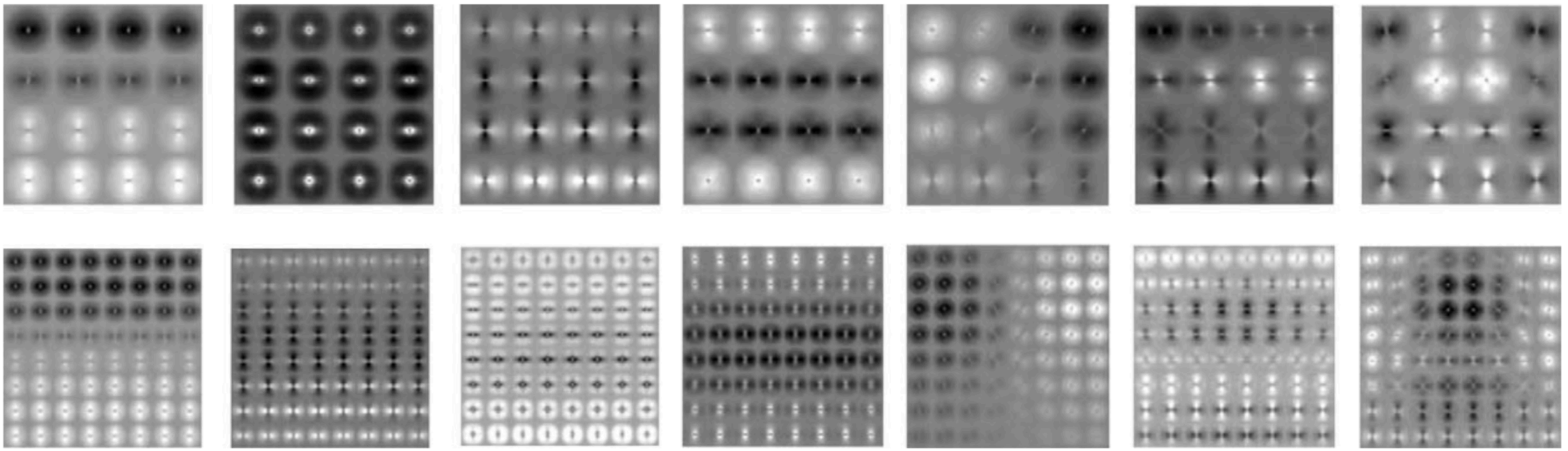
Global descriptors

[Oliva, Torralba'01]: evidence that visual input is processed at different spatial scales (from low to high spatial frequency):

- Low frequency: less sensitive to noise and nuisances, but also less details
- High-frequency: finer details

A. Oliva and A. Torralba, Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope, IJCV, 2001.

GIST



- Compute weights by doing Principal Component on the responses to multi-scale filters
- Weights mapped to properties (openness, naturalness, roughness, expansion, …)

A. Oliva and A. Torralba, Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope, IJCV, 2001.

GIST and Spatial Envelop

Flat view of a man-made environment,
vertically structured.
Small space with large elements.

Flat view of a man-made semi-closed
urban environment.

Flat view of a man-made closed
urban environment.
Large space with small elements.

Perspective view of a man-made
open environment.

Man-made open environment.

Man-made closed urban environment.

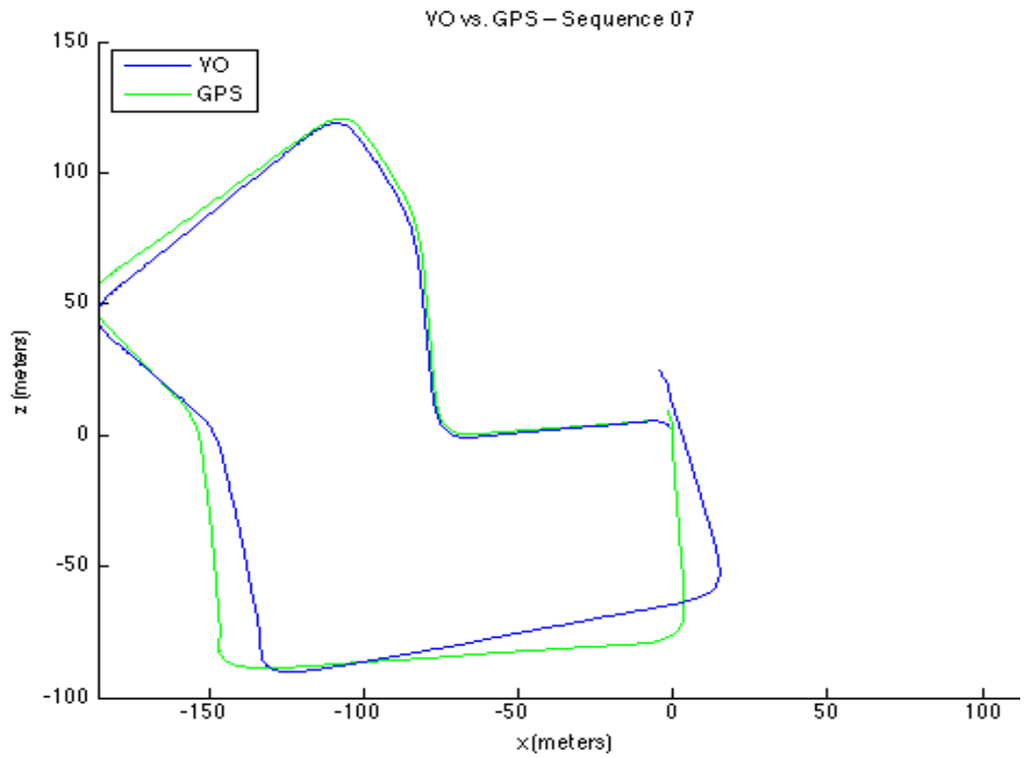
Perspective view of a man-made closed
urban environment.
Large space with small elements.

Flat view of a man-made urban
environment, vertically structured.

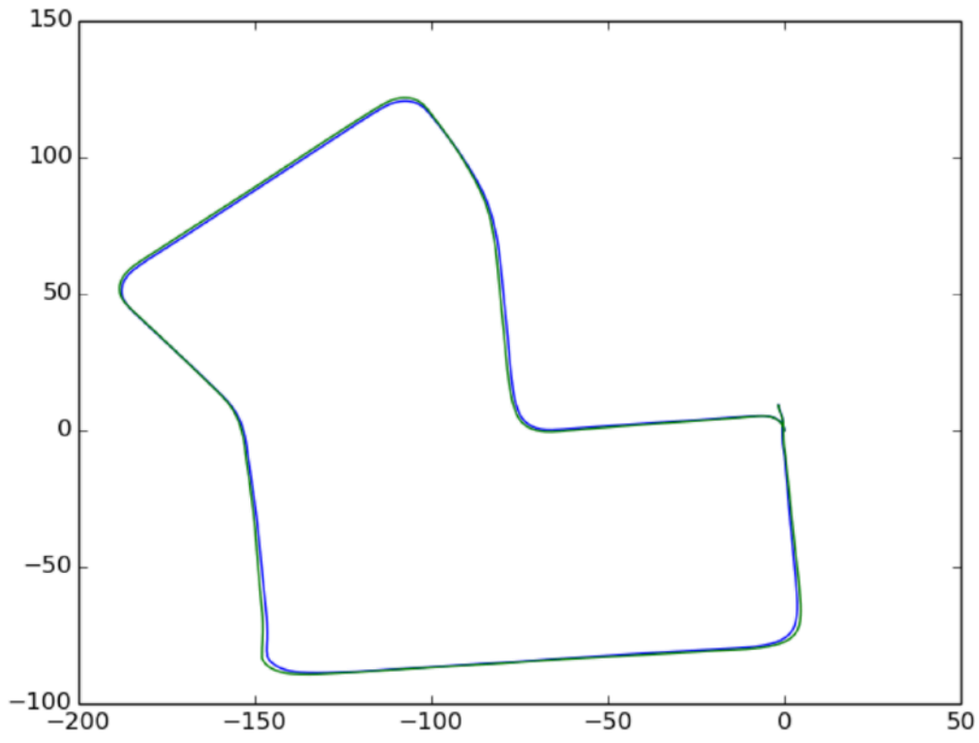
A. Oliva and A. Torralba, Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope, IJCV, 2001.

Recap

Visual odometry



SLAM



SLAM (Simultaneous Localization and Mapping) requires:

- place recognition => loop closure detection

and / or

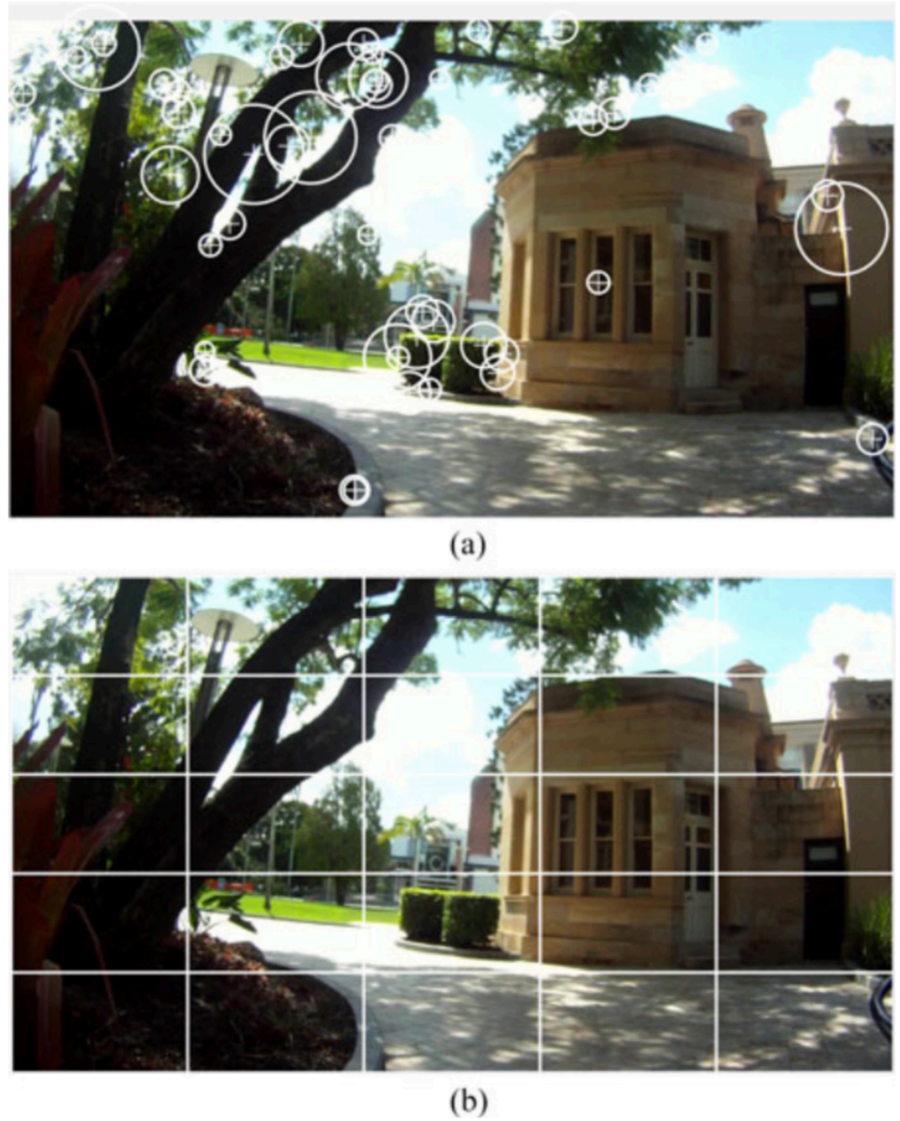
- object detection => landmark detection

Image Retrieval: Approaches

- Local descriptors

- Global descriptors

- Learning-based methods



(a)

(b)

[courtesy of Lowry'16]

Local vs. Global Descriptors

Local descriptors:

- allow estimating feature (and camera) geometry
- sensitive to lighting conditions and seasonal variations

Global descriptors:

- better at handling lighting conditions and seasonal variation
- more sensitive to viewpoint changes

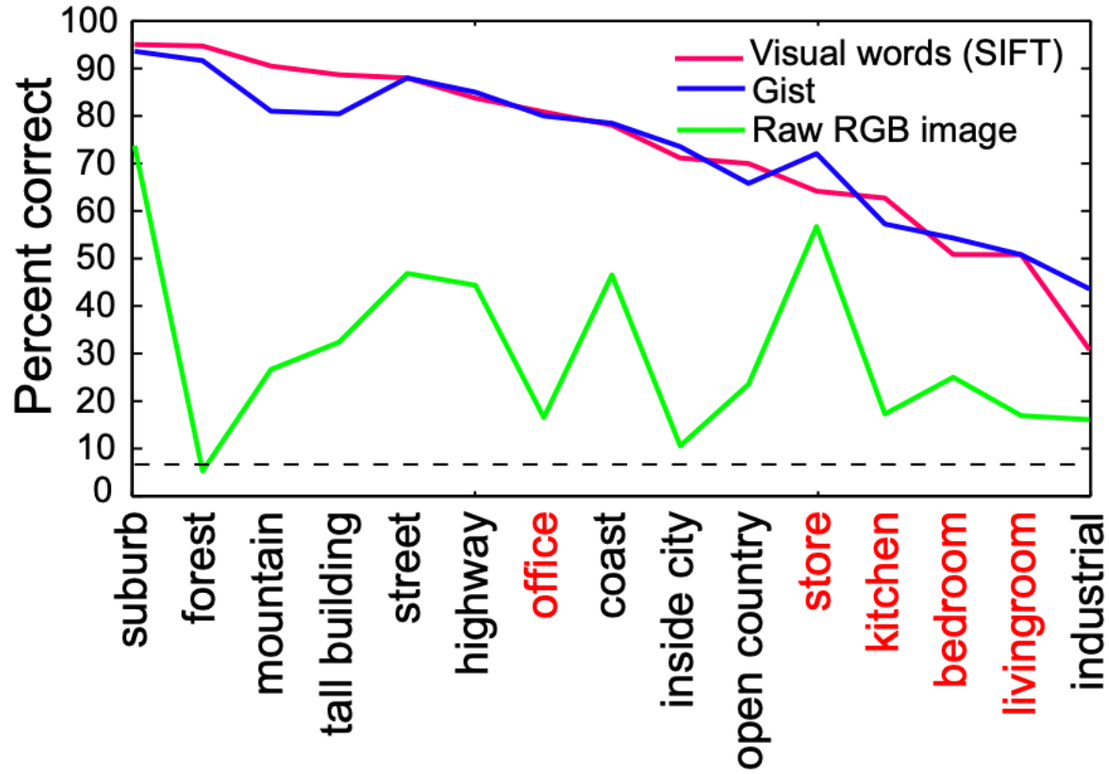


Figure 1. Comparison of Spatial Sift and Gist features for a scene recognition task. Both set of features have a strong correlation in the performance across the 15 scene categories. Average performance for the different features are: Gist: 73.0%, Pyramid matching: 73.4%, bag of words: 64.1%, and color pixels (SSD): 30.6%. In all cases we use an SVM.

A. Quattoni and A. Torralba, Recognizing Indoor Scenes, CVPR'09.

Deep Learning Revolution

-new take on algorithms
-large amount of data
-large amount of computing (GPU)

2010: ImageNet Large Scale Visual Recognition Challenge (ILSVRC) is launched

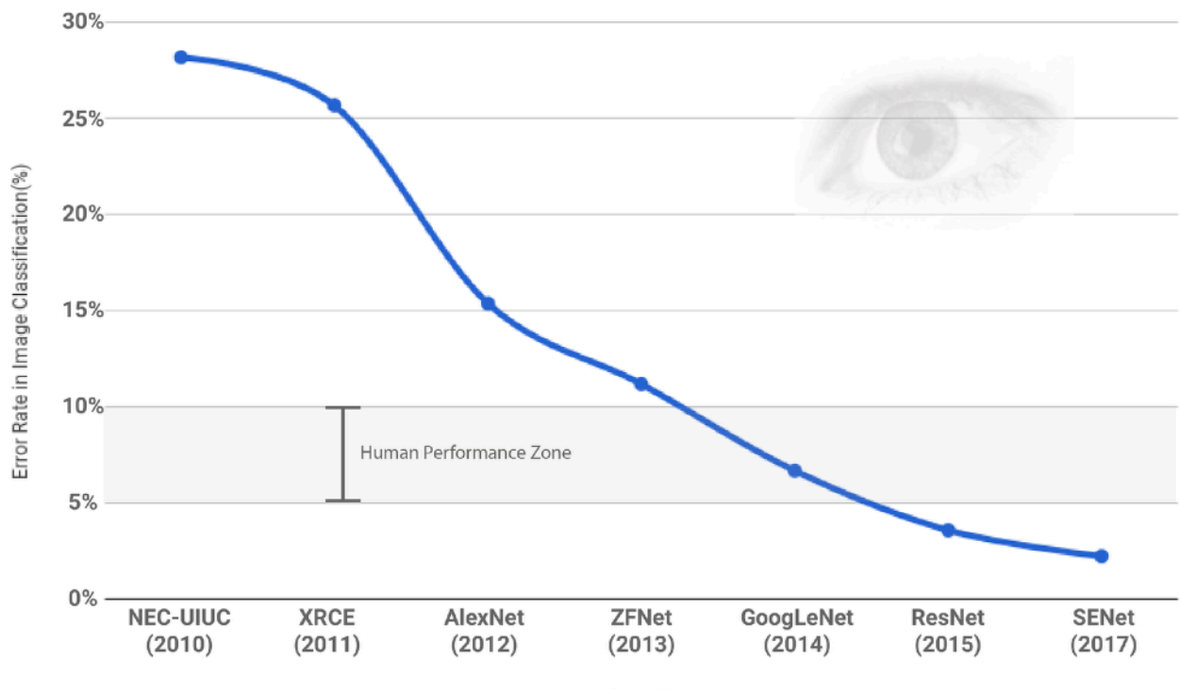
14M images

objects/ bounding boxes

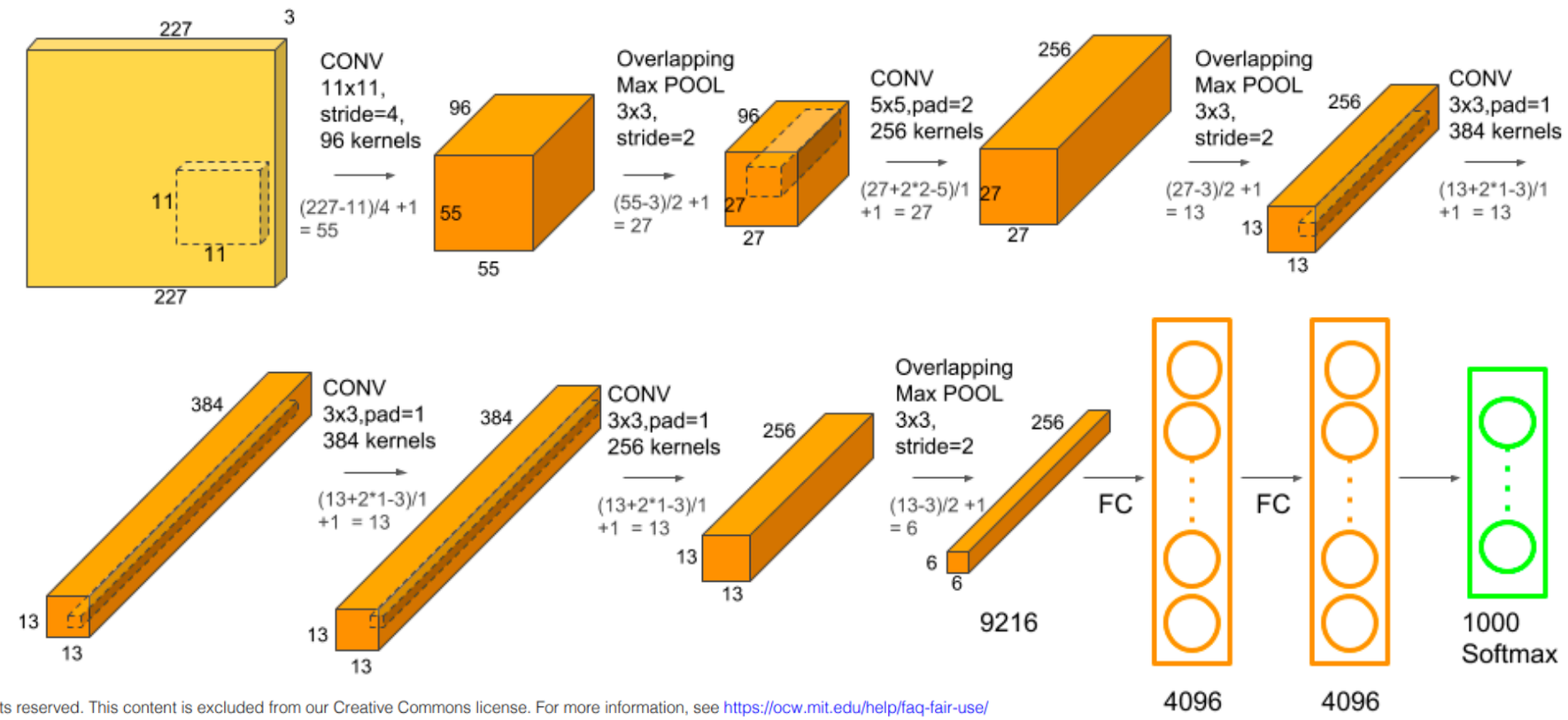
>1k classes

Deep Learning Revolution

AlexNet:
- winning entry in ILSVRC 2012
- CNN
- 10% error reduction



RGB image as input:



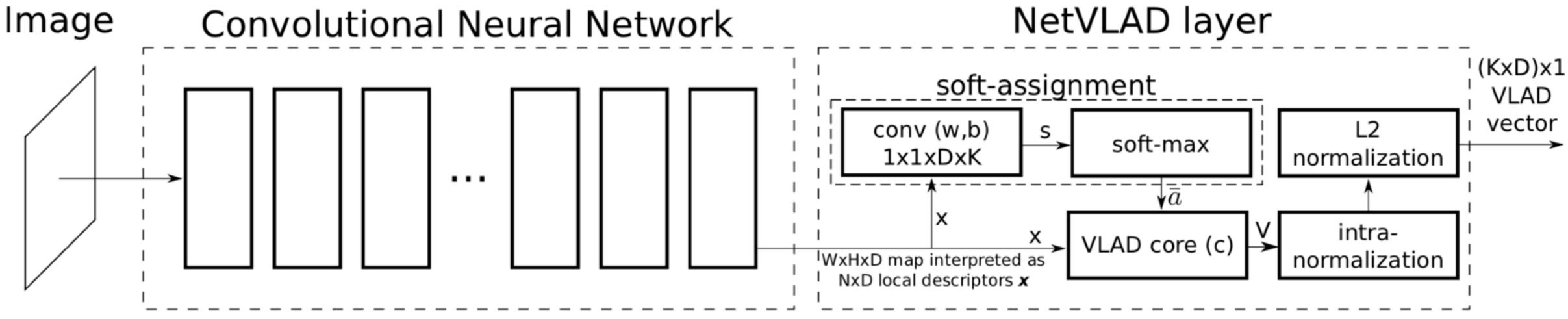
Classification results

Learning-based Descriptors: NetVLAD

Earlier approaches: using AlexNet or similar and use layers activations as descriptors

NetVLAD:

- CNN-based approach
- Trained on the task of place recognition
- Clever use of Internet data for training



R. Arandjelovix, P. Gronat, J. Sivic, NetVLAD: CNN Architecture for Weakly Supervised Place Recognition, CVPR'16.

Learning-based Descriptors: NetVLAD

How to get labeled data?

- a large dataset of panoramic images from the Google Street View Time Machine
- positions based on their (noisy) GPS
- Seasonal variations
- Illumination changes

(a) (b) (c)

Figure 4. **Google Street View Time Machine examples.** Each column shows perspective images generated from panoramas from nearby locations, taken at different times. A well designed method can use this source of imagery to learn to be invariant to changes in viewpoint and lighting (a-c), and to moderate occlusions (b). It can also learn to suppress confusing visual information such as clouds (a), vehicles and people (b-c), and to chose to either ignore vegetation or to learn a season-invariant vegetation representation (a-c). More examples are given in appendix B.

Metrics

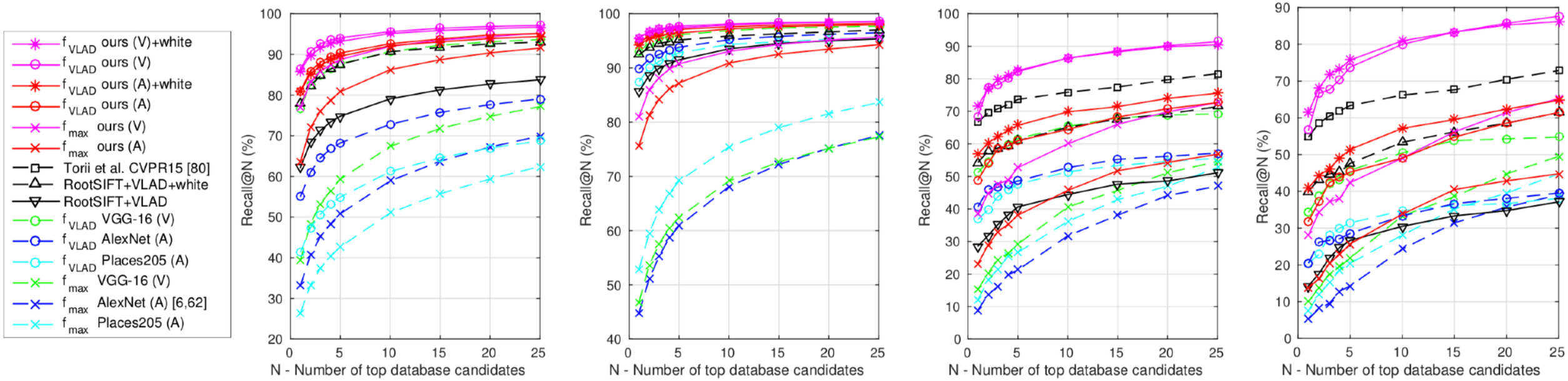
True positives (TP): correct matches
False positives (FP): incorrect matches
False negatives (FN): missed matches

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}.$$

Perfect system:
100% precision (0 FP)
100% recall (0 FN)

Learning-based Descriptors: NetVLAD



(a) Pitts250k-test (b) TokyoTM-val (c) Tokyo 24/7 all queries (d) Tokyo 24/7 sunset/night

Figure 5. **Comparison of our methods versus off-the-shelf networks and state-of-the-art.** The base CNN architecture is denoted in brackets: (A)lexNet and (V)GG-16. Trained representations (red and magenta for AlexNet and VGG-16) outperform by a large margin off-the-shelf ones (blue, cyan, green for AlexNet, Places205, VGG-16), f_{VLAD} (-o-) works better than f_{max} (-x-), and our f_{VLAD}+whitening (-∗-) representation based on VGG-16 sets the state-of-the-art on all datasets. [80] only evaluated on Tokyo 24/7 as the method relies on depth data not available in other datasets. Additional results are shown in appendix C.

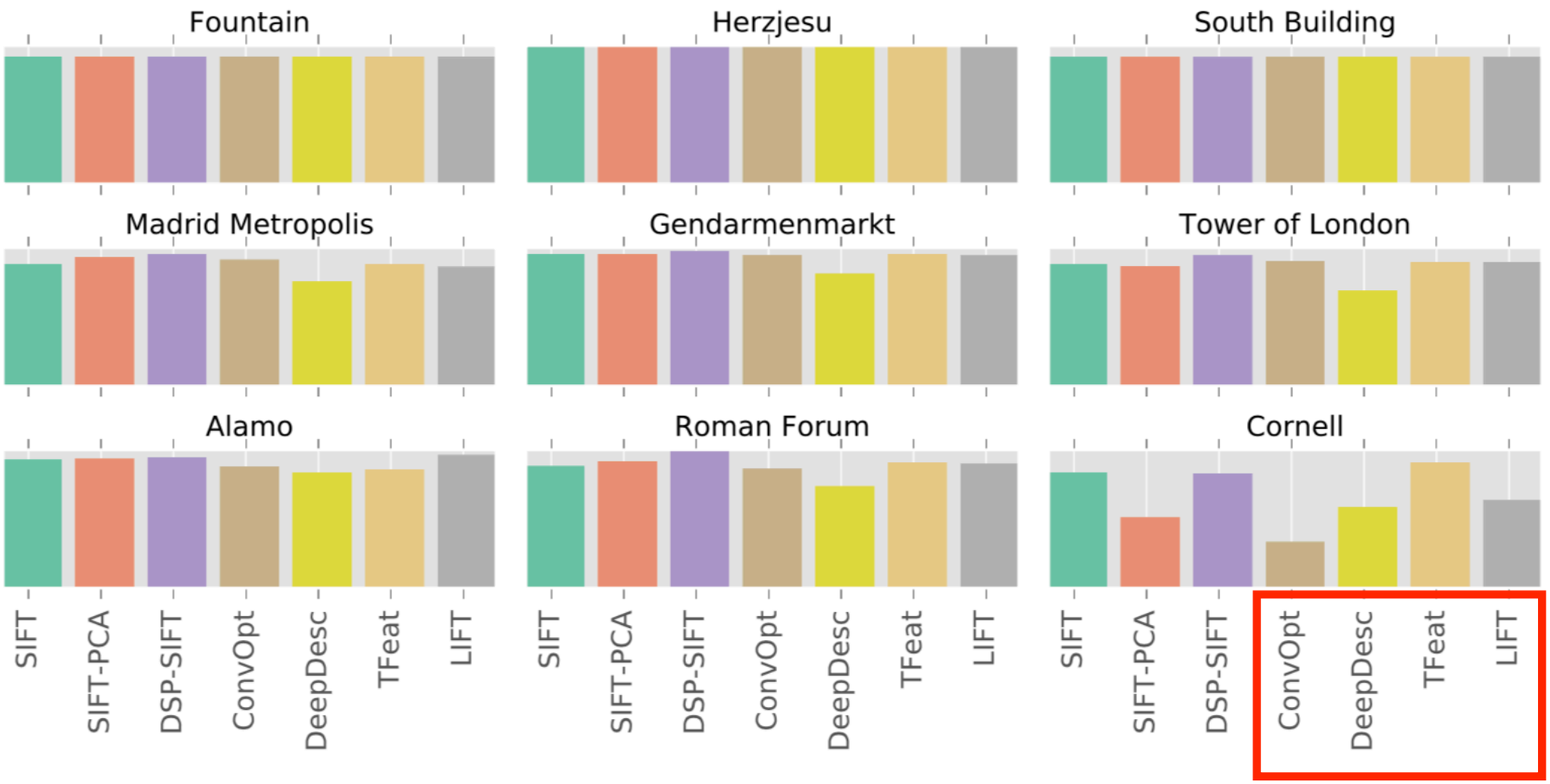
- query image is deemed correctly localized if at least one of the top N retrieved database images is within d = 25 meters from the ground truth position of the query.

- percentage of correctly recognized queries (Recall) is then plotted for different values of N

R. Arandjelovix, P. Gronat, J. Sivic, NetVLAD: CNN Architecture for Weakly Supervised Place Recognition, CVPR'16.

Handcrafted vs. Learned Local Descriptors

- learned descriptors typically outperform SIFT in terms of recall, while SIFT performs better in terms of precision
- advanced SIFT variants outperform learned features
- learned descriptors have high variance across the different datasets (i.e., over-fitting)



Number of registered images for the different methods

Learning-based

Schonberg et al., Comparative Evaluation of Hand-Crafted and Learned Local Features, CVPR'17.

Today + Next Lecture

- **Place recognition**
- **Object detection / recognition**

Visual Place Recognition: A Survey

Stephanie Lowry, Niko Sünderhauf, Paul Newman, *Fellow, IEEE*, John J. Leonard, *Fellow, IEEE*, David Cox, Peter Corke, *Fellow, IEEE*, and Michael J. Milford, *Member, IEEE*

***Abstract*—Visual place recognition is a challenging problem due to the vast range of ways in which the appearance of real-world places can vary. In recent years, improvements in visual sensing capabilities, an ever-increasing focus on long-term mobile robot autonomy, and the ability to draw on state-of-the-art research in other disciplines—particularly recognition in computer vision and animal navigation in neuroscience—have all contributed to significant advances in visual place recognition systems. This paper presents a survey of the visual place recognition research landscape. We start by introducing the concepts behind place recognition—the role of place recognition in the animal kingdom, how a "place" is defined in a robotics context, and the major components of a place recognition system. Long-term robot operations have revealed that changing appearance can be a significant factor in visual place recognition failure; therefore, we discuss how place recognition solutions can implicitly or explicitly account for appearance change within the environment. Finally, we close with a discussion on the future of**

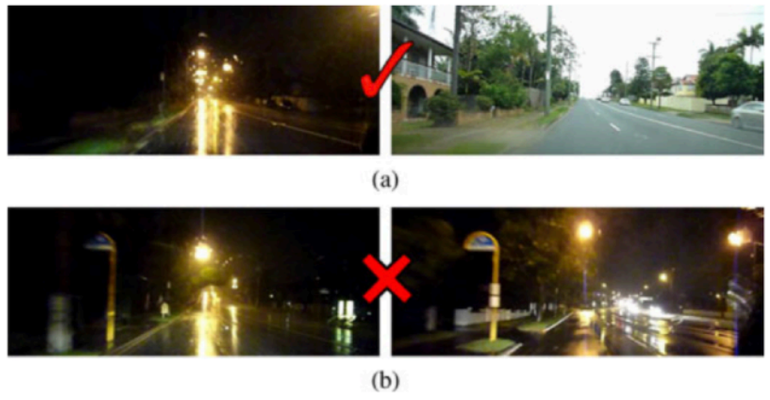


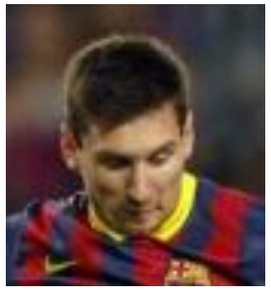
Fig. 1. Visual place recognition systems must be able to (a) successfully match very perceptually different images while (b) also rejecting incorrect matches between aliased image pairs of different places.

+ a few more recent papers

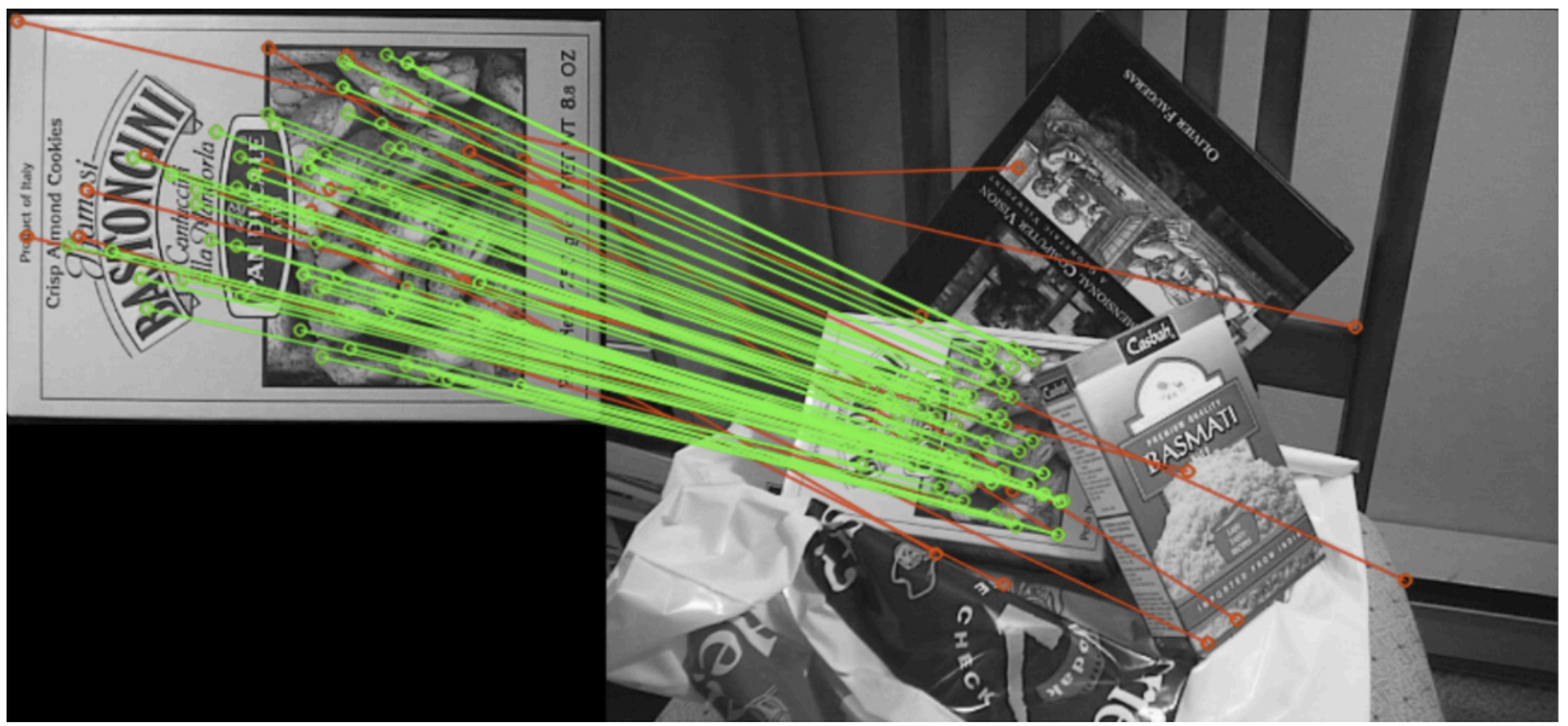
Traditional Object Detectors

- template matching (sliding window)

template



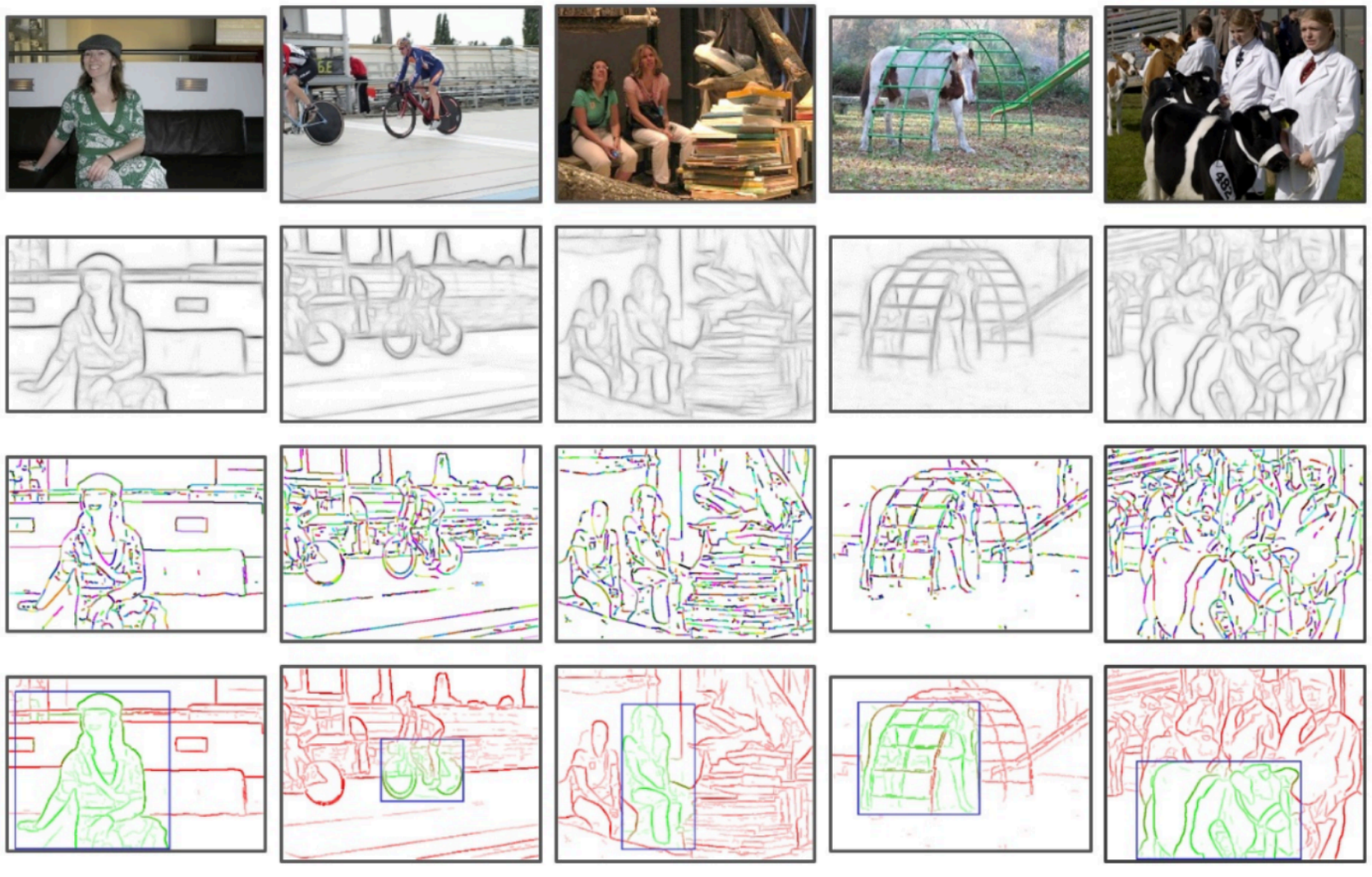
- feature-based

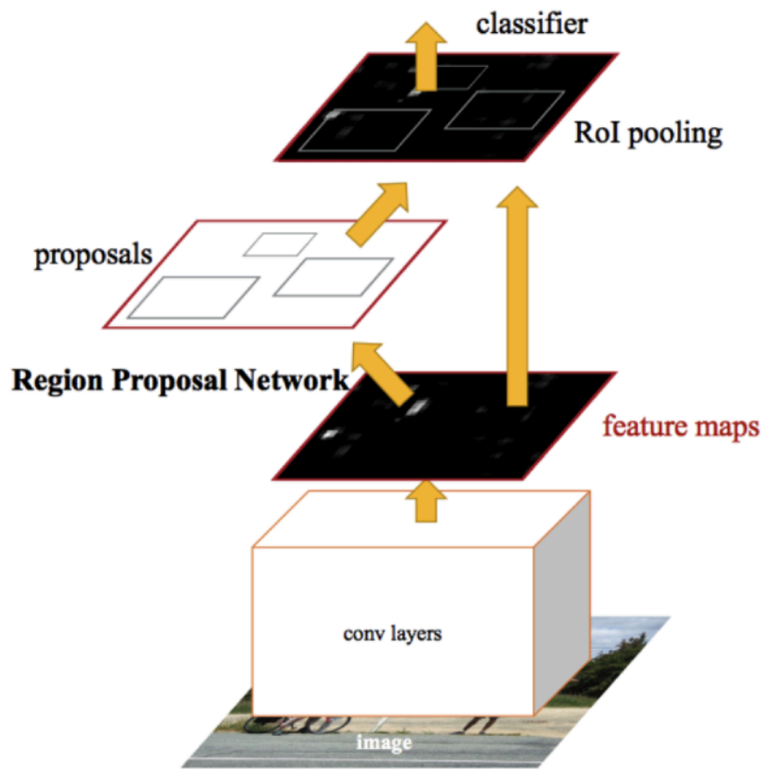


(scalability?)

Traditional Object Detectors

- Object proposal + object classification





(robustness? speed?)

L. Zitnick and P. Dollar, Edge Boxes: Locating Object Proposals from Edges, ECCV'14

Learning-based Object Detection: YOLO

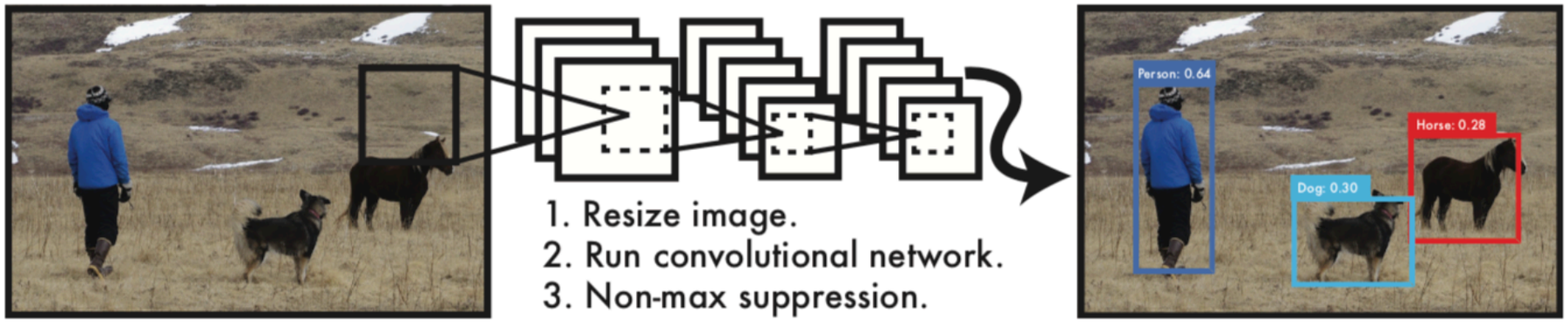


Figure 1: The YOLO Detection System. Processing images with YOLO is simple and straightforward. Our system (1) resizes the input image to 448×448, (2) runs a single convolutional network on the image, and (3) thresholds the resulting detections by the model's confidence.

- YOLO processes images 45 frames per second.
- A smaller version of the network, Fast YOLO, processes an 155fps

Redmond et al, "You Only Look Once: Unified, Real-Time Object Detection", CVPR'16.

Learning-based Object Detection: YOLO

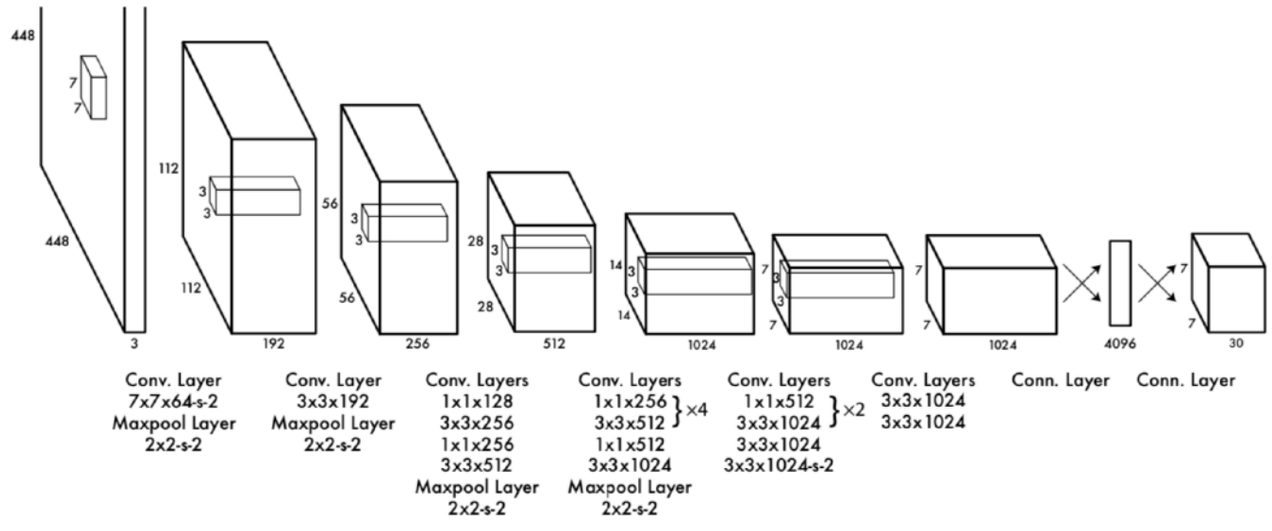


Figure 3: The Architecture. Our detection network has 24 convolutional layers followed by 2 fully connected layers. Alternating 1×1 convolutional layers reduce the features space from preceding layers. We pretrain the convolutional layers on the ImageNet classification task at half the resolution (224×224 input image) and then double the resolution for detection.

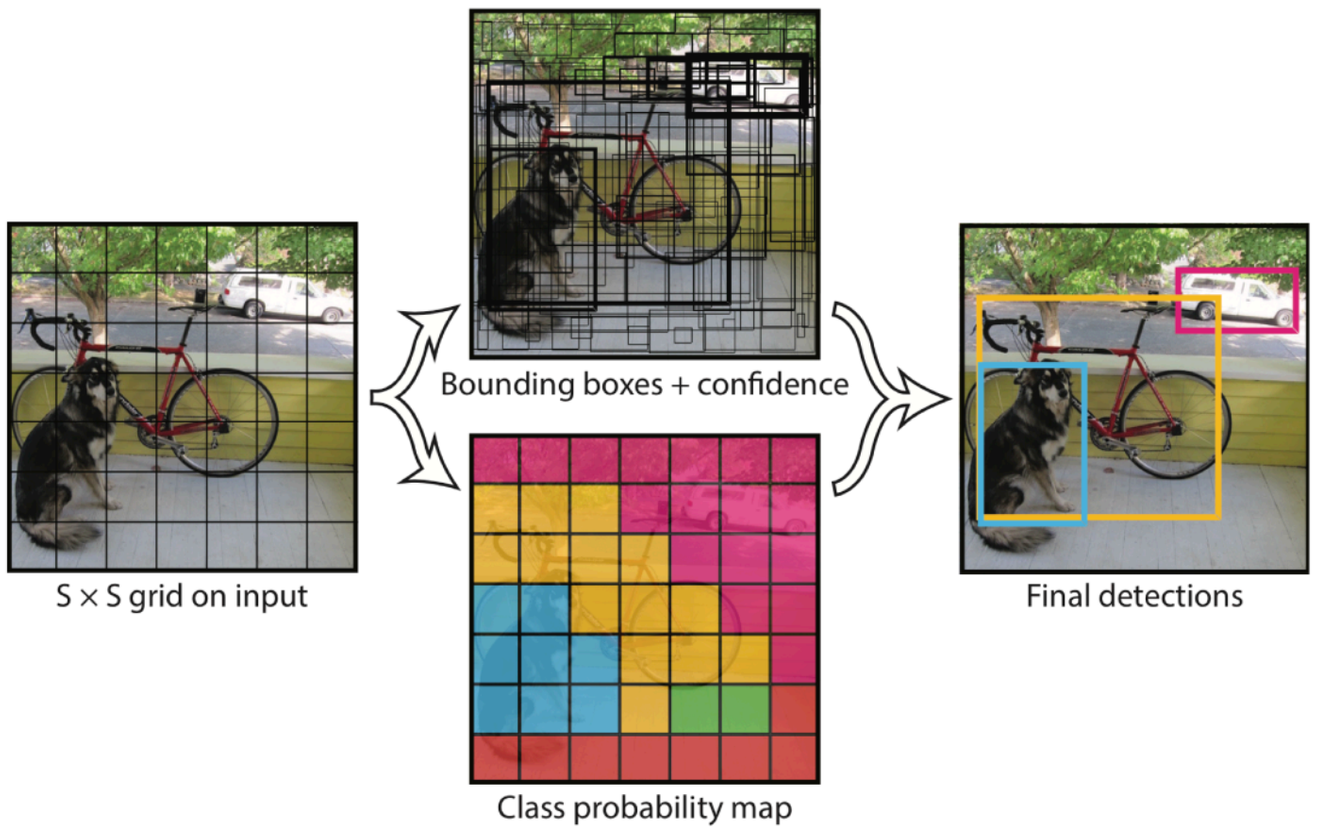


Image is split in S x S grid.

Yolo is trained to predict:

- B bounding boxes in each grid cell (x, y, h, w, confidence)
- A class label for each cell

YOLO

Redmond et al, “You Only Look Once: Unified, Real-Time Object Detection”, CVPR’16.
https://www.youtube.com/watch?v=uG2UOasIx2I

Learning-based Object Detection: YOLO

mAP: mean Average Precision (average precision value for recall value over 0 to 1).

Real-Time Detectors	Train	mAP	FPS
100Hz DPM [31]	2007	16.0	100
30Hz DPM [31]	2007	26.1	30
Fast YOLO	2007+2012	52.7	**155**
YOLO	2007+2012	**63.4**	45
Less Than Real-Time			
Fastest DPM [38]	2007	30.4	15
R-CNN Minus R [20]	2007	53.5	6
Fast R-CNN [14]	2007+2012	70.0	0.5
Faster R-CNN VGG-16[28]	2007+2012	73.2	7
Faster R-CNN ZF [28]	2007+2012	62.1	18
YOLO VGG-16	2007+2012	66.4	21

Table 1: Real-Time Systems on PASCAL VOC 2007. Comparing the performance and speed of fast detectors. Fast YOLO is the fastest detector on record for PASCAL VOC detection and is still twice as accurate as any other real-time detector. YOLO is 10 mAP more accurate than the fast version while still well above real-time in speed.

Limitations of YOLO:

- **small objects**: “each grid cell only predicts B boxes and can only have one class. This spatial constraint limits the number of nearby objects that our model can predict. Our model struggles with small objects that appear in groups, such as flocks of birds."

- **generalization**: fails to detect objects in new or unusual aspect ratios or configurations.