COMPUTER VISION & MACHINE LEARNING

Lecture-3 , Day- 2

STTP on “Deep Learning, Computer Vision and Speech Processing”

By: Suprava, Patnaik, Professor, ExTC, XIE, Mumbai

What is Computer Vision?

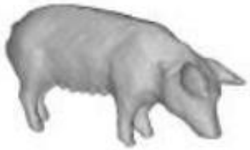
- Want to make computers understand what we are viewing. (image processing-image analysis-feature extraction-pattern recognition, help in various applications)

- **Goal:** Input at high-dimensional visual data, and fit models to summaries the data, based on the fact that computer will understand the input, like human beings (may be better). (from pixels-to-scene)
  - Content based image retrieval
  - Recognizing and learning object categories

- **Model building**
  - Discriminative Model for identifying an object
  - Generative Model for describing an object

Research in CV

Image-based 3D Reconstruction



Shape Analysis

Robot Vision

RGB-D Vision

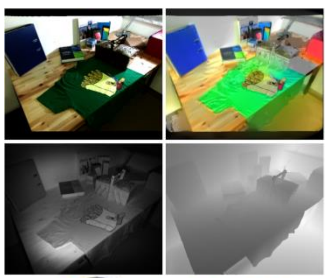
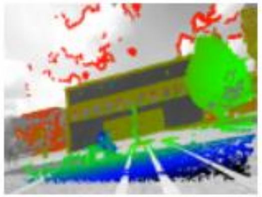


Image Segmentation

Visual SLAM



Optical Flow

Convex Relaxation Methods

Computer Vision Pipeline

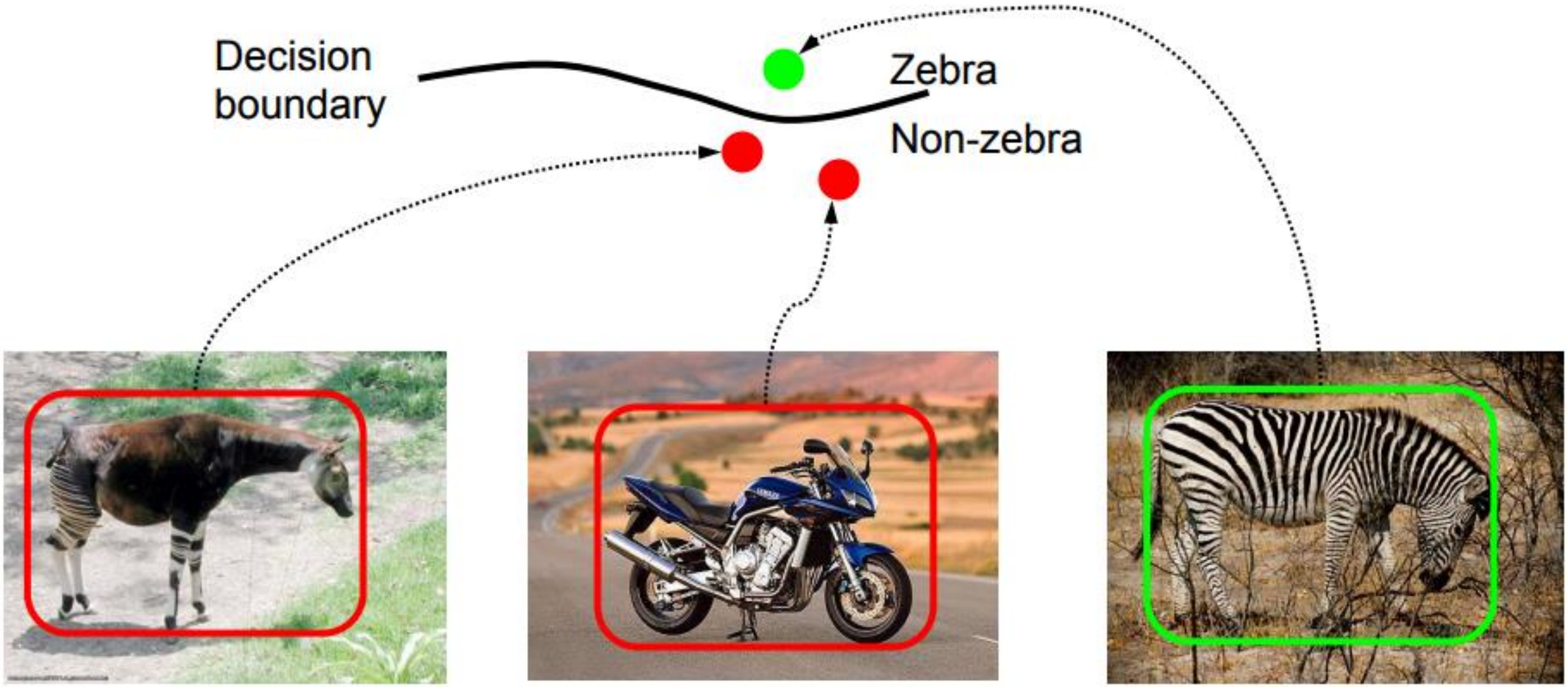
- Representation:
  - How to define object category
- Learning:
  - Defining a model to respond to the category specific inputs only
- Recognition:
  - How to use models

Representation

A picture is worth a thousand words

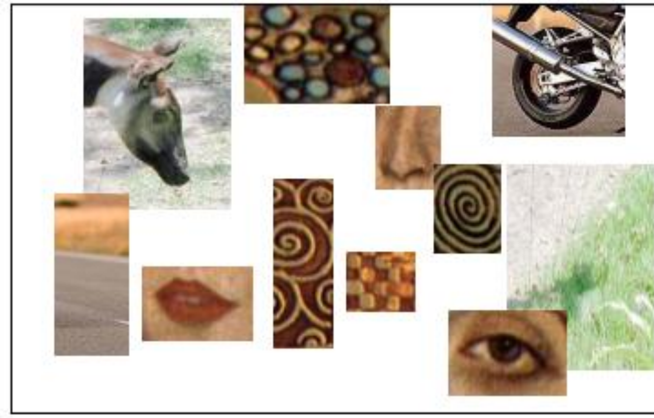
Discriminative

- Direct modeling of $\dfrac{p(zebra \mid image)}{p(no\ zebra \mid image)}$



Generative Model (Bag of words)

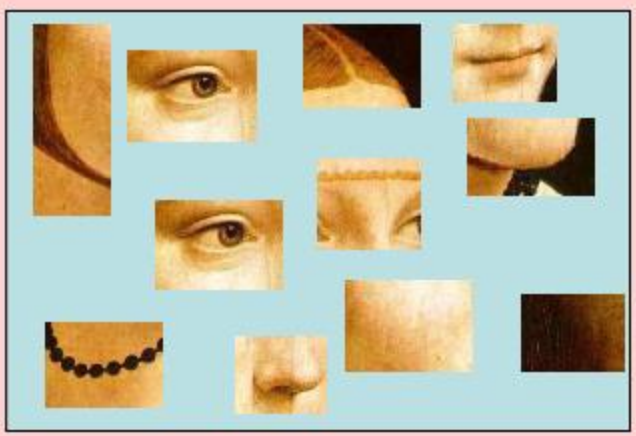
- Model $p(image \mid zebra)$ and $p(image \mid no\ zebra)$

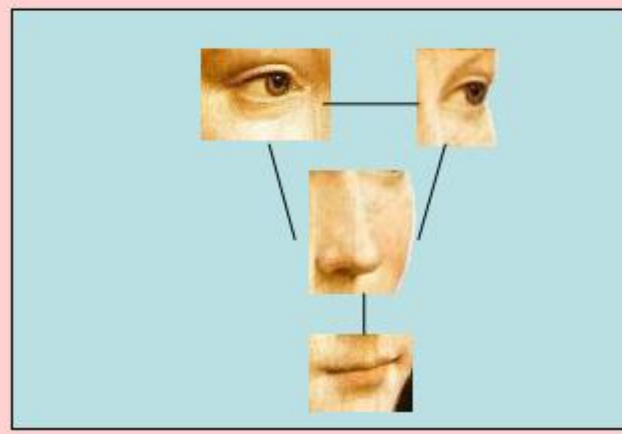


	$p(image \mid zebra)$	$p(image \mid no\ zebra)$
	Low	Middle
	High	Middle→Low

Hybrid Model

- Face Recognition
  - Appearance only or location and appearance





Some Literature

Discriminative Approaches:

Perceptron and Neural networks (Rosenblatt 1958, Windrow and Hoff 1960, Hopfiled 1982, Rumelhart and McClelland 1986, Lecun et al. 1998)

Nearest neighborhood classifier (Hart 1968)

Fisher linear discriminant analysis(Fisher)

Support Vector Machine (Vapnik 1995)

Bagging, Boosting,… (Breiman 1994, Freund and Schapire 1995, Friedman et al. 1998,)

…

Generative Approaches:

PCA, TCA, ICA (Karhunen and Loeve 1947, H´ erault et al. 1980, Frey and Jojic 1999)

MRFs, Particle Filtering (Ising, Geman and Geman 1994, Isard and Blake 1996)

Maximum Entropy Model (Della Pietra et al. 1997, Zhu et al. 1997, Hinton 2002)

Deep Nets (Hinton et al. 2006)

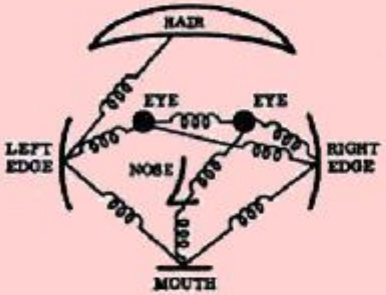
….

Representation

– Use set of features or each pixel in image

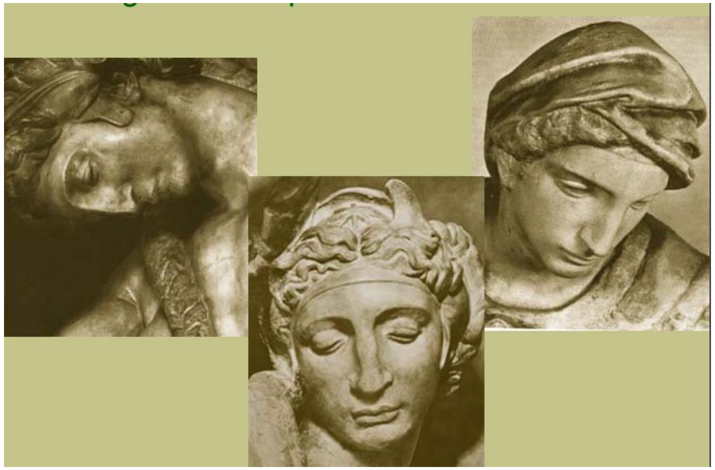
– Invariances
  - View point
  - Illumination
  - Occlusion
  - Scale
  - Deformation
  - Clutter
  - etc.

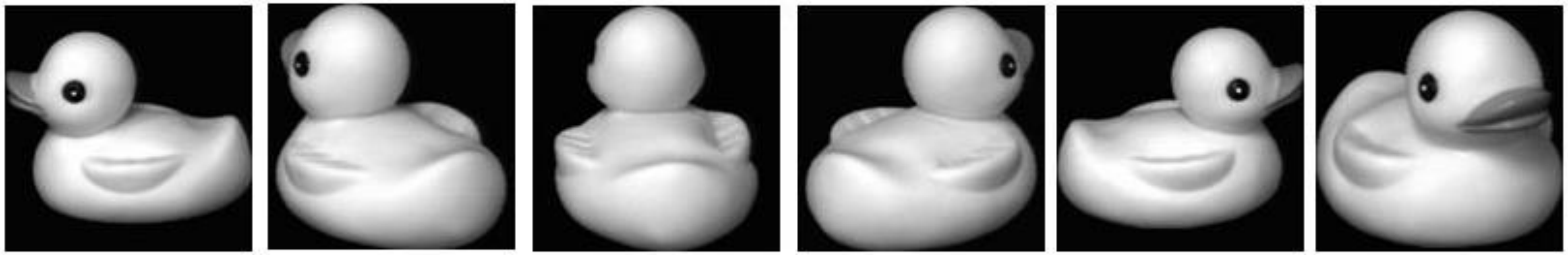
– Part-based or global w/sub-window



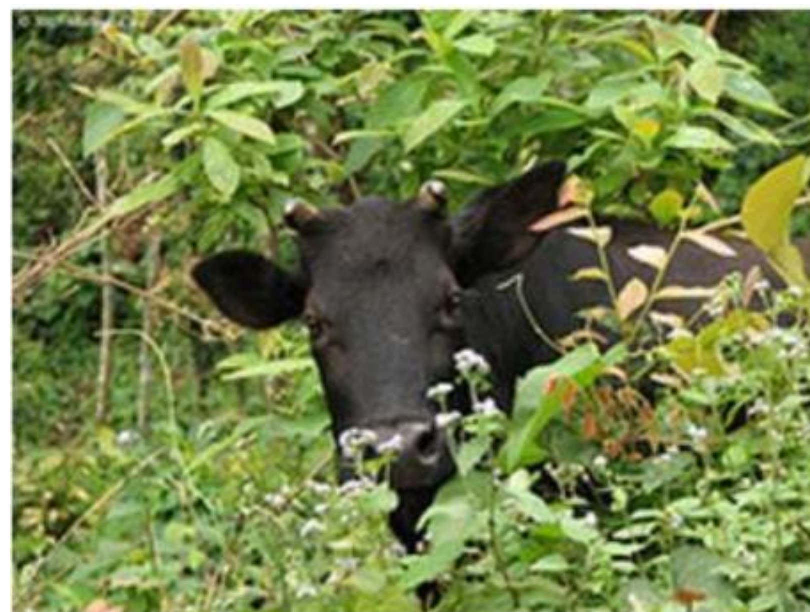
View point variation

- Appearance of 3D object can change dramatically with variations in view angle or object orientation

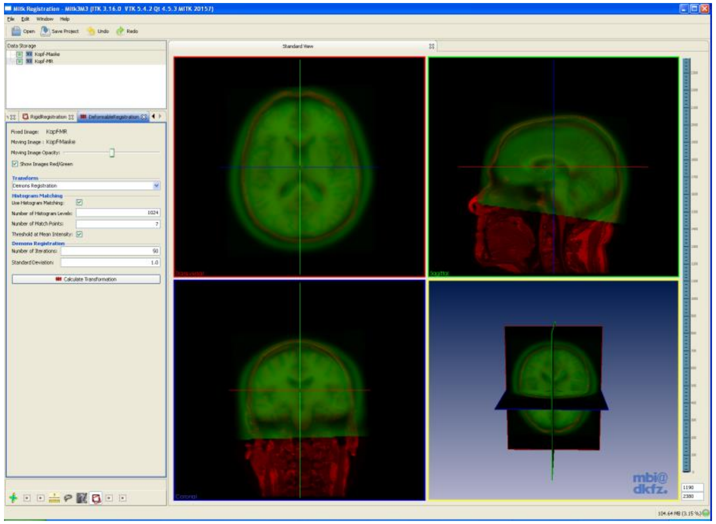




Illumination, Occlusion



Scale, Deformation



Motion (Source: S. Lazebnik)

Back-ground clutter and context

Intra class variation

Model Learning (feature based)

Edges and beyond edges?

- Mid-level cues

"More Tokens"

- High-level object parts:

- Difficult to hand-engineer → What about learning them?

Feature hierarchy?

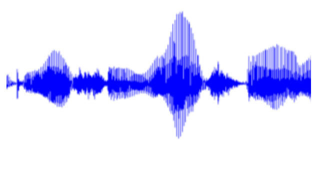
Object detection

Image

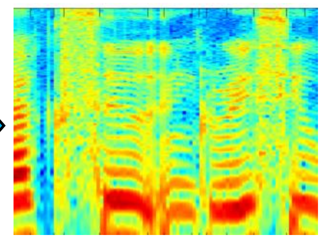
Low-level vision features

Recognition

Audio classification



Audio

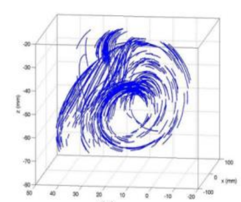


Low-level audio features

Speaker identification

Helicopter control

Helicopter



Low-level state features

Action

CV Before 2012

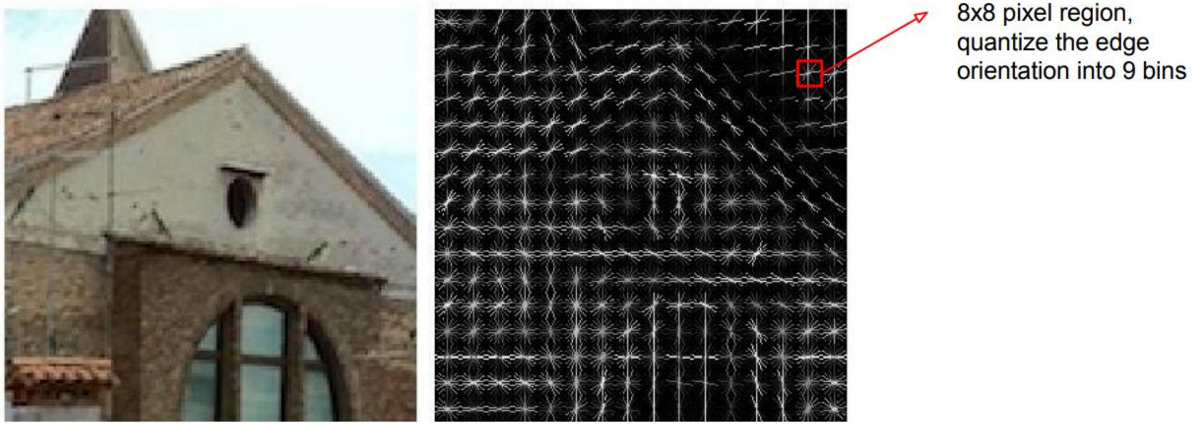
- Feature based recognition
- HOG, SIFT

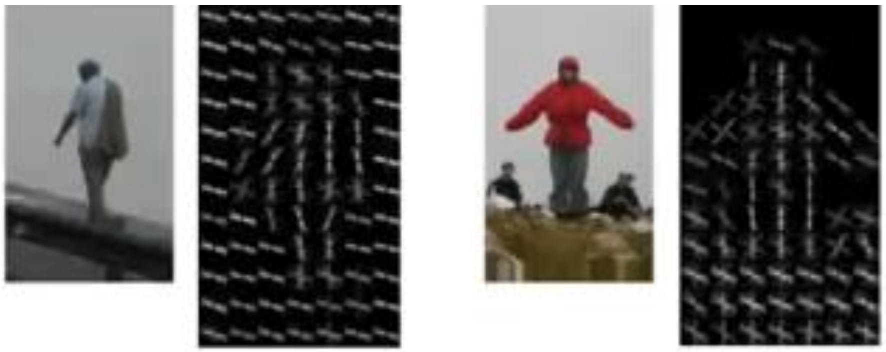
1. Divide into overlapping patches.
2. Extract HOG on the patches
3. Consider orientation bins
4. Concatenate HOGs from all the batches

repeat for each detected feature

Traditional object recognition

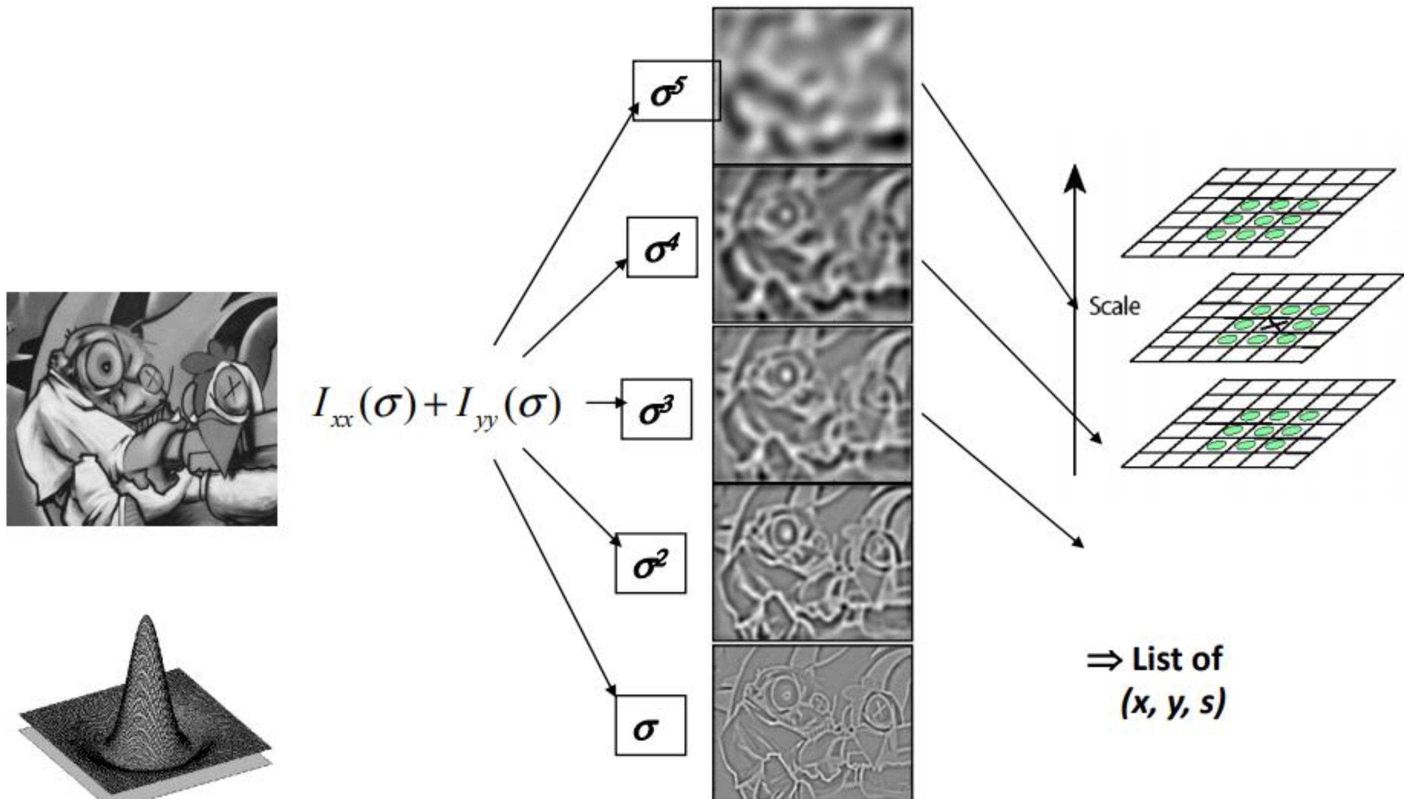
Example: HOG features





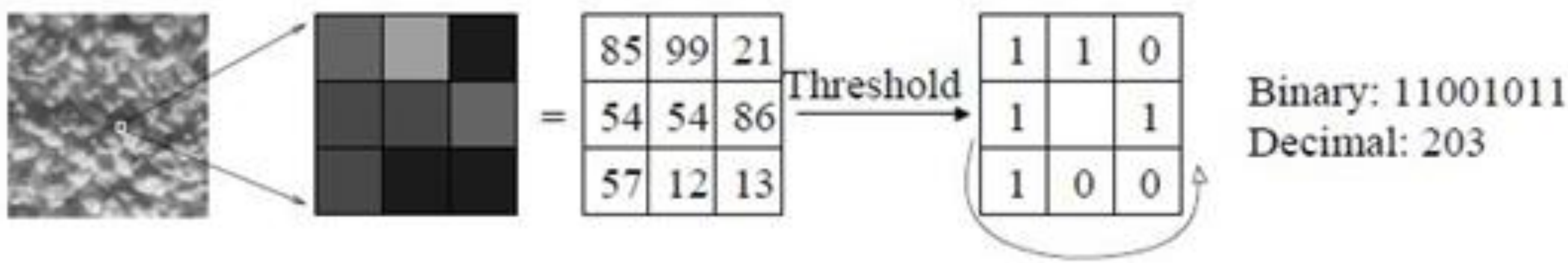
SIFT (Scale Invariant Feature Transform)

- Difference of Gaussian blurring of an image with two different variances , let it be φ and kφ
- Search for local extrema over scale and space.
- If the ir
  value, i
- Apply c
  points.
- A 16x1
  devided
  block, &
  128 bir
  to form

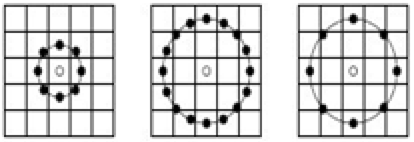


LBP (Local Binary Pattern)

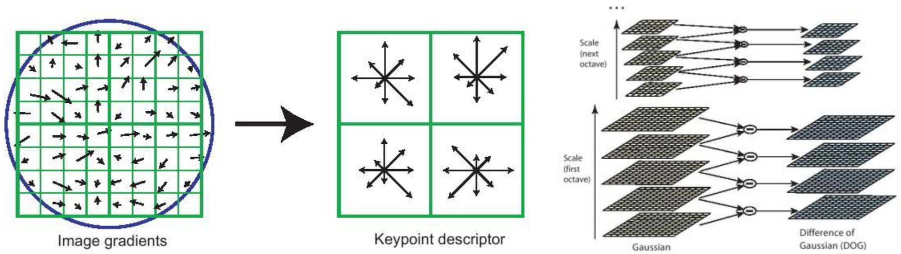
- The histogram of the labels used as a descriptor(texture).



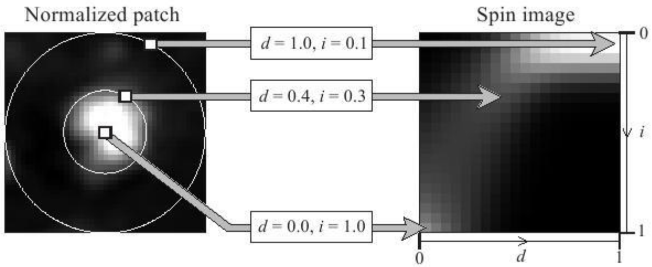
- Neighborhood of different sizes



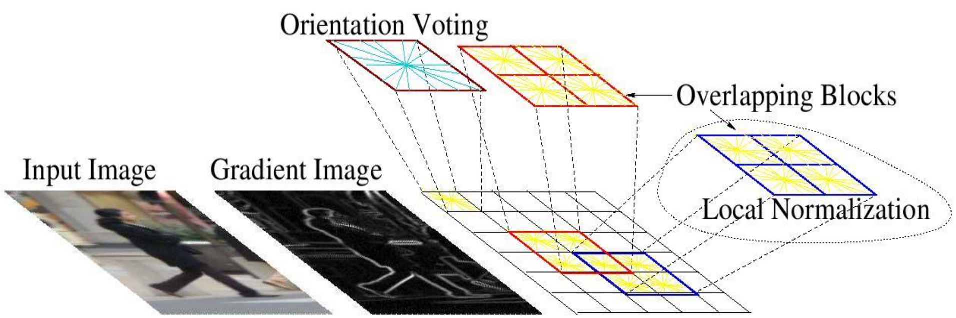
Common CV features



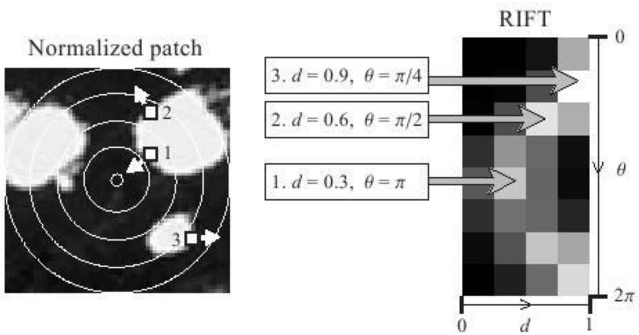
SIFT



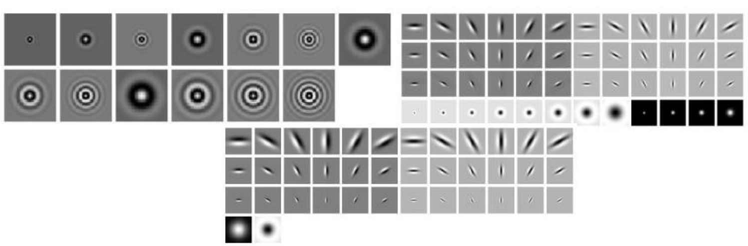
Spin image



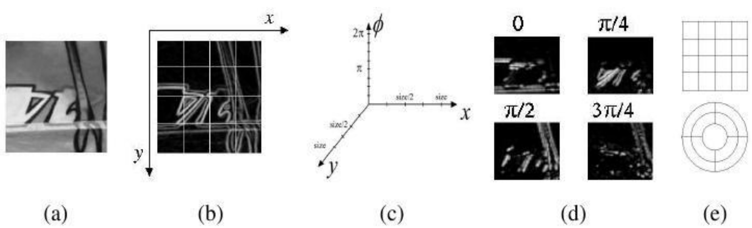
HoG



RIFT



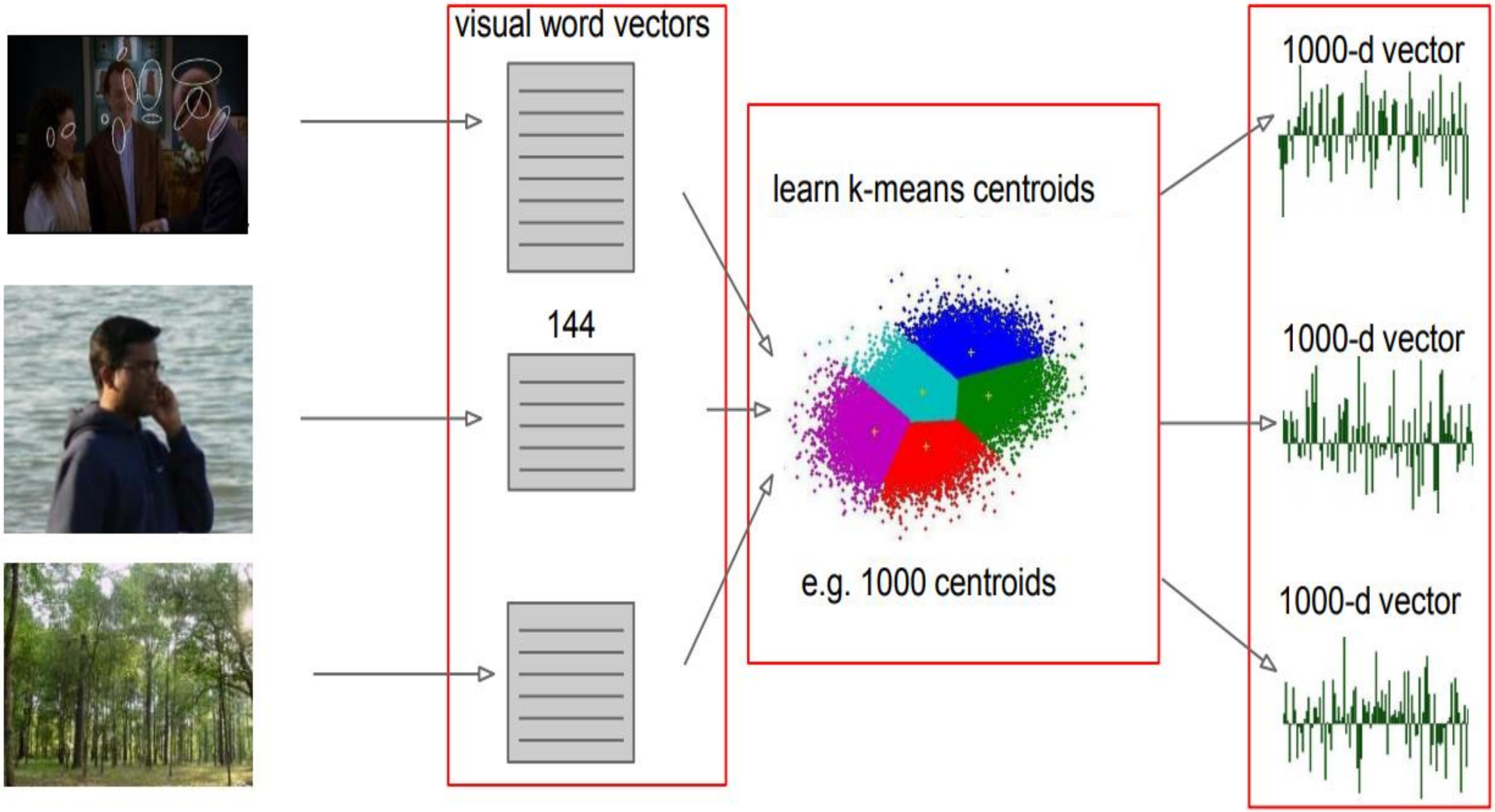
Textons



GLOH

Features as Bag of words

Example: Bag of Words



Requirements of a local feature

- **Repetitive :** Detect the same points independently in each image.
- **Invariant** to translation, rotation, scale, i.e invariant to affine transformation.
- **Invariant** to presence of noise, blur etc.
- **Locality :**Robust to occlusion, clutter and illumination change.
- **Distinctiveness :** The region should contain “interesting” structure.
- **Quantity :** There should be enough points to represent the image. •
- **Time efficient**.

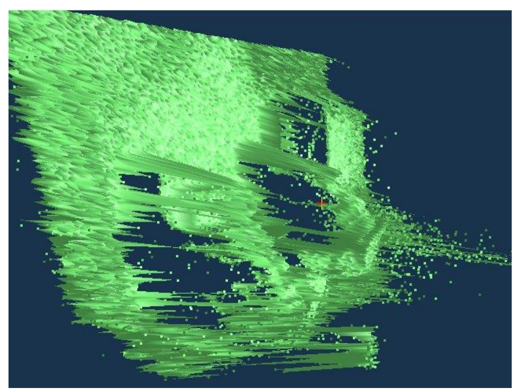
Computer vision is more than pictures (after 2012)

Images

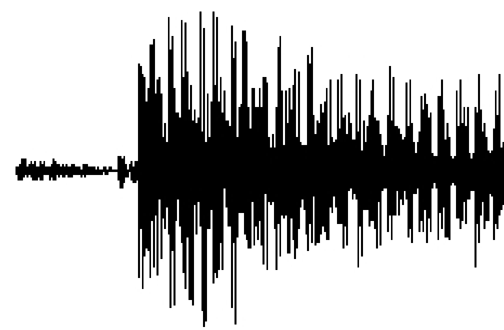
Video

Camera array

Thermal
Infrared



3d range scans
(flash lidar)



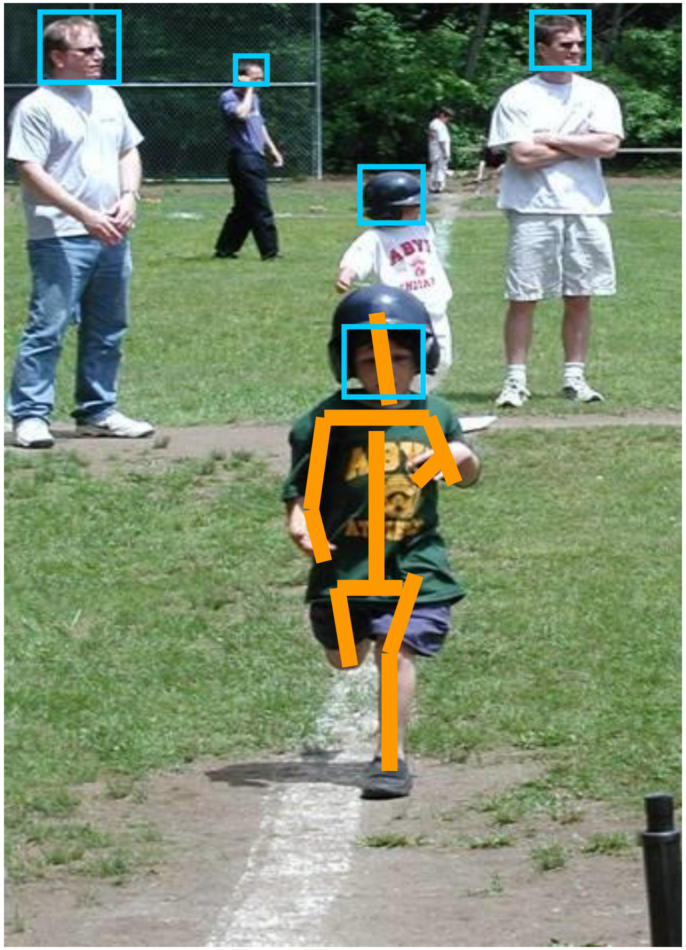
Audio

Discriminative v.s. Generative Models

Generative and discriminative learning are key problems in machine learning and computer vision.

If you are asking, "Are there any faces in this image?", then you would probably want to use discriminative methods.

If you are asking, "Find a 3-d model that describes the runner", then you would use generative methods.



ICCV W. Freeman and A. Blake

Intuition about Margin: data augmantation

Infant

?

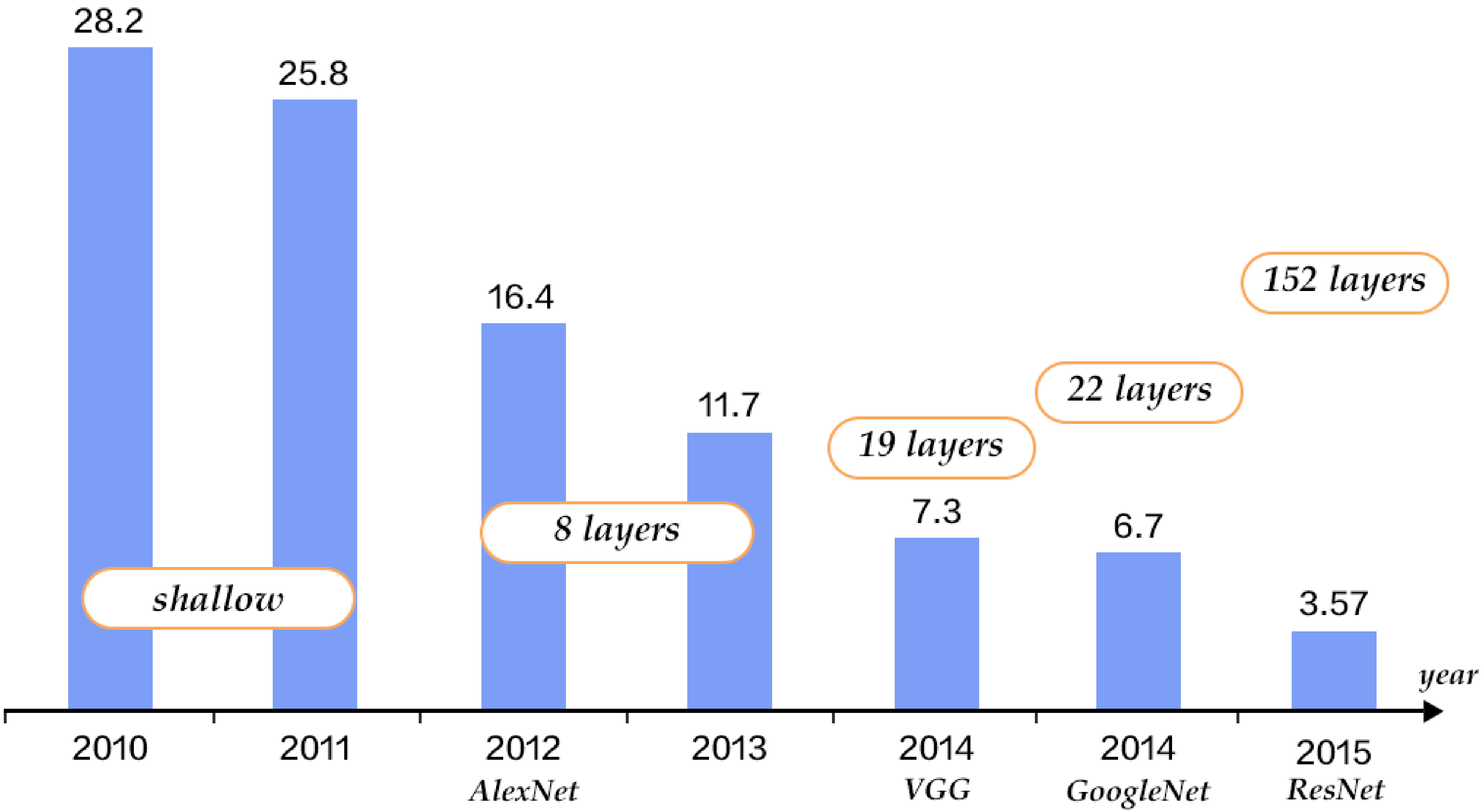
Elderly

Man

?

Woman

Era after AlexNet

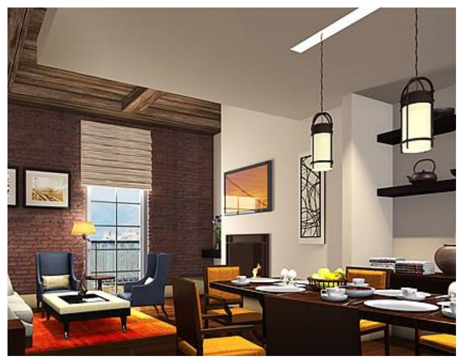


Recent Developments

1. Sparse coding for feature learning: Dictionary based learning
2. Compressed Sensing : Next session
3. Advanced classification

No more words ~ knowledge abstraction

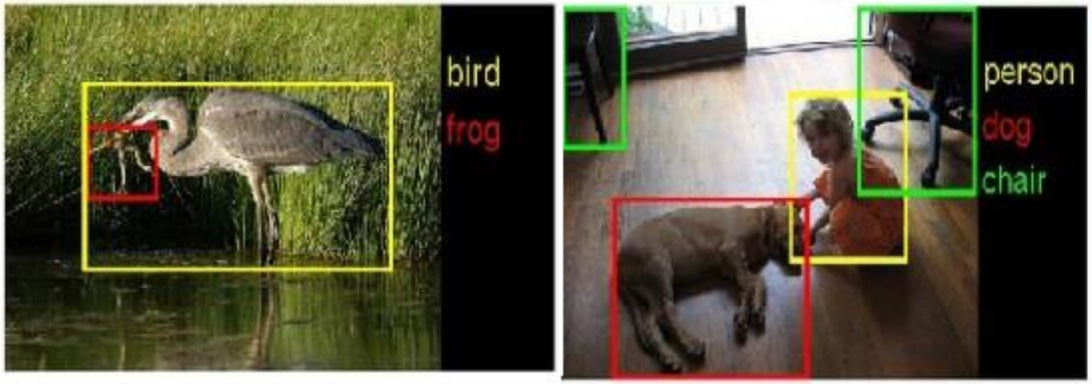
- Key question: Can we automatically learn a good feature representation?
- Find a better way to represent images than pixels.



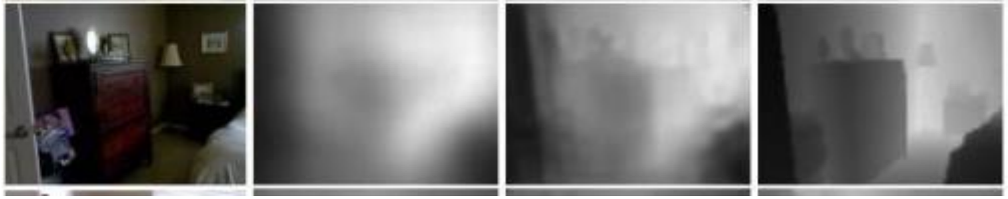
Going beyond Classification:

- Localization
- Detection
- Depth Estimation (from single image)



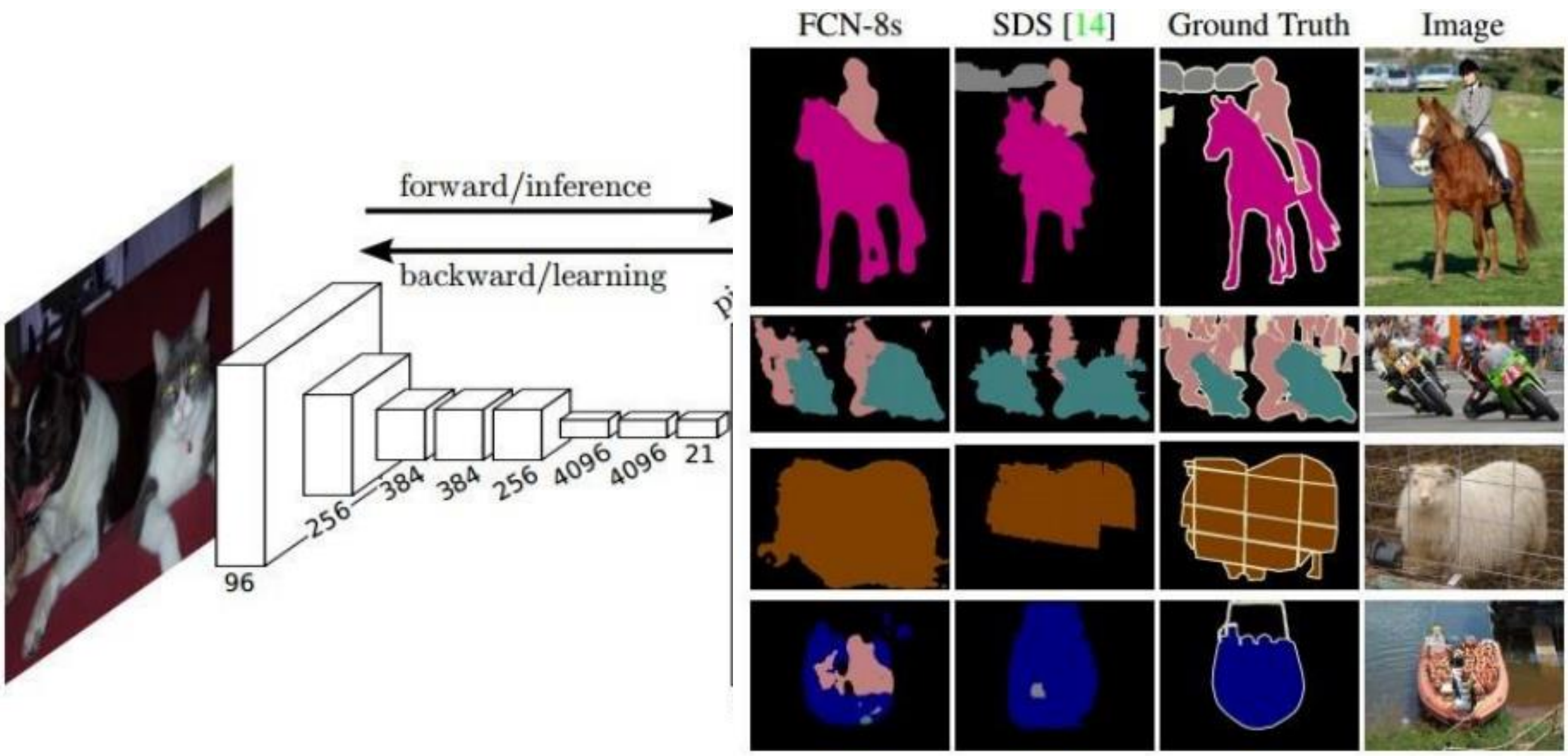
Each detection has:

- confidence
- class (integer)
- x1,y1,x2,y2 bounding box coordinates



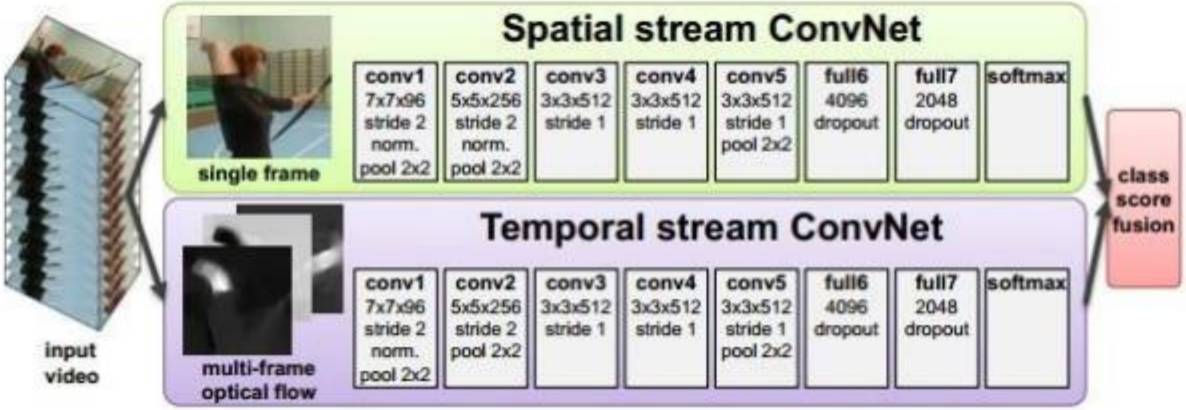
Going beyond Classification:

- Semantic Segmentation



Going beyond Classification:

- **Video Classification** (action detection from Multitask learning)
- The spatial stream performs action recognition from still video frames, whilst the temporal stream is trained to recognise action from motion in the form of dense optical flow



Two-Stream Convolutional Networks for Action Recognition in Videos *[Simonyan et al.], 2014*

Going beyond Classification:

- Image Captioning

"man in black shirt is playing guitar."

"construction worker in orange safety vest is working on road."

"two young girls are playing with lego toy."

"girl in pink dress is jumping in air."

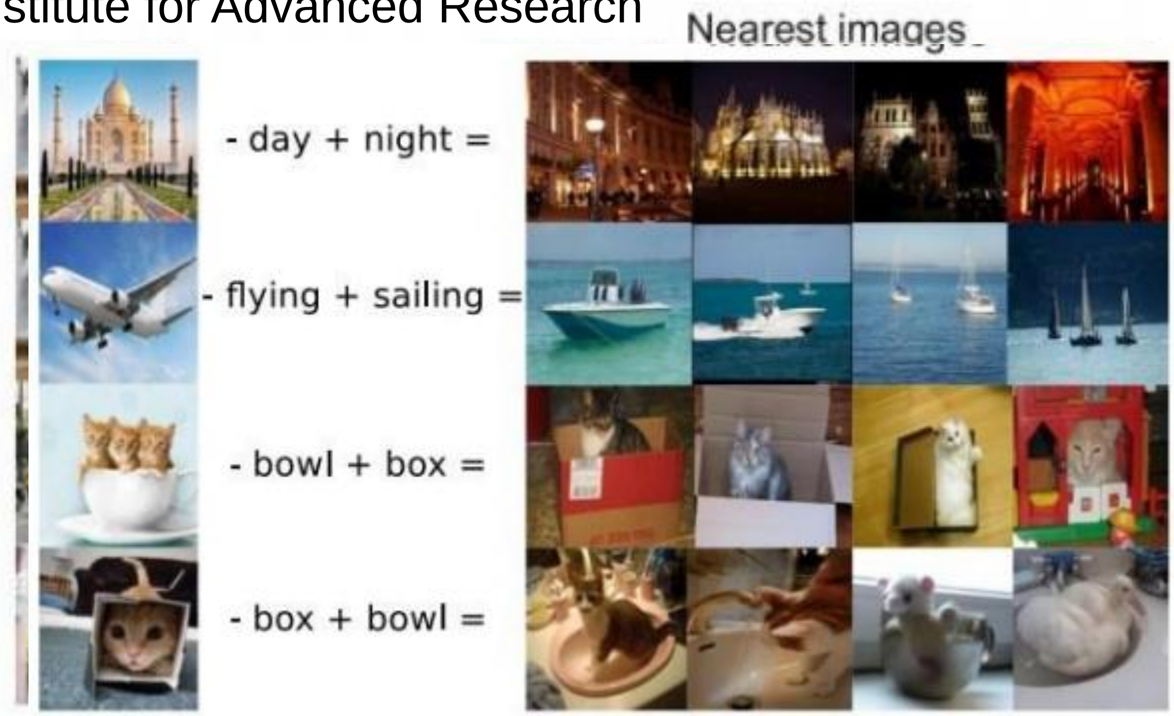
"black and white dog jumps over bar."

"young girl in pink shirt is swinging on swing."

Going beyond Classification:

- Image Ranking and retrieval
- Unifying Visual-Semantic Embedding with Multimodal Neural Language Models

Ryan Kiros, Ruslan Salakhutdinov, Richard S. Zemel University of Toronto
Canadian Institute for Advanced Research



Going beyond Classification:

- Visual Question Answering

Rotation Invariant Featur × Microsoft PowerPoint - F × CNN × https://arxiv.org/pdf/141 × CloudCV: Visual Questic

Secure | https://vqa.cloudcv.org

CloudCV EvalAI Origami Fabrik GSoC Demos

CloudCV: Visual Question Answering (V

CloudCV can answer questions you ask about an image

More details about the VQA dataset can be found here. Torch code for VQA is available here.

Browsers currently supported by the demo: Google Chrome, Mozilla Firefox.

Try VQA on Sample Images

Click on one of these images to send it to our servers (Or upload your own images below)



- Still have many orders of magnitude to go in order to match the infero-temporal(IT) pathway of the human visual system.

Questions?? Comments?? Thoughts?? Ideas??