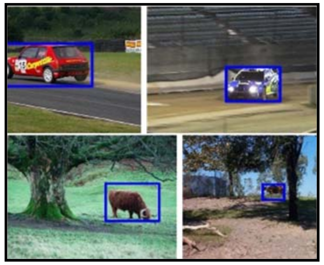
Machine visual perception

Cordelia Schmid
INRIA Grenoble

Machine visual perception

- Artificial capacity to ***see***, understand the visual world

Image or sequence of images

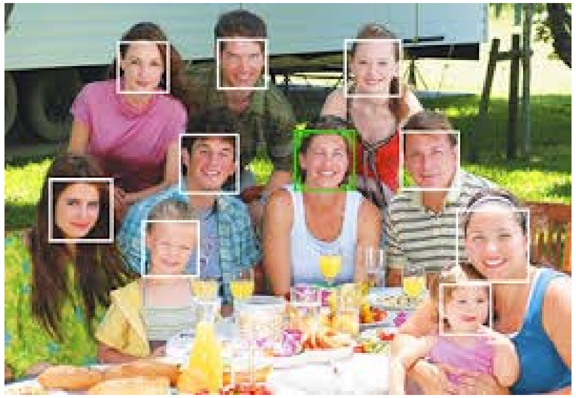


Object recognition

Action recognition

Machine visual perception - applications

- Face detection
  - Available in many cameras for autofocus
  - First step for face recognition



Courtesy Fujifilm

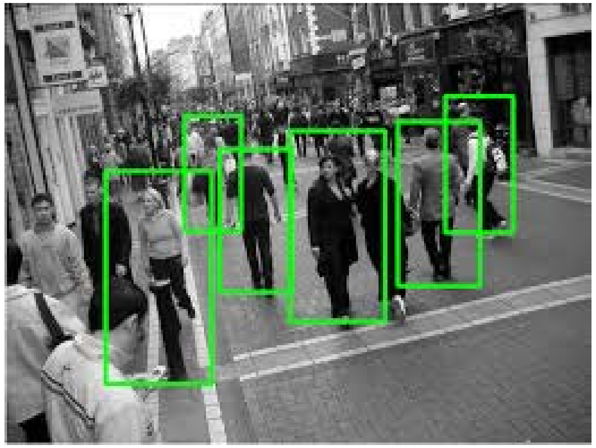
Face Detection function keeps subjects' faces in **sharp focus**

Courtesy Ricoh

Machine visual perception - applications

- Pedestrian detection
  - Applicable to car safety and video surveillance

Courtesy Volvo



Courtesy Embedded Vision Alliance

Machine visual perception - applications

- Search for places and particular objects
  - For example on a smart phone

Courtesy Google

Machine visual perception - applications

- Complete description (story) of a video

As the headwaiter takes them to a table they pass by the piano, and the woman looks at Sam. Sam, with a conscious effort, keeps his eyes on the keyboard as they go past. The headwaiter seats Ilsa...

Machine visual perception - applications

- Complete description (story) of a video

As the headwaiter takes them to a table they pass by the piano, and the woman looks at Sam. Sam, with a conscious effort, keeps his eyes on the keyboard as they go past. The headwaiter seats Ilsa...

Machine visual perception - applications

- Complete description (story) of a video

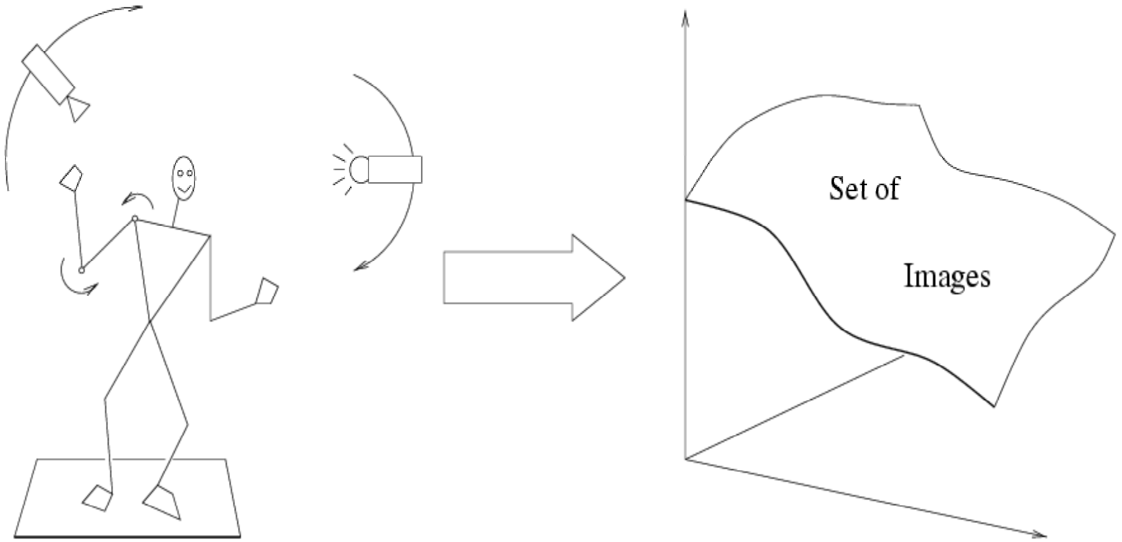
As the headwaiter takes them to a table they pass by the piano, and the woman looks at Sam. Sam, with a conscious effort, keeps his eyes on the keyboard as they go past. The headwaiter seats Ilsa...

Machine visual perception - applications

- Complete description (story) of a video

As the headwaiter takes them to a table they pass by the piano, and the woman looks at Sam. Sam, with a conscious effort, keeps his eyes on the keyboard as they go past. The headwaiter seats Ilsa...

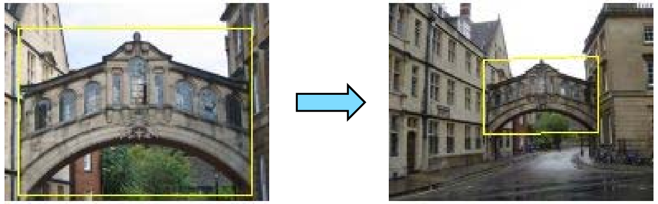
Difficulties: within-object variations



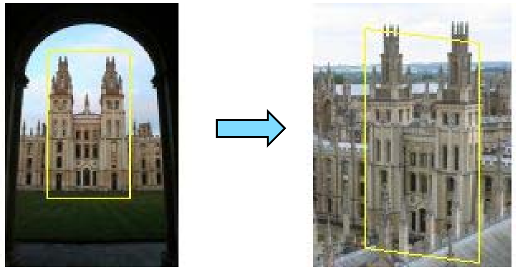
Variability: Camera position, illumination, internal parameters

Within-object variations

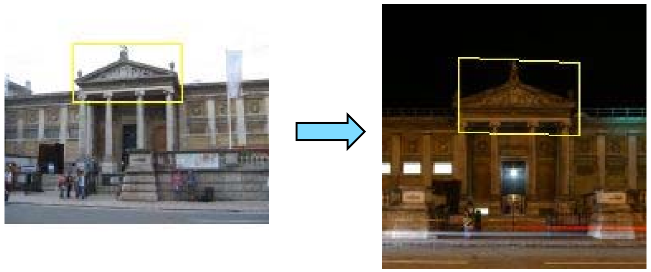
Difficulties: within-object variations



Scale



Viewpoint

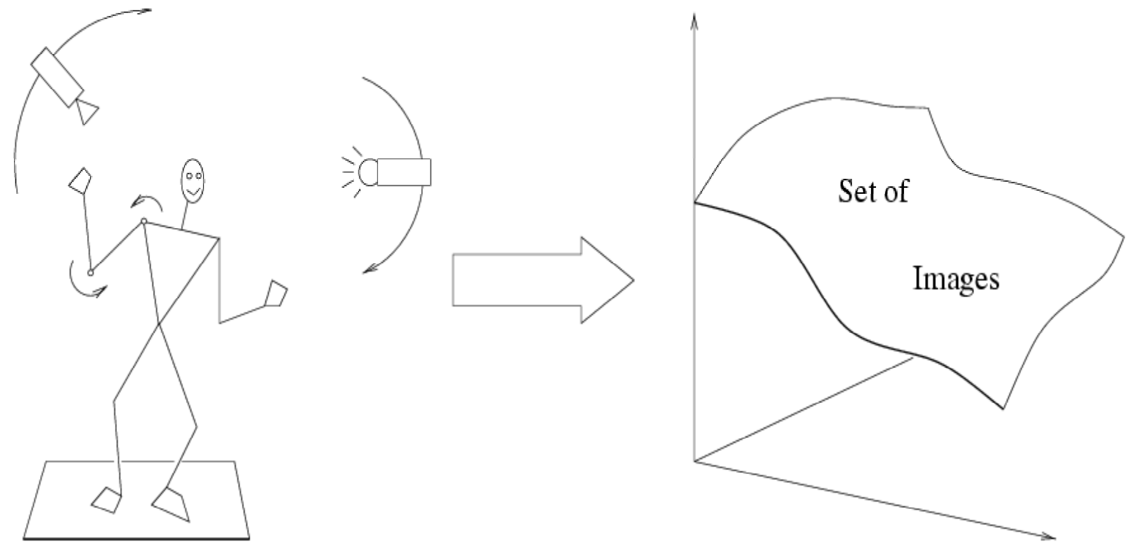


Lighting

Occlusion

informatics mathematics

Difficulties: within-class variations



Variability: many different objects belong to a class

Within-class variations

Difficulties: within-class variations





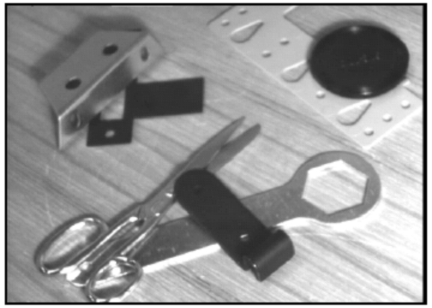

Difficulties: within-class variations

Overview

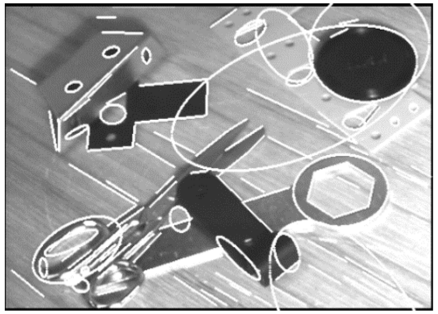
- ***History of machine visual perception***
- State of the art for visual perception
- Practical matters

Machine vision late 80s to early 90s

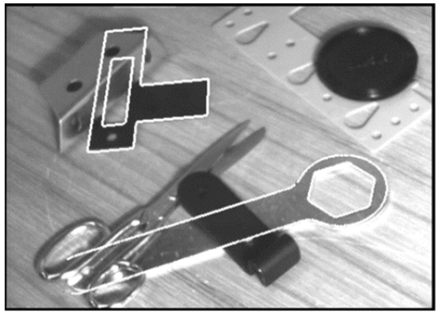
- Simple features, handcrafted models, few images, simple tasks



original image



detected features

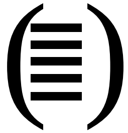


objects recognized
with projective invariants

Rothwell, Zisserman, Mundy and Forsyth, *Efficient Model Library Access by Projectively Invariant Indexing Functions,* CVPR 1992

Machine vision early 90s to early 2000s

- Local appearance-based descriptors (> 1000 images/objects)



local descriptor

$$\begin{bmatrix} L \\ L_i L_i \\ L_i L_{ij} L_j \\ L_{ii} \\ L_{ij} L_{ji} \\ \varepsilon_{ij}(L_{jkl} L_i L_k L_l - L_{jkk} L_i L_l L_l) \\ L_{iij} L_j L_k L_k - L_{ijk} L_i L_j L_k \\ -\varepsilon_{ij} L_{jkl} L_i L_k L_l \\ L_{ijk} L_i L_j L_k \end{bmatrix}$$

differential invariants, local jet

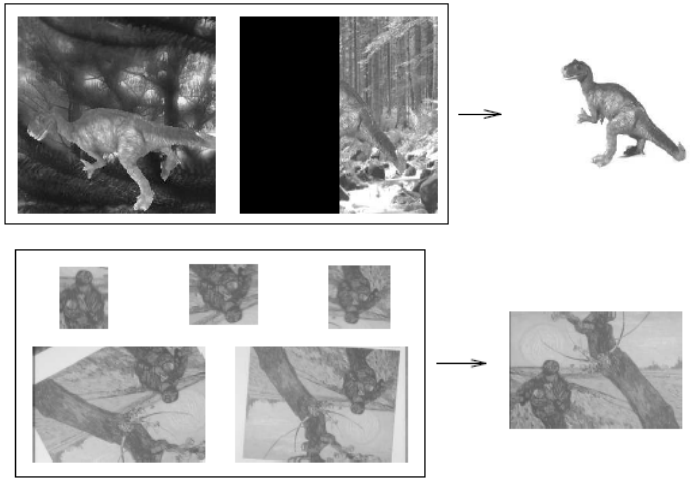
- Voting scheme to find most similar scene/object

Schmid and Mohr, *Local grayvalue invariants for image*, IEEE Trans. on Pattern Analysis & Machine Intelligence, 1997; **Longuet-Higgins Prize 2006**

Experimental results

- Local appearance-based descriptors (> 1000 images/objects)

Database

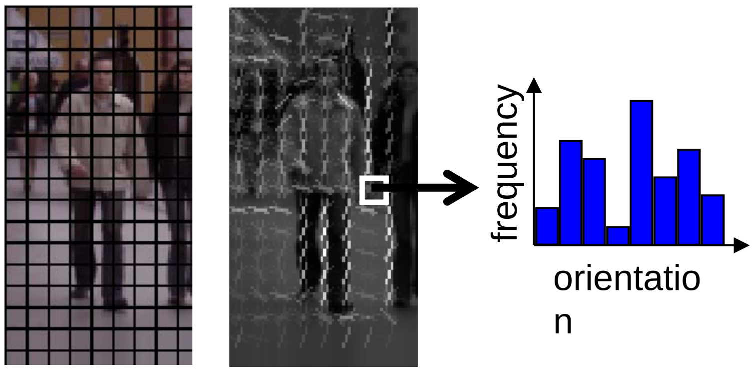


Search / recognition results

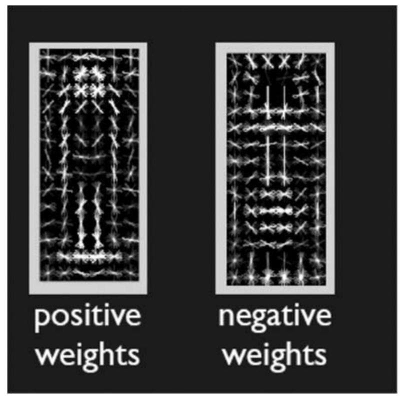
Schmid and Mohr, *Local grayvalue invariants for image*, IEEE Trans. on Pattern Analysis & Machine Intelligence, 1997; **Longuet-Higgins Prize 2006**

Machine vision early 2000s to early 2010s

- Machine learning based approach for categories (pedestrians)



Histogram of oriented gradients



Support vector machine classifier

Dalal and Triggs, *Histograms of oriented gradients for human detection*,
CVPR’05; **Longuet-Higgins Prize 2015**

Results for pedestrian localization

Dalal and Triggs, *Histograms of oriented gradients for human detection*, CVPR’05

Machine vision starting early 2010s

- End-to-end learning, deep convolutional neural networks
  [LeCun'98, …, Krizhevsky'12]

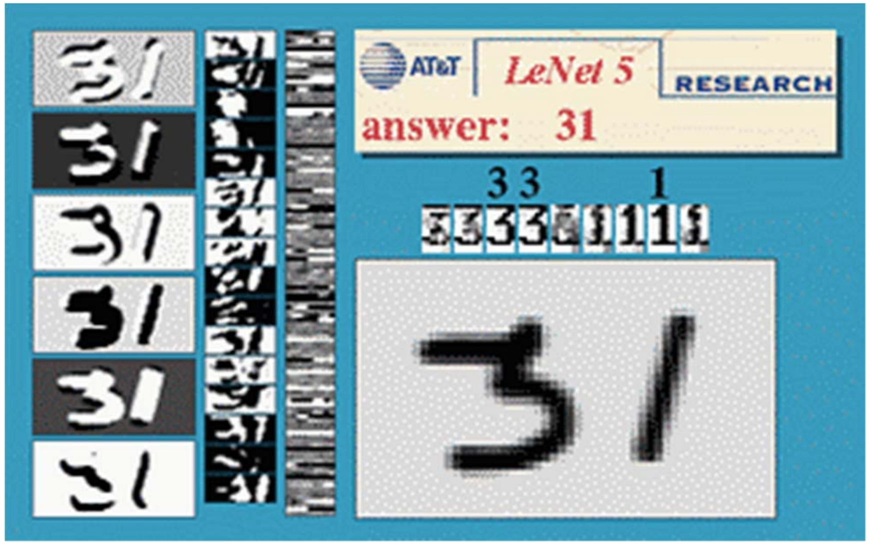
- State of the art result on ImageNet challenge
  - 1000 categories and 1.2 million images

(a) Star Anise (92.45%) (b) Geyser (85.45%) (c) Pulp Magazine (83.01%) (d) Carrycot (81.48%)

(e) European gallinule (15.00%) (f) Sea Snake (10.00 %) (g) Paintbrush (4.68 %) (h) Mountain Tent (0.00%)

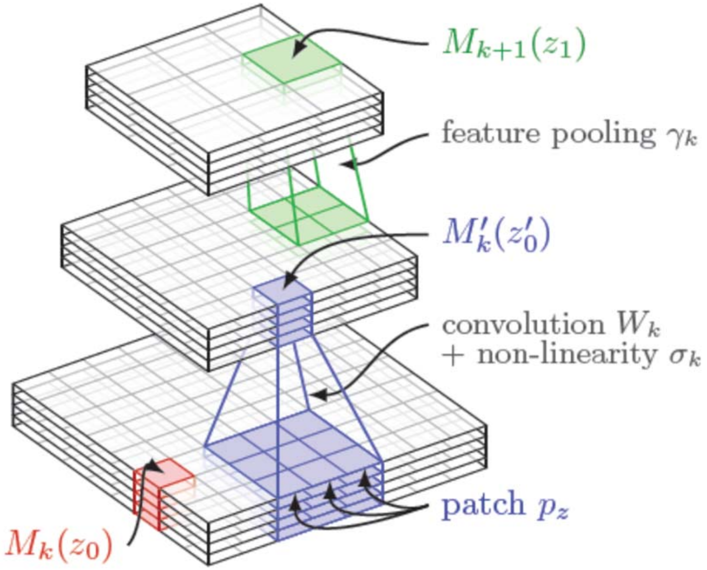
Machine vision starting early 2000s

- End-to-end learning, deep convolutional neural networks
  [LeCun'98, …, Krizhevsky'12]



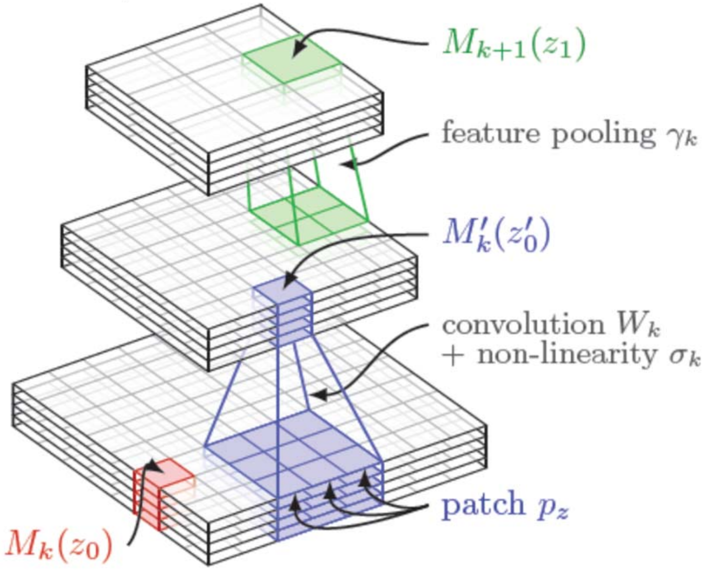
Deep convolutional neural networks

- Convolutional neural network – one layer



Deep convolutional neural networks

- Convolutional neural network – one layer



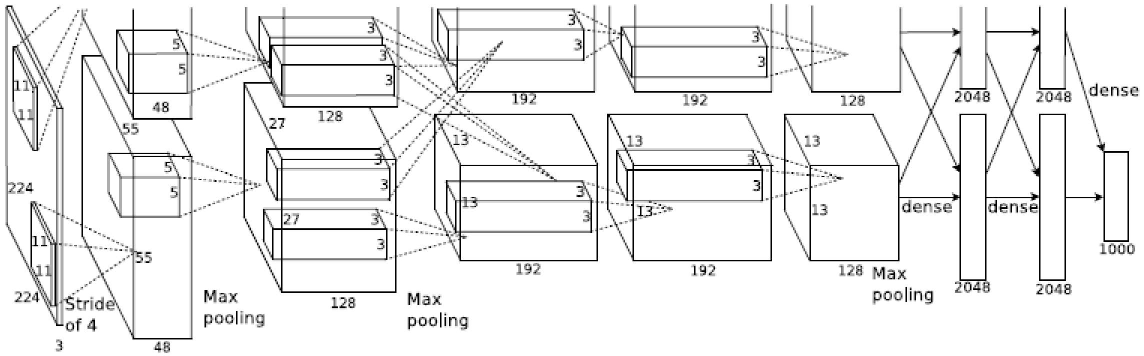
Convolutions
- Learned convolutional filters
- Translation invariant
- Several filters at each layer
- From simple to complex filters

Non-linearity (sigmoid, RELU)

Pooling (average, max)

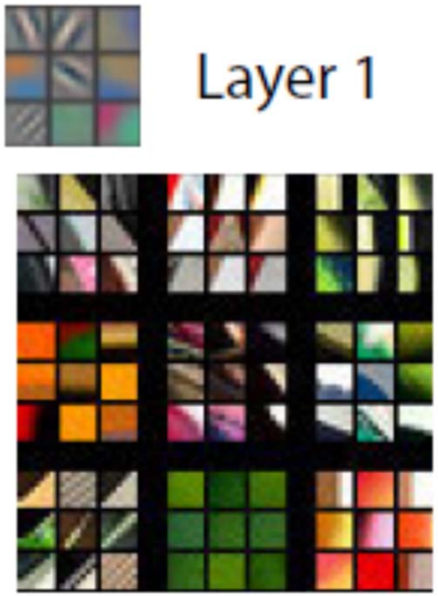
Deep convolutional neural networks

- First 5 layers: convolutional layer, last 2: full connected
- Large model (7 hidden layers, 650k units, 60M parameters)
- Requires large training set (ImageNet)
- GPU implementation (50x speed up over CPU)

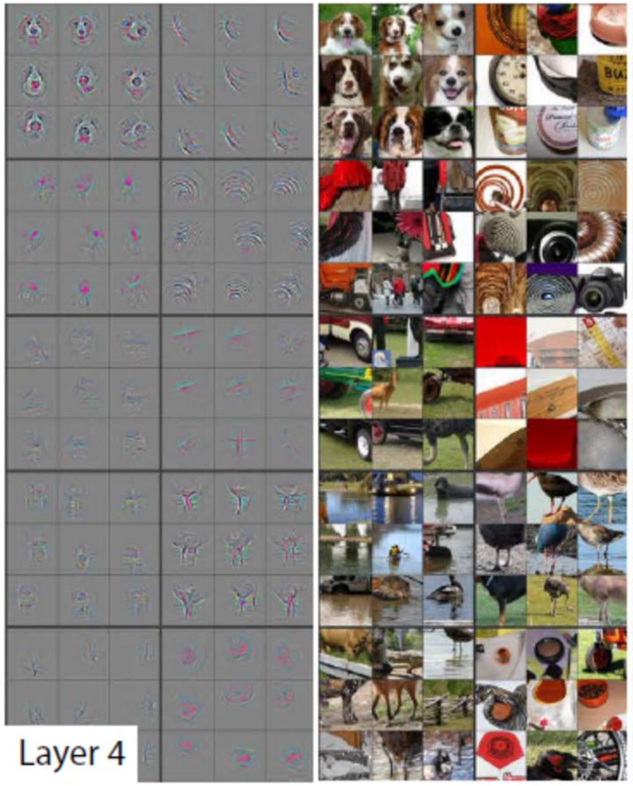


Krizhevsky, Sutskever, Hinton, *ImageNet classification with deep convolutional neural networks,* NIPS'12

Visualization of the convolution filters



Zeiler and Fergus, *Visualizing and Understanding Convolutional Networks*, ECCV'14



Top nine activations

Zeiler and Fergus, *Visualizing and Understanding Convolutional Networks*, ECCV'14

Overview

- History of machine visual perception
- ***State of the art for visual perception***
- Weakly supervised learning and synthetic data

Today's machine visual perception

Machine learning

Large-scale & deep learning

Learning with noisy labels

Data (images & videos)

Large quantity, but quality?

Manual / weakly-supervised annotation, synthetic data

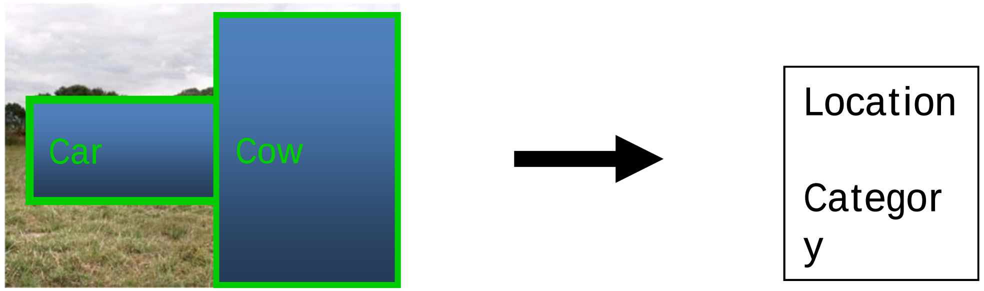
Machine visual perception

Understanding of the visual world

Design of models

Current state of the art – object localization

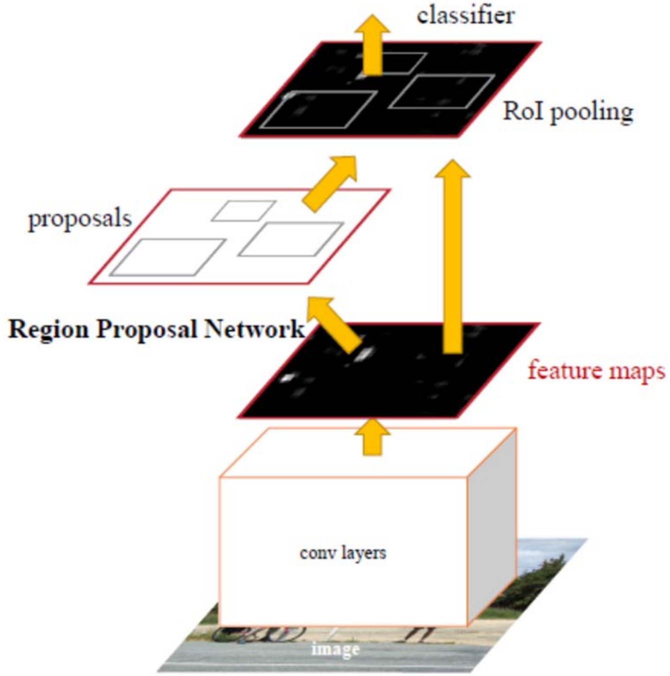
- Object localization



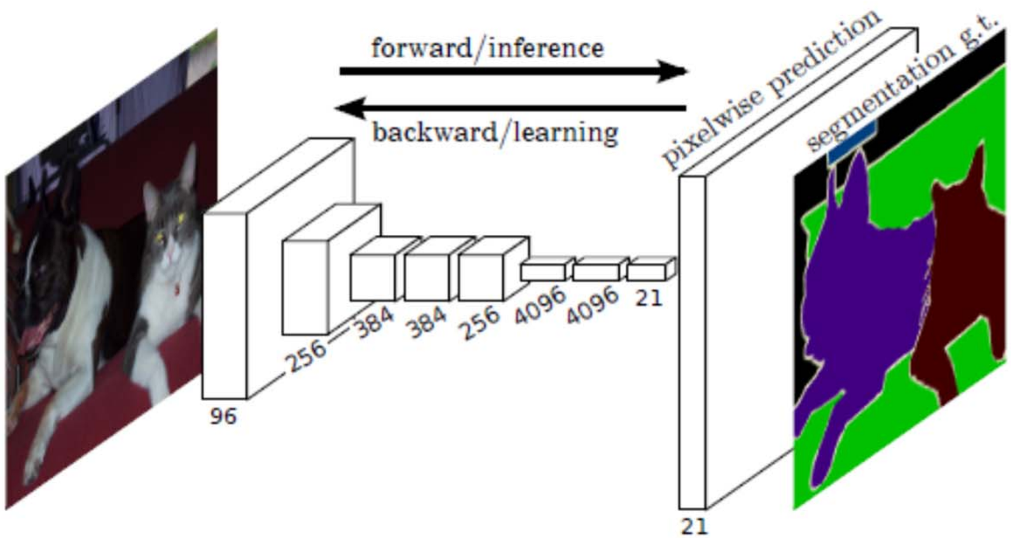
- Region-based CNN features [Girshick'15]

Faster R-CNN for object localization [Girshick'15]

- Region Proposal Network
- ROI pooling
- Classification in object category & background

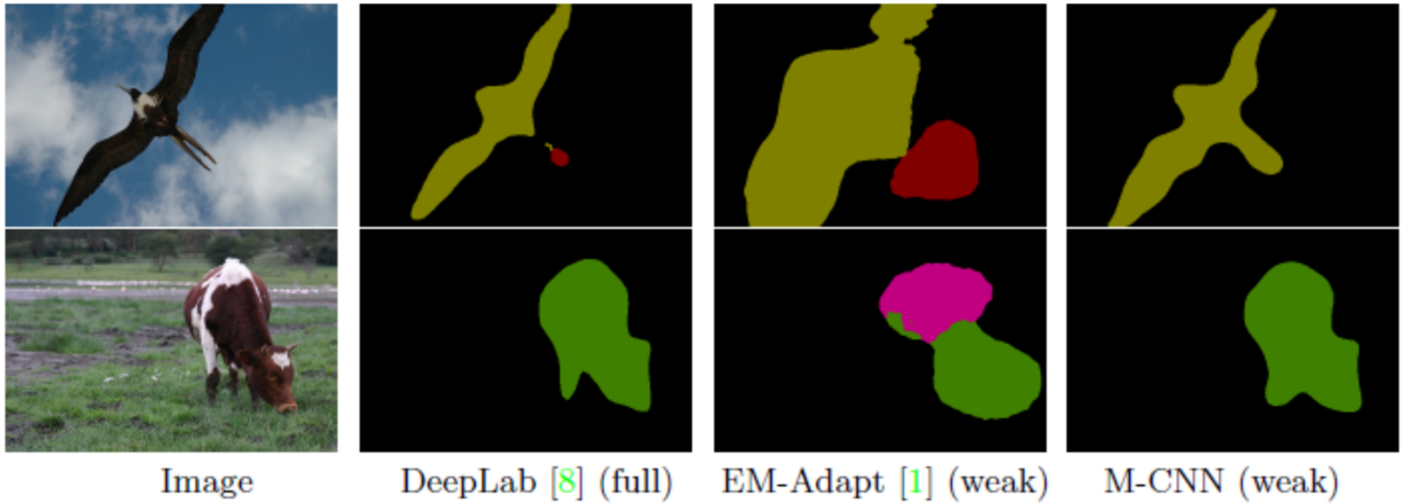


Current state of the art – semantic segmentation



Fully convolutional networks for semantic segmentation [Long et al.'15]

Current state of the art – semantic segmentation



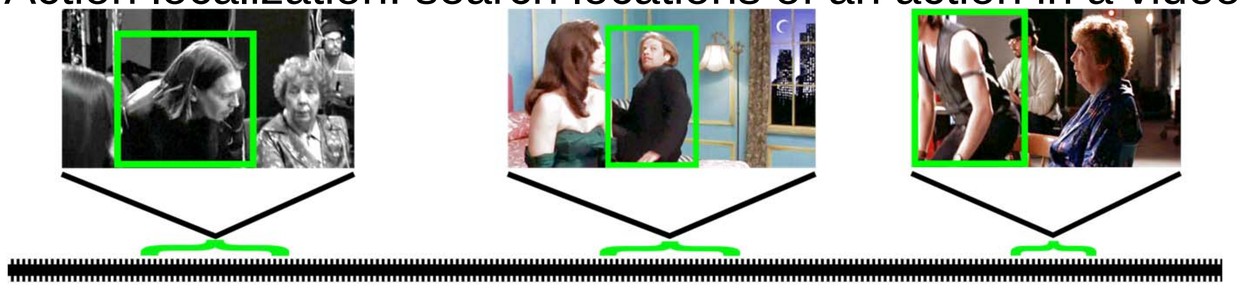
Results for fully- and weakly-supervised semantic segmentation

Current state of the art - action recognition

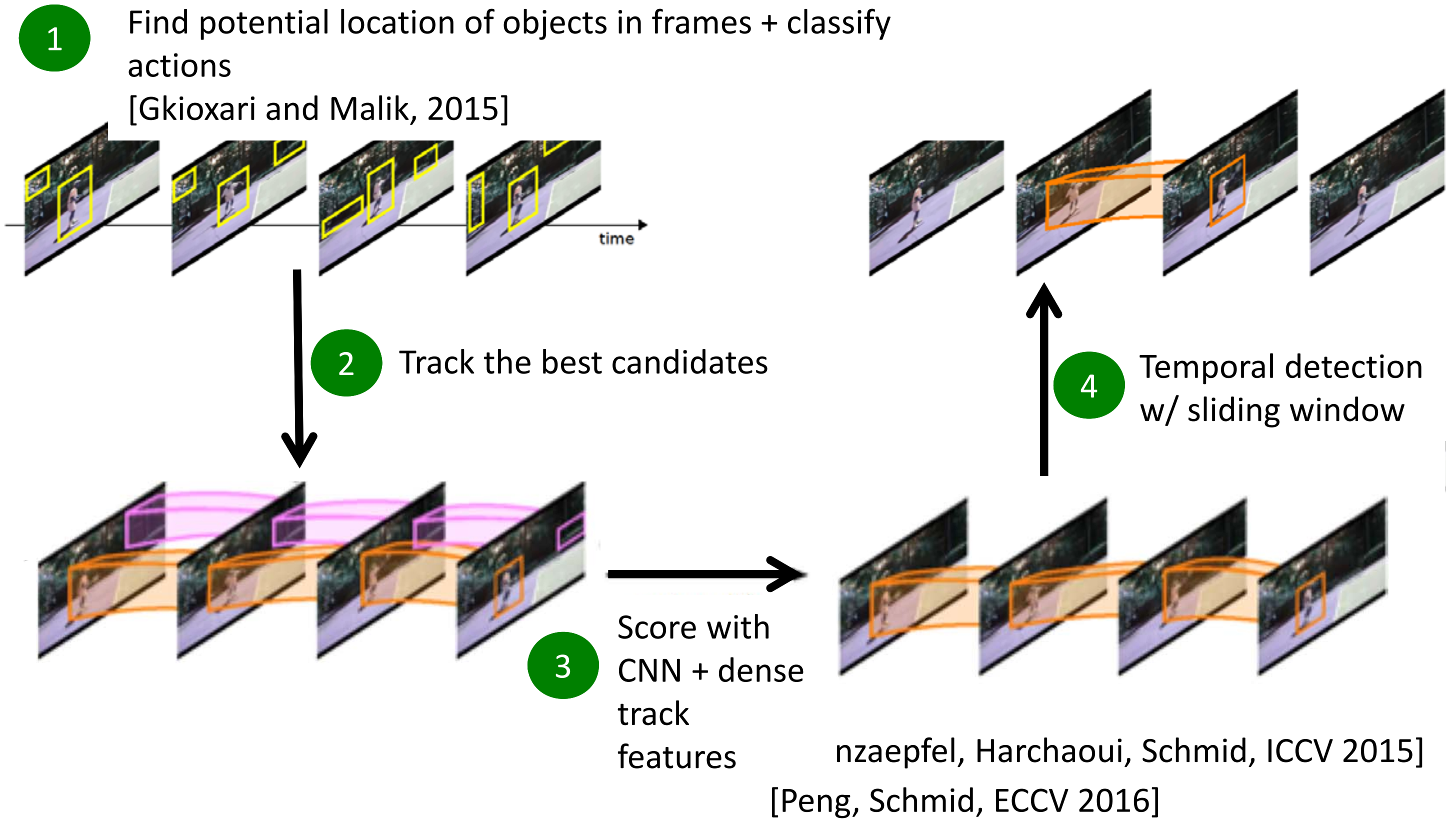
- Action classification: assigning an action label to a video clip

Making sandwich: present
Feeding animal: not present
...

- Action localization: search locations of an action in a video



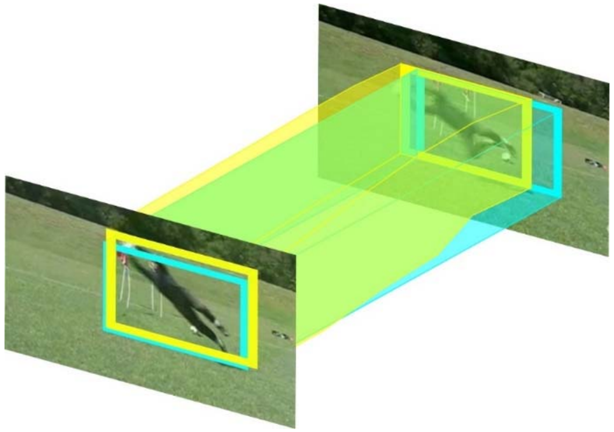
Spatio-temporal action localization

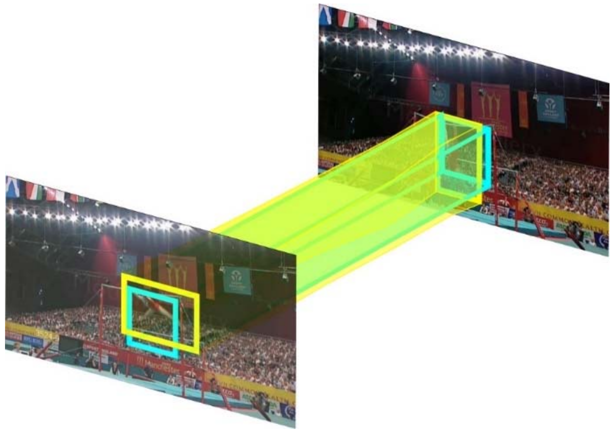


ACtion tubelet detector

Classify and regress spatio-temporal volumes

- *Anchor cuboids*: fixed spatial extent over time
- *Regressed tubelets*: score + deform the cuboid shape



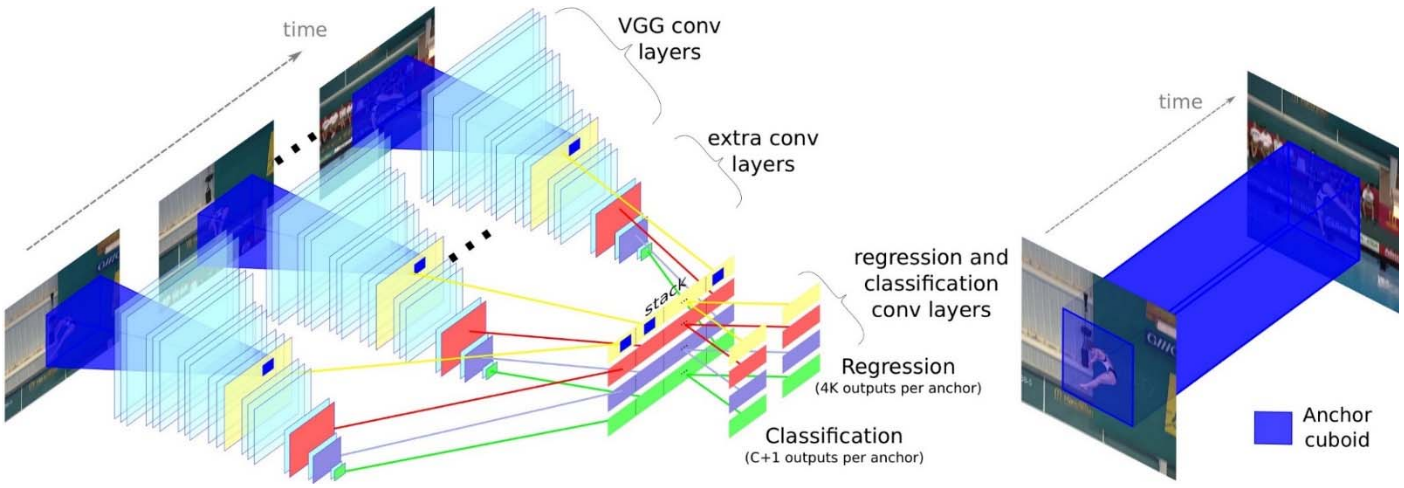


[Action tublet detector for spatio-temporal action localization,
V. Kalogeiton, P. Weinzaephel, V. Ferrari, C. Schmid, ICCV'17]

informatics mathematics
Inria

ACtion tubelet detector

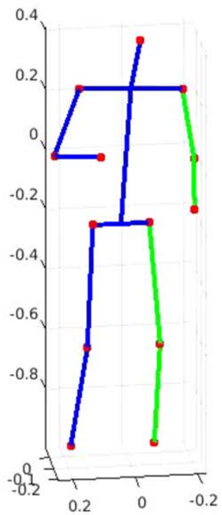
Use sequences of frames to detect *tubelets*: anchor cuboids



SSD detector [Liu et al., ECCV'16]

Current state of the art - 2D & 3D human pose

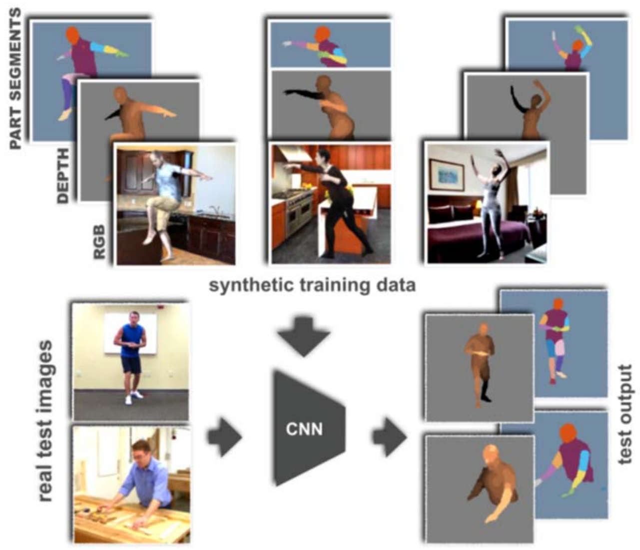
- Impact of human / pose detection
  - Design of more accurate models, with 2D and 3D pose
  - Model interactions with objects



[LCR-Net: Localization-Classification-Regression for Human Pose,
G. Rogez, P. Weinzaepfel, C. Schmid, CVPR'17]

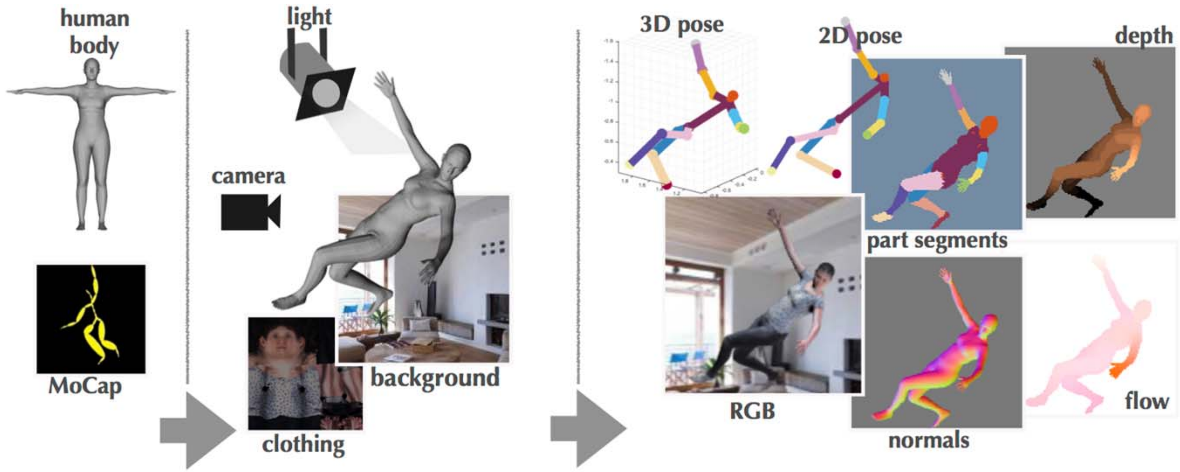
Training with synthetic data

- Learning from Synthetic Humans [Varol, Romero, Martin, Mahmood, Black, Laptev, Schmid, CVPR'17]



SURREAL dataset
Synthetic hUmans foR REAL tasks

a body with *random* 3D shape + 3D pose from MoCap data ➔2D image is rendered with a *random* camera + *random* lighting + *random cloth* texture + a *random* static scene image



Output: RGB image, 2D/3D pose, optical flow, depth image, segmentation map for body parts

Overview

- History of machine visual perception
- State of the art for visual perception
- ***Practical matters***

Practical matters

- Lectures by Cordelia Schmid and Jakob Verbeek
- Online course information
  - Schedule, slides, papers
  - http://thoth.inrialpes.fr/~verbeek/MLOR.17.18.php
- Grading
  - 50% written exam
  - 50% quizes on the presented papers
  - optionally paper presentation, grade for presentation can replace worst grade among quizes

Practical matters

- Paper presentations
  - Each paper is presented by two or three students
  - Presentations last for 15~20 minutes
  - Send email with your choice of presentation
  - Papers on the web site

Master internships

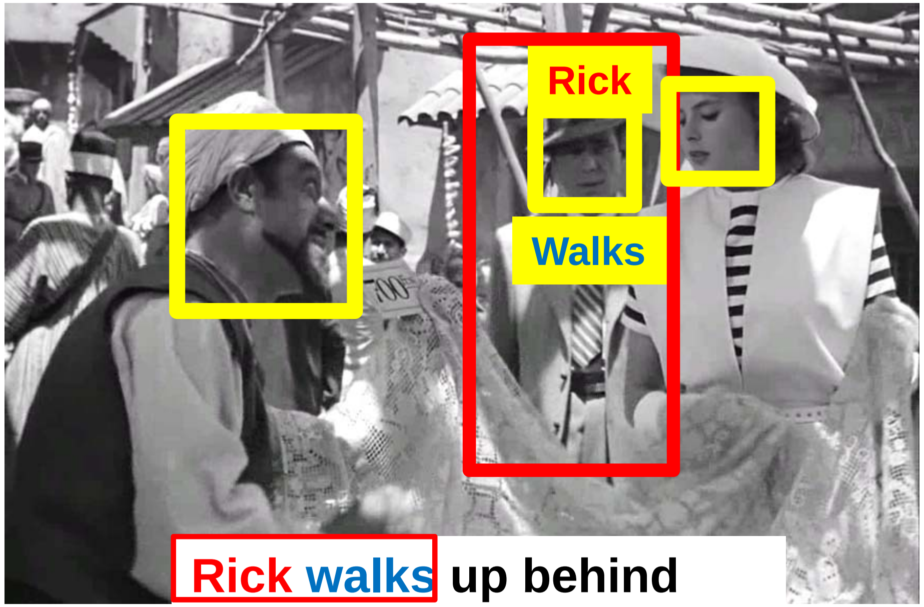
- See https://thoth.inrialpes.fr/jobs
- Or contact the members of the team directly

Cross-modal learning for scene understanding

Supervisors: K. Alahari & C. Schmid

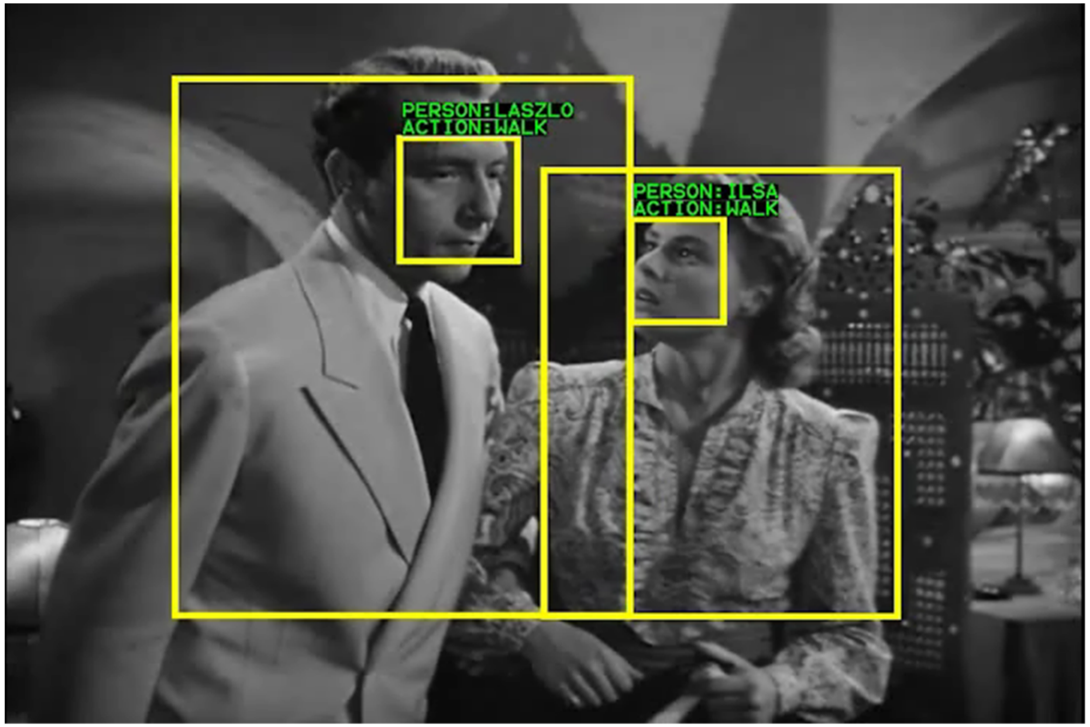
[Bojanowski et al., ICCV 2013]

Cross-modal learning for scene understanding

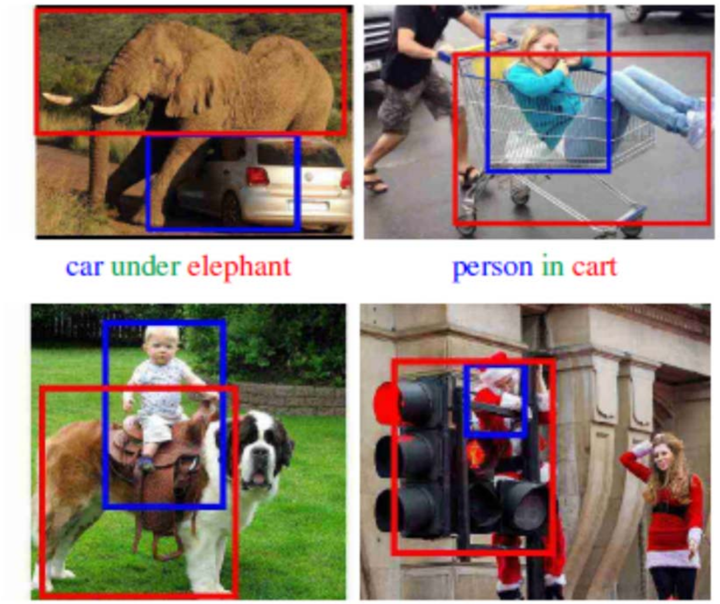


Rick walks up behind Ilsa

[Bojanowski et al., ICCV 2013]


PERSON:LASZLO
ACTION:WALK
PERSON:ILSA
ACTION:WALK

Cross-modal learning for scene understanding



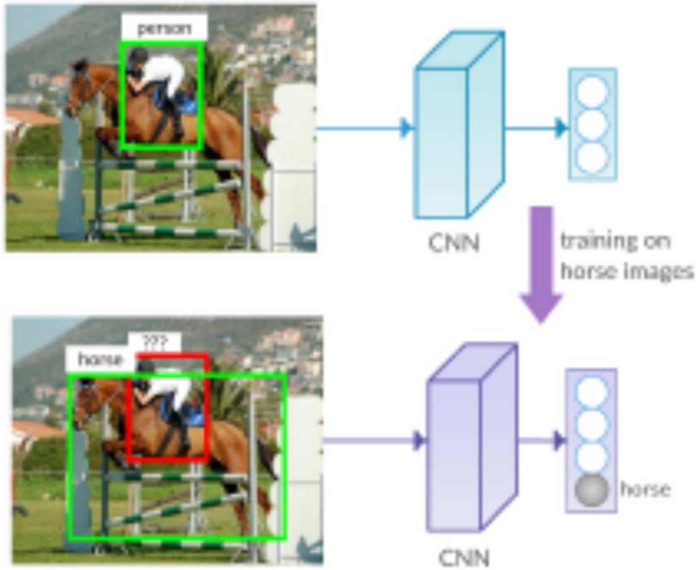
car under elephant | person in cart

person ride dog | person on top of traffic light

[Weakly-supervised learning of visual relations,
J. Peyre, I. Laptev, C. Schmid, J. Sivic, ICCV'17]

Incremental learning for scene understanding

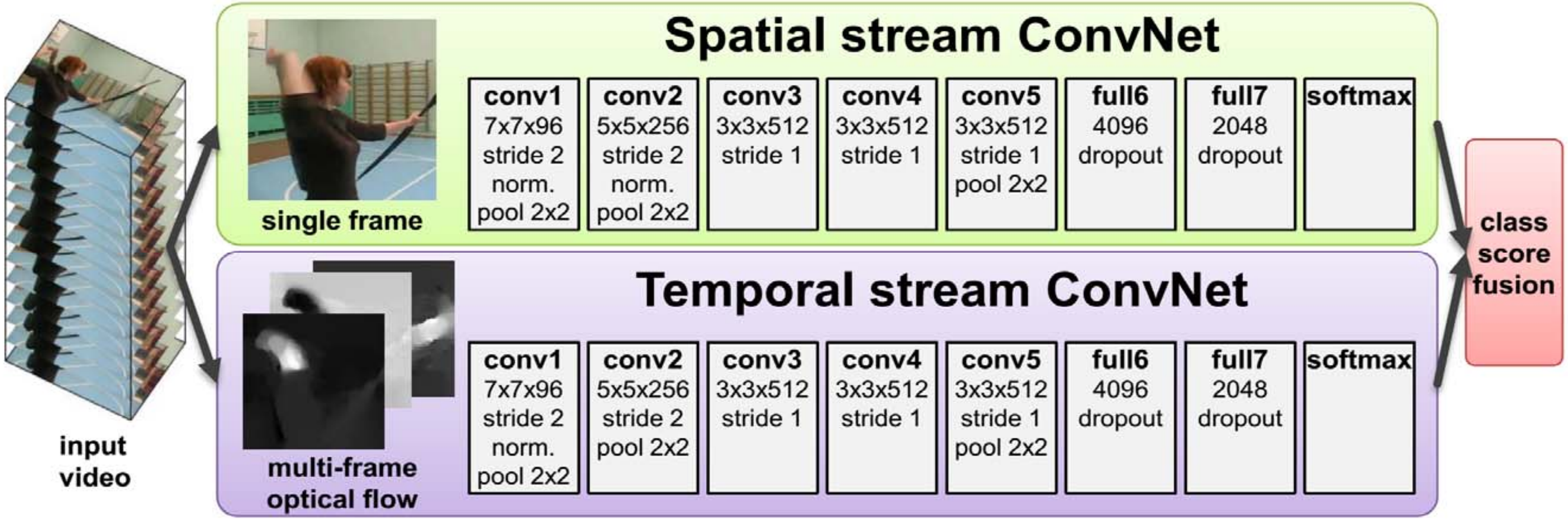
Supervisors: K. Alahari & C. Schmid



[Incremental learning of object detectors without catastrophic forgetting, K. Shmelkov, C. Schmid, K. Alahari, ICCV'17]

End-to-end architectures for large-scale video recognition

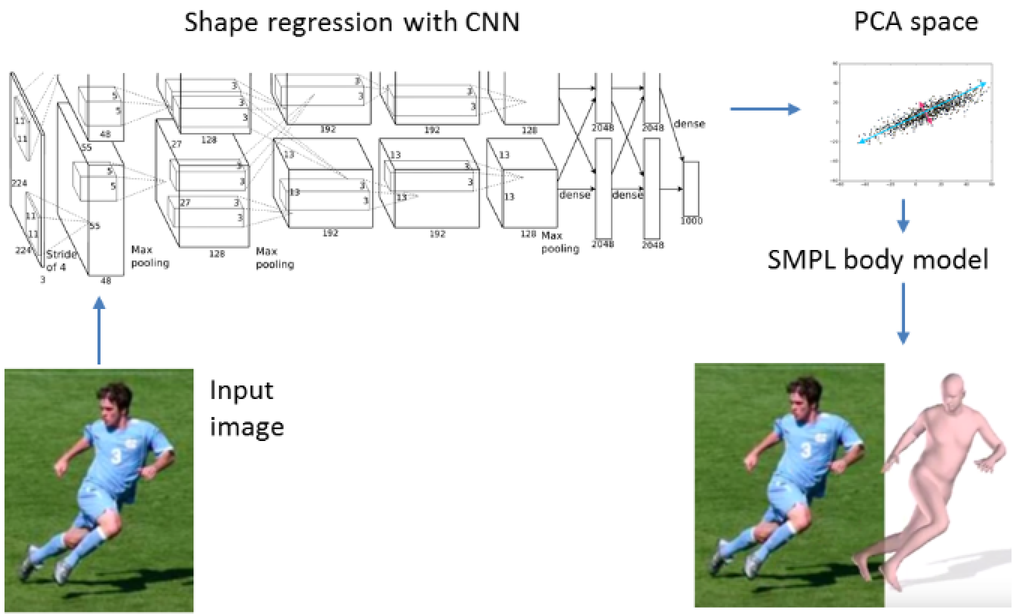
Supervisors: P. Weinzaephel (NAVER) & C. Schmid



[Simonyan, K., & Zisserman, A. Two-stream convolutional networks for action recognition in videos. NIPS 2014.]

Human 3D shape estimation from a single image

Supervisors: G. Rogez, JS Franco & C. Schmid



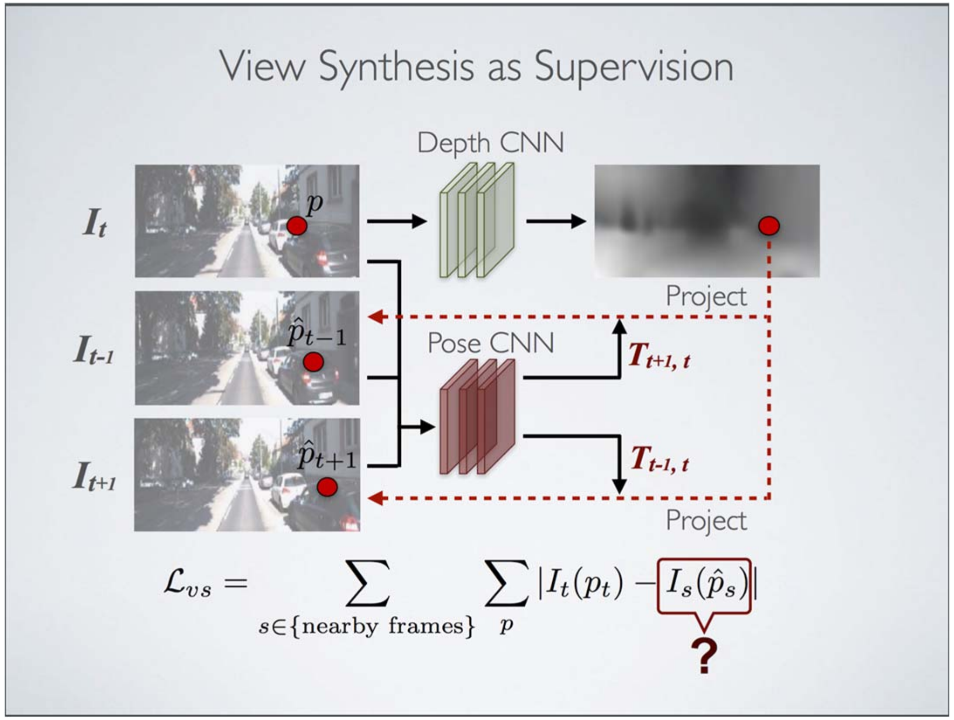
Learning to grasp with visual guidance

Supervisors: C. Schmid, A. Pashevich

- Design hierarchical reinforcement learning techniques
- Integrate object category model information into grasping

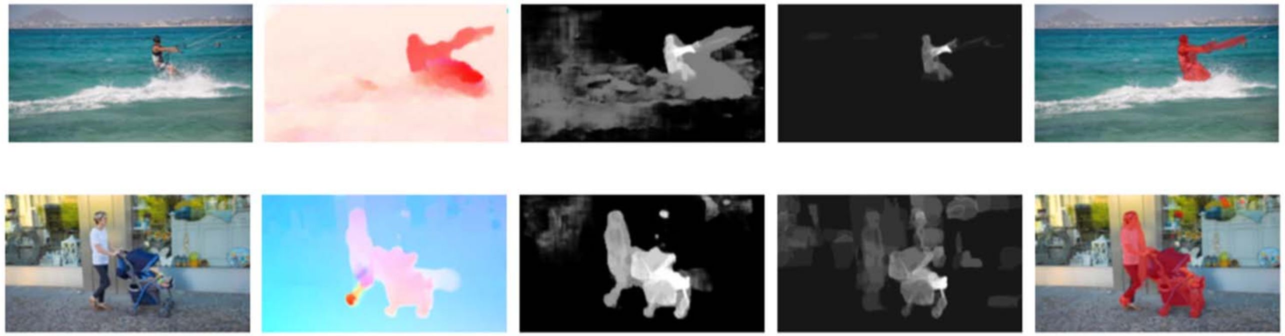
Motion estimation from 3D depth maps

Supervisors: C. Schmid



[Zhou, Brown, Snavely, Lowe, CVPR'17]

Motion estimation in real videos



DAVIS LDOF MP-Net Objectness Result

[Learning Motion Patterns, Tokmakov, Alahari, Schmid, CVPR’17]