

Object recognition and Computer Vision 2017

<http://www.di.ens.fr/willow/teaching/recvis17/>

Introduction and basic camera geometry

Josef Sivic

<http://www.di.ens.fr/~josef>

INRIA, WILLOW, ENS/INRIA/CNRS UMR 8548

Departement d'Informatique, Ecole Normale Supérieure, Paris

With slides from: **S. Lazebnik, J. Ponce, S. Seitz, R. Szeliski**

Object recognition and Computer Vision 2017

<http://www.di.ens.fr/willow/teaching/recvis17/>

Lectures:



Jean Ponce



Cordelia Schmid



Ivan Laptev



Armand Joulin

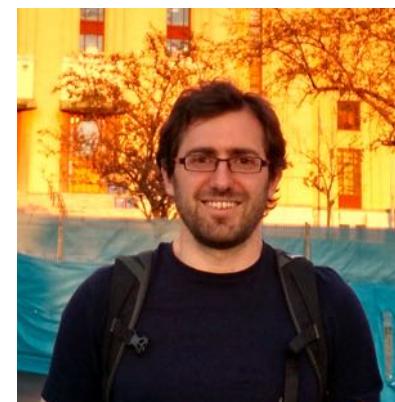


Mathieu Aubry

Teaching assistants:



Gul Varol



Ignacio Rocco

Lecture	Date	Topic and reading materials.	Slides
1	Oct 3	Introduction; Camera geometry (3hrs, J. Sivic)	
2	Oct 10	Instance-level recognition I. - Local invariant features, correspondence, image matching (3hrs, J. Sivic);	
3	Oct 17	Instance-level recognition II. - Efficient visual search (1.5hrs, C. Schmid) Bag-of-feature models for category-level recognition (1.5hrs, C. Schmid)	
4	Oct 24	ICCV 2017. No lecture. Assignments: Assignment 1 due.	
5	Oct 31	Sparse coding and dictionary learning for image analysis (3hrs, J. Ponce)	
6	Nov 7	Neural networks; Optimization methods (3hrs, A. Joulin) Assignments: Assignment 2 due.	
7	Nov 14	Convolutional neural networks for visual recognition I. (I. Laptev) Final project topics are out. Due date for project proposals: Nov 28.	
8	Nov 21	Convolutional neural networks for visual recognition II. (J. Sivic) Assignments: Assignment 3 due.	
9	Nov 28	Motion and human actions I. (C. Schmid) Assignments: Final project proposal due.	
10	Dec 5	Human pose estimation; Weakly-supervised learning I (I. Laptev)	
11	Dec 12	3D object recognition and Convolutional neural networks (M. Aubry) Weakly-supervised learning II (I. Laptev)	
12	Jan 15 Jan 16 Jan 17	Final project presentations and evaluation (I. Laptev, J. Sivic) Jan 15: 13:00-17:00 Jan 16: 13:00-17:00 Jan 17: 13:00-17:00 The presentations will take place at Salle Alan Turing - 1st floor at Inria Paris research center, 2 Rue Simone Iff, 75012, Paris. Directions are here. When you enter the building tell the receptionist you are going for the presentation and go directly to the first floor (no special access card is needed).	

Object recognition and computer vision 2017

Class webpage:

<http://www.di.ens.fr/willow/teaching/recvis17>

Grading:

- 3 programming assignments (50%)
 - Instance-level recognition
 - Image classification
 - Convolutional Neural Networks
- Final project (50%)
More independent work, resulting in a report and a class presentation.

Policy

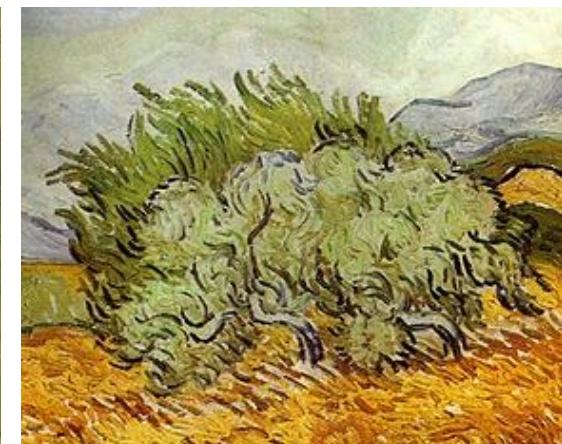
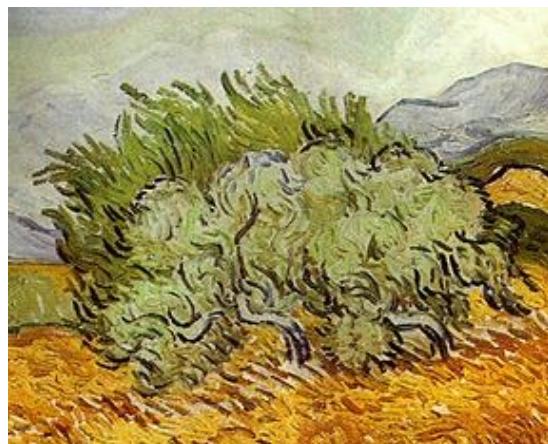
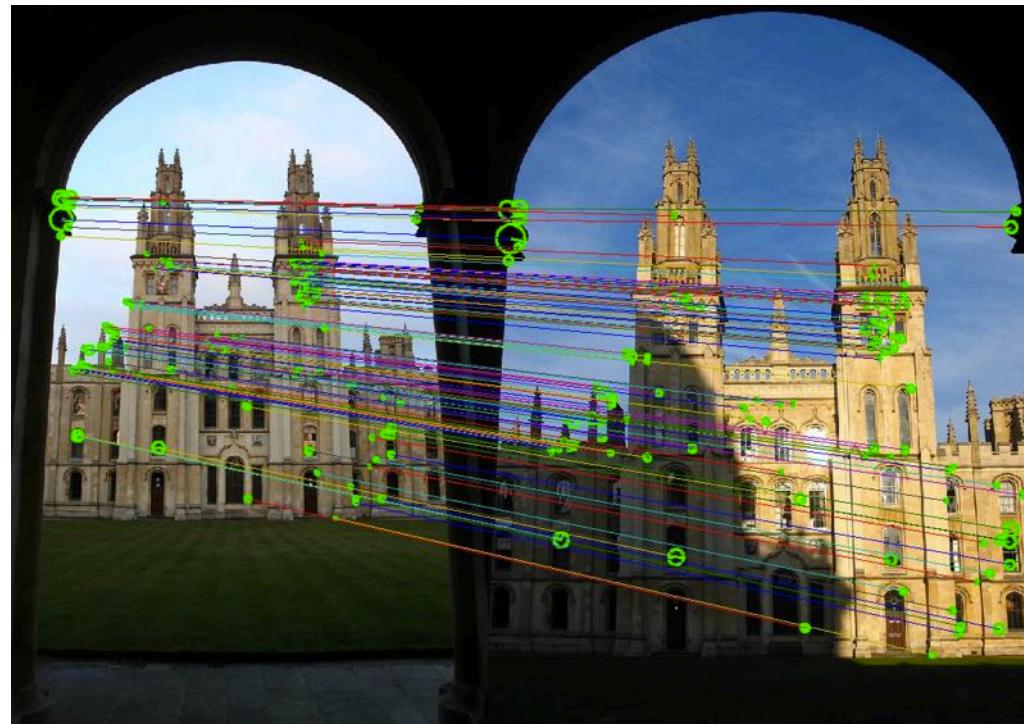
Assignments are strictly individual

Copy-paste of the code, results, parts → 0p. of the report

FPs can be done in groups of max 2 people

Assignment I: Instance level recognition

- Part I: Sparse features for matching specific objects in images
- Part II: Affine co-varient detectors
- Part III: Towards large scale retrieval
- Part IV: Large scale retrieval



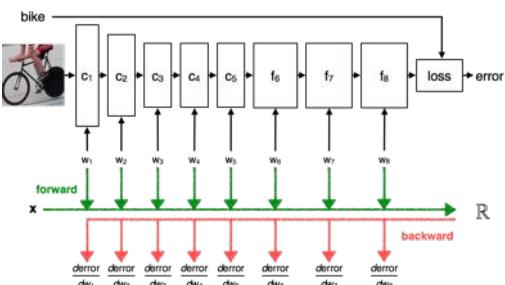
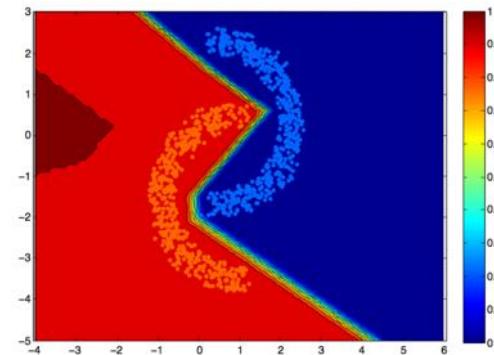
Assignment II: Image Classification

- Part 1: Training and testing an Image Classifier
- Part 2: Training an Image Classifier for Retrieval using Google images



Assignment III: Image Classification with Convolutional Neural Networks

- Part 1: Neural Network's theory:
 - Forward pass, Backward pass
 - Parameter update
- Part 2: PASCAL VOC Image classification with ConvNet features
- Part 3: Convolutional neural networks



Final project

- Select the topic + write project proposal
 - Present the work in the class
 - Write project report
-
- Can be done individually or as a **group of max 2 people**
 - The proposed project topics are from the recent top-conference publications in computer vision, see example topics from 2016 here: <http://www.di.ens.fr/willow/teaching/recvis16/>
 - Student-defined projects are welcome
 - Final project can be joint with another MVA course

Register on the class webpage

On the class website, fill-in the **registration form** (name, email, school)

News:

- Fill in [this form](#) if you want to follow the course.

We will create an account for you on the **Moodle**.

Assignments / reports will be collected using the **Moodle**.

Assignment 1 is out. Due on Oct 24th.

Recap: TODO

1. Register by filling-in form on the class webpage (asap)
2. Fill-in Doodle for Matlab tutorial (by Wed Oct 10 6pm)
3. Assignment 1 – Instance-level recognition
<http://www.di.ens.fr/willow/teaching/recvis17/assignment1/>
Due in 3 weeks: Oct 24 2017

Matlab tutorial

Please fill-in a Doodle linked from the class webpage
by **Wed Oct 10 6pm.**

The tutorial will be at:

INRIA/Willow, 2 Rue Simone Iff, Paris

Who should participate?

- Students with no or limited experience with Matlab.

Introduction to computer vision

<http://imagine.enpc.fr/~aubrym/lectures/introvis17/>

Thursdays 9:00 - 12:00 at Salle R, ENS ULM.

Taught by Mathieu Aubry.

M1 course

Covers the basics of computer vision in detail.



Mathieu Aubry

Research

Both WILLOW (J. Ponce, I. Laptev, J. Sivic) and LEAR (C. Schmid) groups are active in computer vision and visual recognition research.

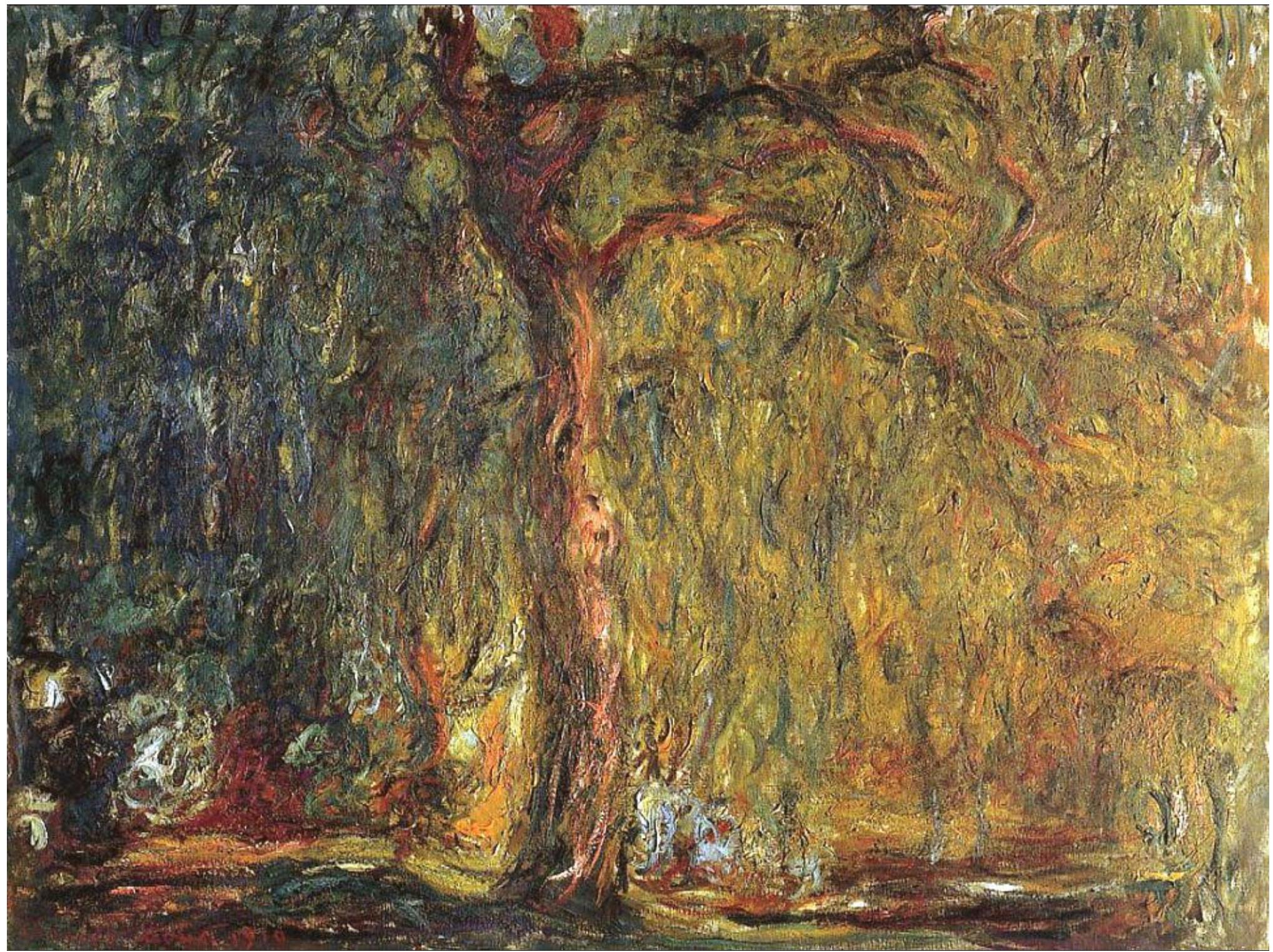
<http://www.di.ens.fr/willow/>

<http://lear.inrialpes.fr/>

with close links to SIERRA – machine learning (F. Bach)

<http://www.di.ens.fr/sierra/>

There will be master internships available.
Talk to us if you are interested!

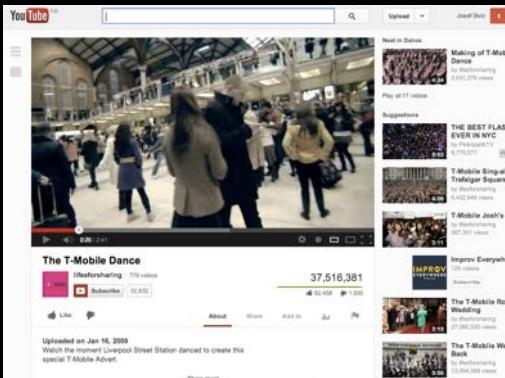


Outline

- What computer vision is about
- What this class is about
- A brief history of visual recognition
- A brief recap on geometry

Why is visual recognition important? Images are all around us

Archives of visual information



Internet videos



10,000+ TV channels



Historical imagery

Cameras around us



2M+ surveillance cameras



Car cameras

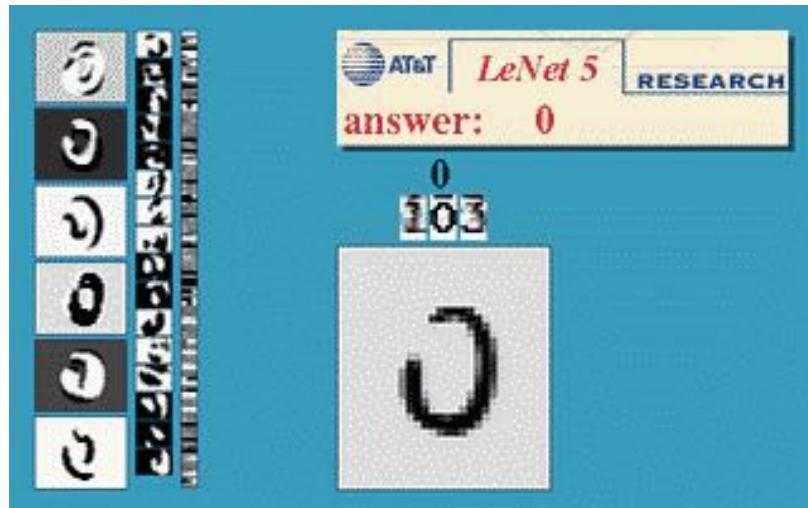


Personal cameras

Optical character recognition (OCR)

Technology to convert scanned docs to text

- If you have a scanner, it probably came with OCR software



Digit recognition, AT&T labs
<http://www.research.att.com/~yann/>



License plate readers
http://en.wikipedia.org/wiki/Automatic_number_plate_recognition

Face detection



- Many new digital cameras now detect faces
 - Canon, Sony, Fuji, ...

Smile detection

The Smile Shutter flow

Imagine a camera smart enough to catch every smile! In Smile Shutter Mode, your Cyber-shot® camera can automatically trip the shutter at just the right instant to catch the perfect expression.



[Sony Cyber-shot® T70 Digital Still Camera](#)

Object recognition (in supermarkets)



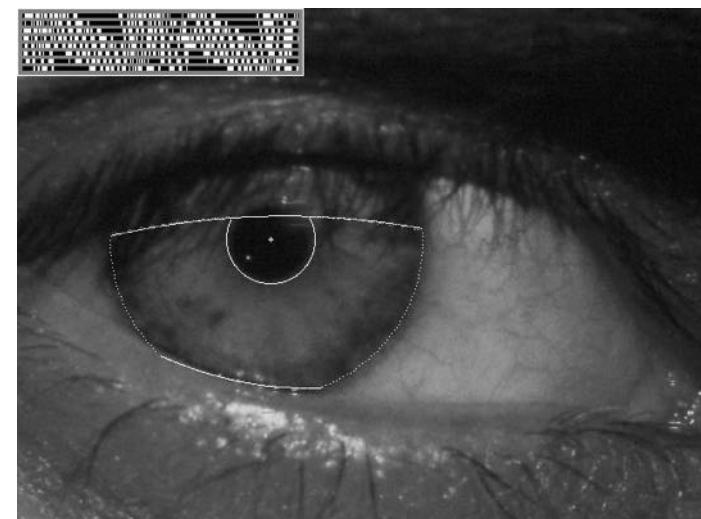
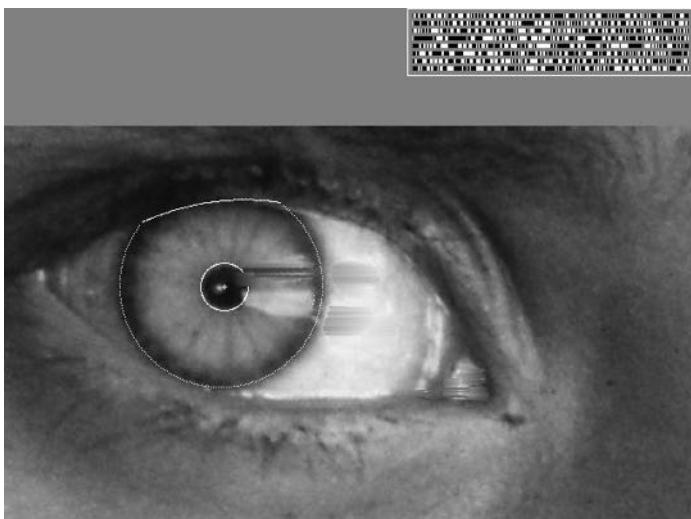
[LaneHawk by EvolutionRobotics](#)

“A smart camera is flush-mounted in the checkout lane, continuously watching for items. When an item is detected and recognized, the cashier verifies the quantity of items that were found under the basket, and continues to close the transaction. The item can remain under the basket, and with LaneHawk, you are assured to get paid for it... “

Vision-based biometrics



“How the Afghan Girl was Identified by Her Iris Patterns” Read the [story](#)
[wikipedia](#)



Login without a password...



Fingerprint scanners on
many new laptops,
other devices



Face recognition systems now
beginning to appear more widely
<http://www.sensiblevision.com/>



FaceID on the latest iPhone

Object recognition (in mobile phones)



Point & Find, Nokia

Google Goggles

Helping to find content

- Adobe Stock auto-keywording

Adobe Stock | Portfolio | Uploaded Files | Sales | Contributor Account | Upload | BUY | MORGAN | Adobe

New In review Rejected

Status: All (6) ▾ File type: All (6) ▾

Select all

29041982.jpg

What kind of media is this?

Photo

Title (max 200 characters)

Woman blowing soap bubbles

Keywords language

English

Keywords (min 5 - max 50)

1 Woman

2 Bubbles

3 Soap

4 Colorful

5 Summer

6 Enter keyword No.6

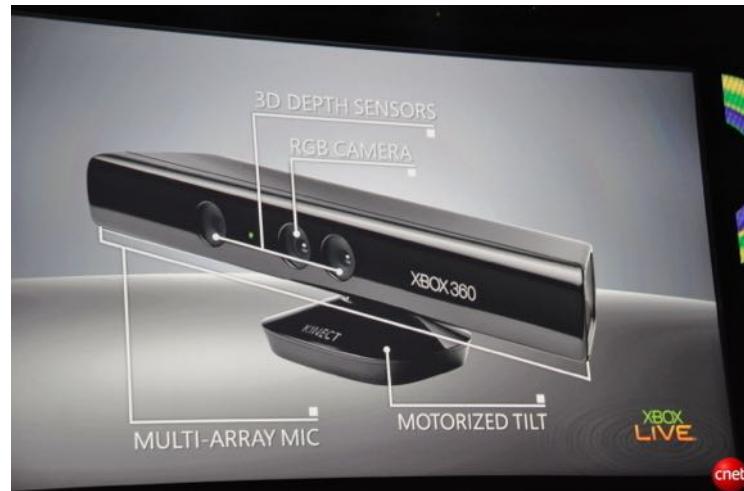
Category

Select a category

The screenshot shows the Adobe Stock contributor dashboard. On the left, there are six thumbnail images with their respective status (Incomplete), rating (0), and download count (0). To the right of each image is a detailed view panel. The first image shows silhouettes of people against a sunset/sunrise sky, with keywords: Woman, Bubbles, Soap, Colorful, Summer, and Enter keyword No.6. The second image shows a group of people from behind, hugging, with keywords: Woman, Bubbles, Soap, Colorful, Summer, and Enter keyword No.6. The third image shows a woman blowing colorful soap bubbles, with keywords: Woman, Bubbles, Soap, Colorful, Summer, and Enter keyword No.6. The fourth image shows a carved pumpkin with a scary face, with keywords: Woman, Bubbles, Soap, Colorful, Summer, and Enter keyword No.6. The fifth image shows a skier performing a jump against a bright sun, with keywords: Woman, Bubbles, Soap, Colorful, Summer, and Enter keyword No.6. The sixth image shows a person standing with arms raised in a field of tall grass under a clear blue sky, with keywords: Woman, Bubbles, Soap, Colorful, Summer, and Enter keyword No.6.

Interactive Games: Kinect

- Object Recognition:
<http://www.youtube.com/watch?feature=iv&v=fQ59dXOo63o>
- Mario: <http://www.youtube.com/watch?v=8CTJL5IUjHg>
- 3D: <http://www.youtube.com/watch?v=7QrnwoO1-8A>
- Robot: <http://www.youtube.com/watch?v=w8BmgtMKFbY>



Smart cars

Slide content courtesy of Amnon Shashua

The screenshot shows the Mobileye website homepage. At the top, there are tabs for "manufacturer products" and "consumer products". A main banner features a car from above with three cameras highlighted: "rear looking camera" at the back, "forward looking camera" at the front, and "side looking camera" on the sides. Below the banner, there are three main sections: "EyeQ Vision on a Chip" (with an image of a chip), "Vision Applications" (with an image of a person walking across a crosswalk), and "AWS Advance Warning System" (with an image of a display screen). To the right, there are two boxes: "News" containing links to articles about Volvo's collision warning system and a general news link, and "Events" containing links to Mobileye's participation in Equip Auto and SEMA shows, along with a "read more" link.

- > **EyeQ** Vision on a Chip
- > **Vision Applications**
Road, Vehicle, Pedestrian Protection and more
- > **AWS** Advance Warning System

News

- > **Mobileye Advanced Technologies Power Volvo Cars World First Collision Warning With Auto Brake System**
- > **Volvo: New Collision Warning with Auto Brake Helps Prevent Rear-end**

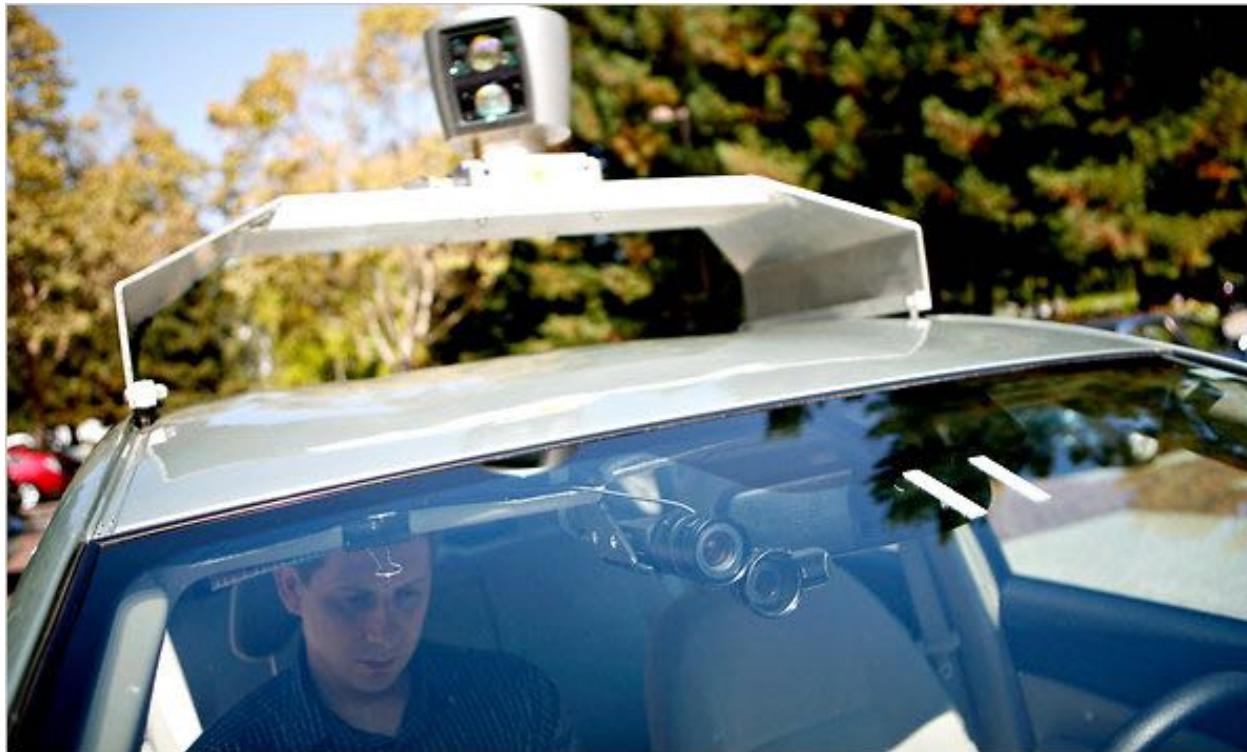
Events

- > [Mobileye at Equip Auto, Paris, France](#)
- > [Mobileye at SEMA, Las Vegas, NV](#)

> [read more](#)

- [Mobileye](#)
 - Bought by Intel: 15.3 Billion USD
 - See also CVPR 2016 [keynote](#)

Google cars



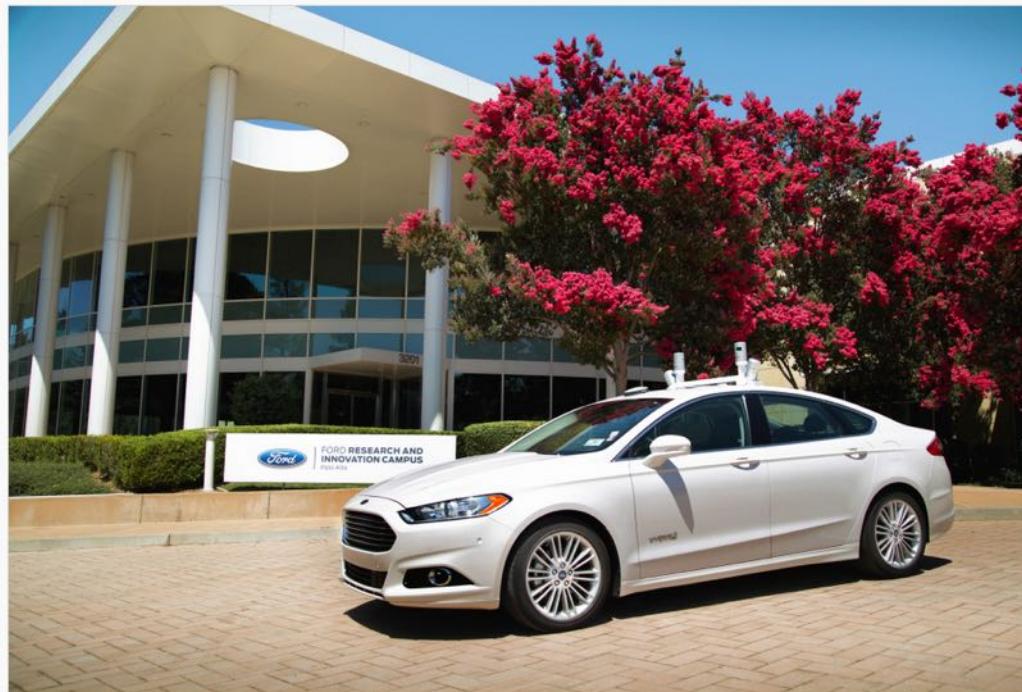
Oct 9, 2010. "[Google Cars Drive Themselves, in Traffic](#)". *The New York Times*. John Markoff

June 24, 2011. "[Nevada state law paves the way for driverless cars](#)". *Financial Post*. Christine Dobby

Aug 9, 2011, "[Human error blamed after Google's driverless car sparks five-vehicle crash](#)". *The Star (Toronto)*

Ford acquires SAIPS for self-driving machine learning and computer vision tech

Posted Aug 16, 2016 by [Darrell Etherington \(@etherington\)](#)



Ford outlined a few of the ways it's aiming to [ship driverless cars by 2021](#), and part of the plan involves acquisitions. CEO Mark Fields revealed at a press event in Palo Alto today that the automaker [acquired SAIPS](#), an Israeli company focusing on machine learning and computer vision. It's also partnering exclusively with Nirenberg Neuroscience, to bring more "humanlike intelligence" to machine learning components of driverless car systems.

SAIPS' technology brings image and video processing algorithms, as well as deep learning tech focused on processing and classifying input signals, all key ingredients in the special sauce that makes up autonomous vehicle tech. This company's expertise should help with on-board interpretation of data captured by sensors on Ford's self-driving cars, and turning that data into usable info for the car's virtual driver system. SAIPS' offerings include detection of anomalies, persistent tracking of objects detected by sensors, and much more. The company's past clients include HP and Trax, but its partner group doesn't appear to have included much in the way of driving-specific applications.

CrunchBase

Ford Motor Company

FOUNDED
1903

OVERVIEW
Ford is an automotive company that develops, manufactures, distributes, and services vehicles, parts, and accessories worldwide. It operates through two sectors: automotive and financial services. The automotive sector offers vehicles primarily under the Ford and Lincoln brand names. This sector markets cars, trucks, parts, and accessories through retail dealers in North America and distributors ...

LOCATION
Dearborn, MI

CATEGORIES
Automotive

WEBSITE
<http://www.ford.com/>

[Full profile for Ford Motor Company](#)

TC NEWSLETTERS

The Daily Crunch

Our top headlines
Delivered daily

TC Week-in-Review

Top stories of the week
Delivered weekly

CrunchBase Daily

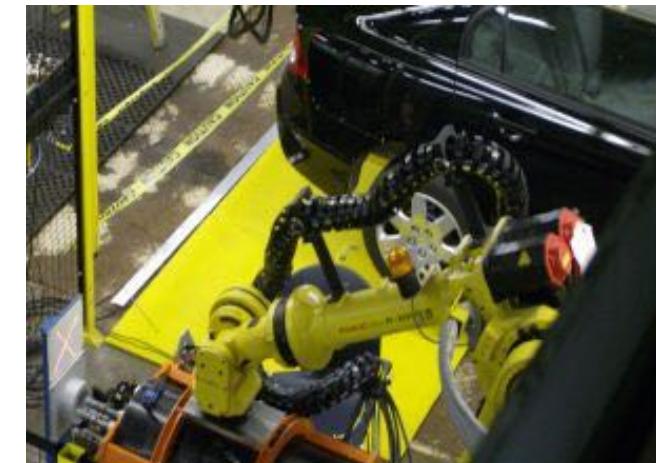
The latest

VALEO LAUNCHES VALEO.AI, THE FIRST GLOBAL RESEARCH CENTER IN ARTIFICIAL INTELLIGENCE AND DEEP LEARNING FOR AUTOMOTIVE APPLICATIONS BASED IN PARIS

Valeo Group | #OpenInnovation

Paris, 14 June 2017 – Valeo launches the first global research center in artificial intelligence and deep learning dedicated to automotive applications.

Industrial robots



Vision-guided robots position nut runners on wheels

Vision in space

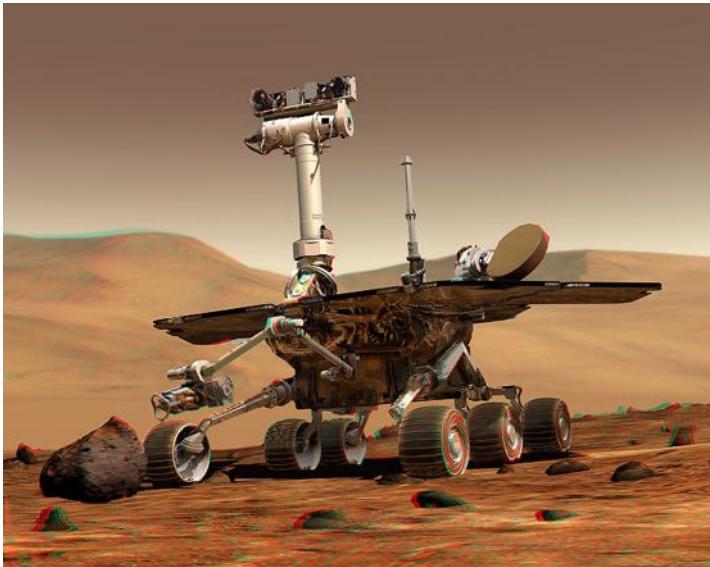


[NASA'S Mars Exploration Rover Spirit](#) captured this westward view from atop a low plateau where Spirit spent the closing months of 2007.

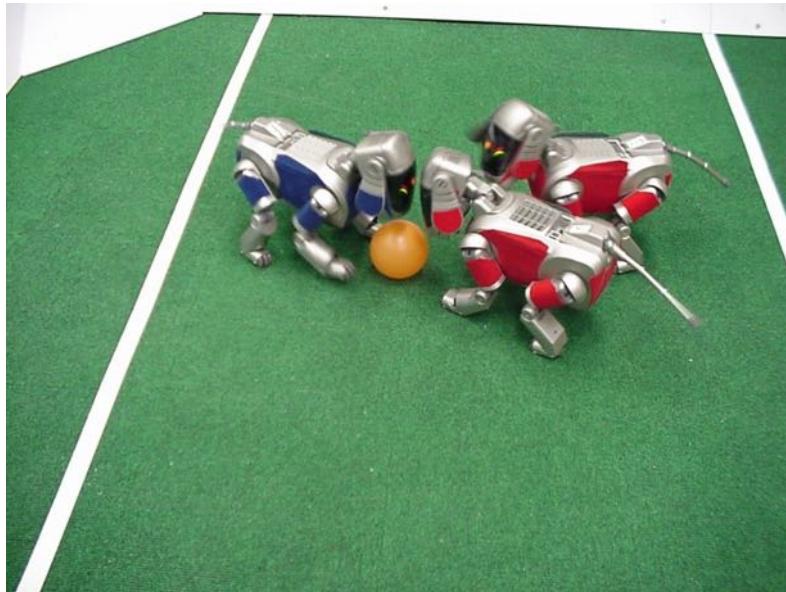
Vision systems (JPL) used for several tasks

- Panorama stitching
- 3D terrain modeling
- Obstacle detection, position tracking
- For more, read "[Computer Vision on Mars](#)" by Matthies et al.

Mobile robots



NASA's Mars Spirit Rover
http://en.wikipedia.org/wiki/Spirit_rover



<http://www.robocup.org/>

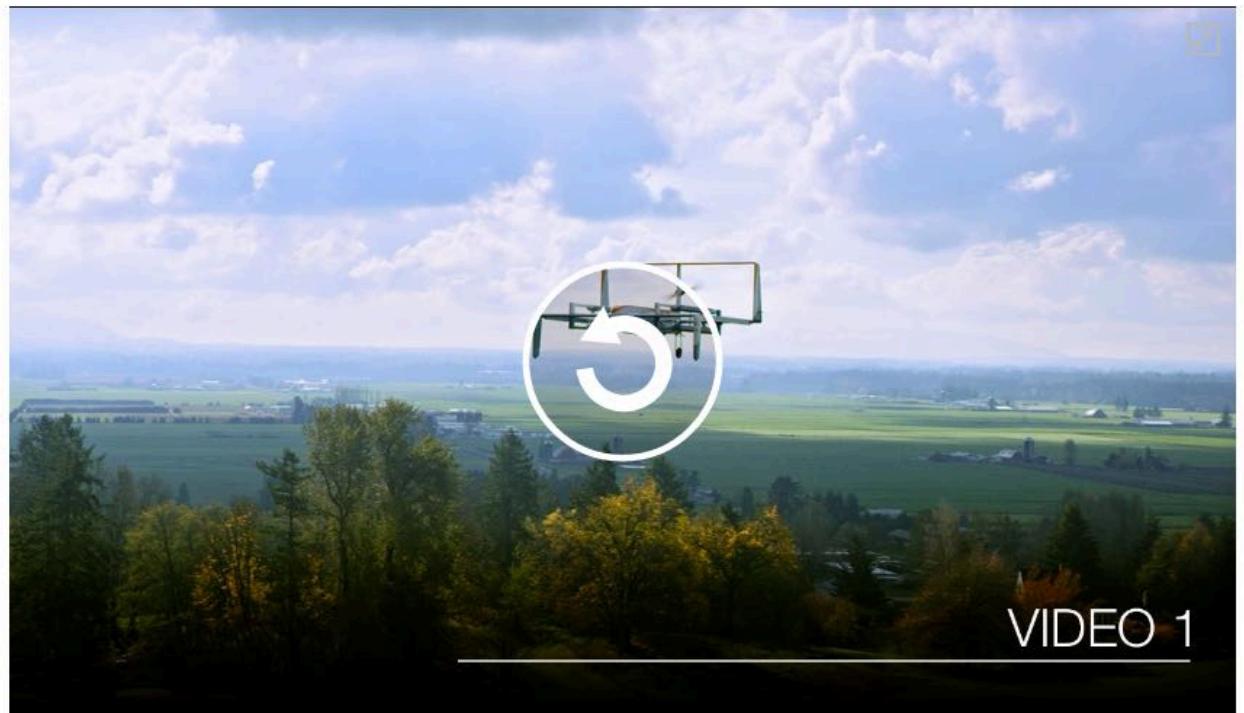
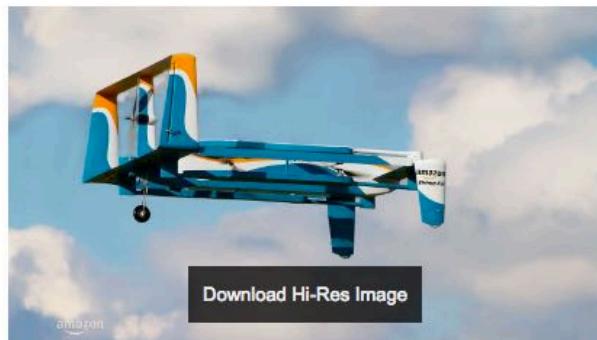


Saxena et al. 2008
[STAIR](#) at Stanford

Amazon Prime Air

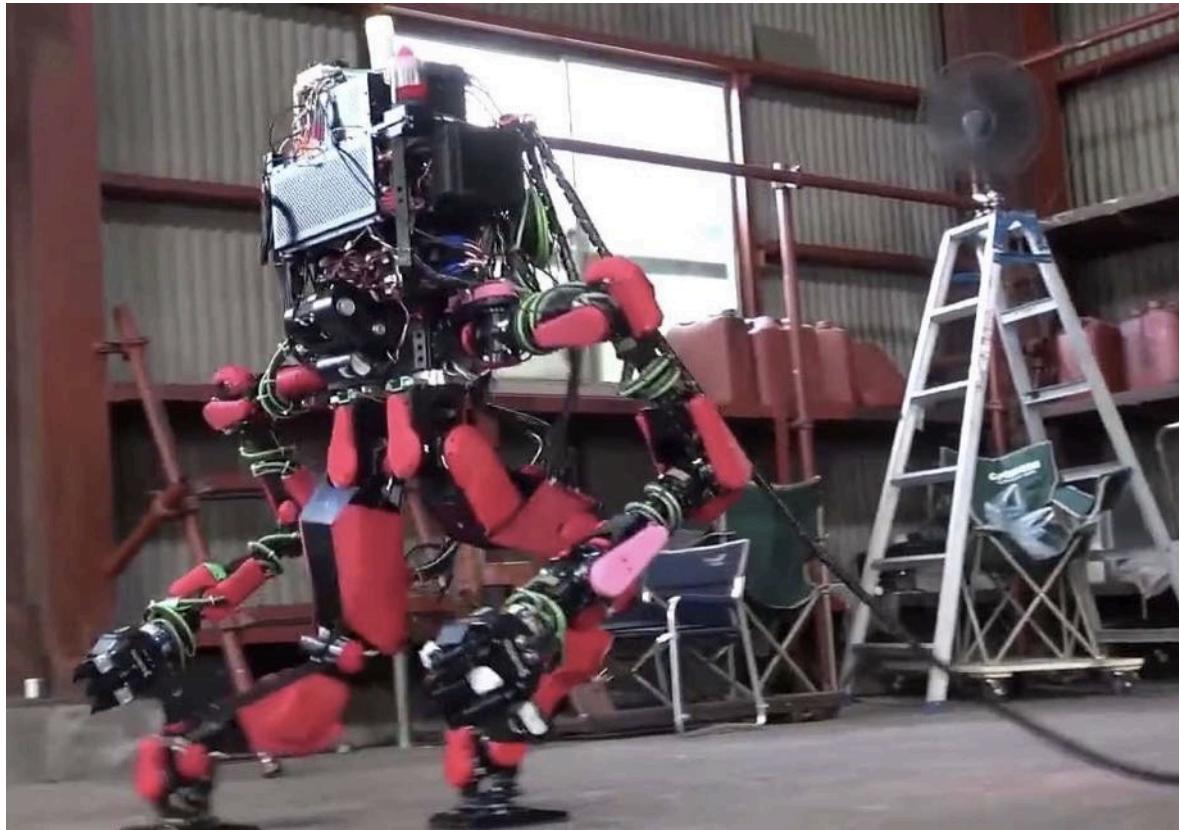


We're excited about Prime Air — a future delivery system from Amazon designed to safely get packages to customers in 30 minutes or less using small unmanned aerial vehicles, also called drones. Prime Air has great potential to enhance the services we already provide to millions of customers by providing rapid parcel delivery that will also increase the overall safety and efficiency of the transportation system. Putting Prime Air into service will take some time, but we will deploy when we have the regulatory support needed to realize our vision.



<https://www.amazon.com/b?node=8037720011>

Assistive robotics for industry and home



[Robot by Schaft, DARPA robot challenge 2015]

Augmented Reality and Virtual Reality



Magic Leap, Oculus, Hololens, etc.

Virtual assistants at home and industry



To assist people
[Microsoft HoloLens 2015]

Computer vision as a job

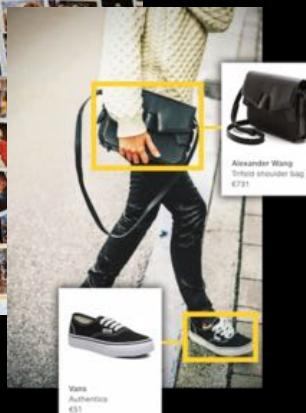
Vincent Delaitre

Phd, 2015

Start-up Deepomatic.com



<https://www.inria.fr/centre/paris/actualites/deepomatic-un-shazam-de-l-image-pour-le-e-commerce-et-les-medias>



Guillaume Seguin

Phd, 2016

Start-up regaind.io



Piotr Bojanowski

Phd, 2016

Facebook AI Research



Oliver Whyte

Phd, 2012

Engineer at Microsoft



Mathieu Aubry

Phd, 2015

Faculty at ENPC



Relja Arandjelovic

Post-doc, 2016

Google DeepMind

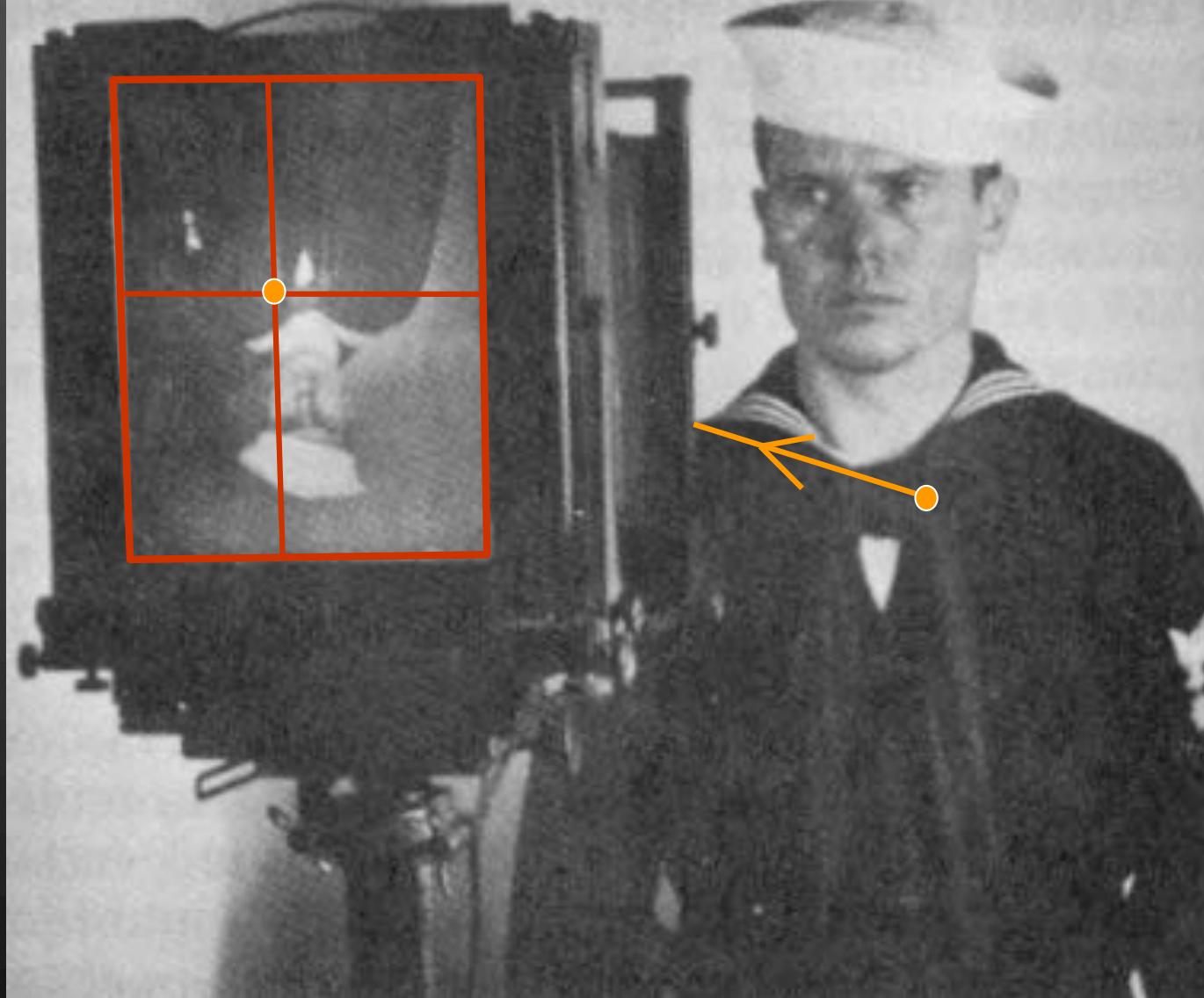


Why is computer vision difficult?

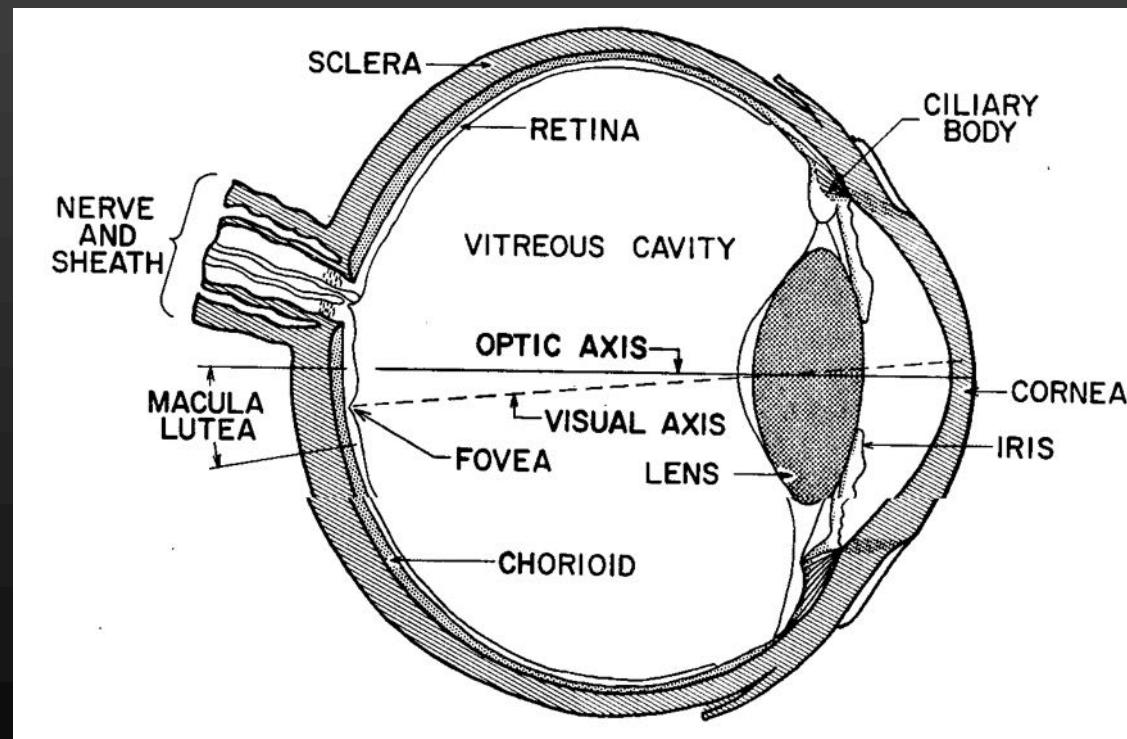
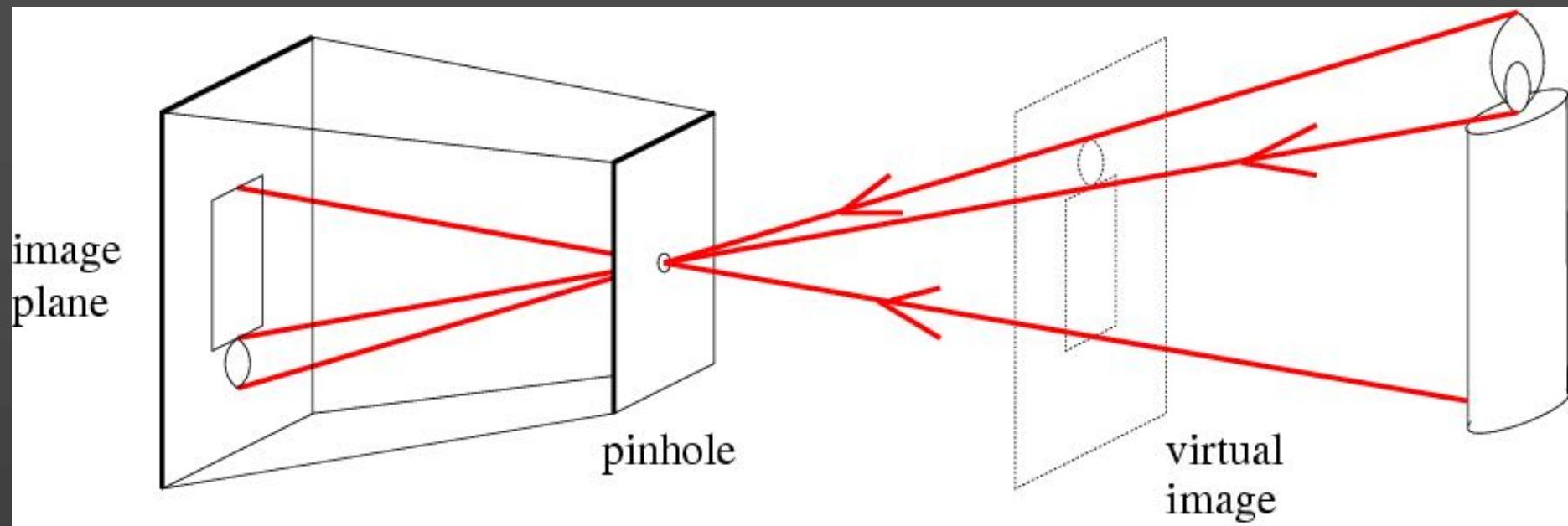


Image source: J. Koenderink

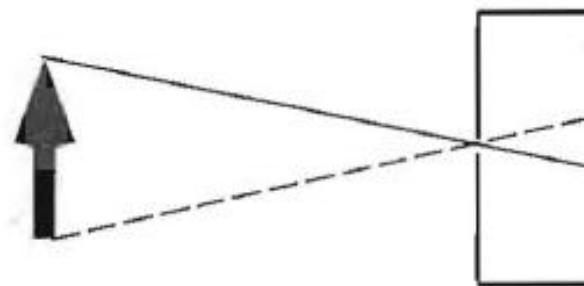
They are formed by the projection of three-dimensional objects.



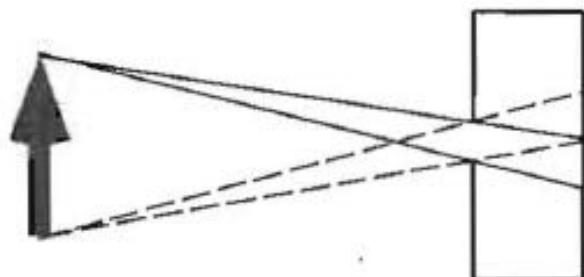
Images are brightness/color patterns drawn in a plane.



Pinhole camera: trade-off between sharpness and light transmission



A. Pinhole Aperture without Lens --> Sharp Image



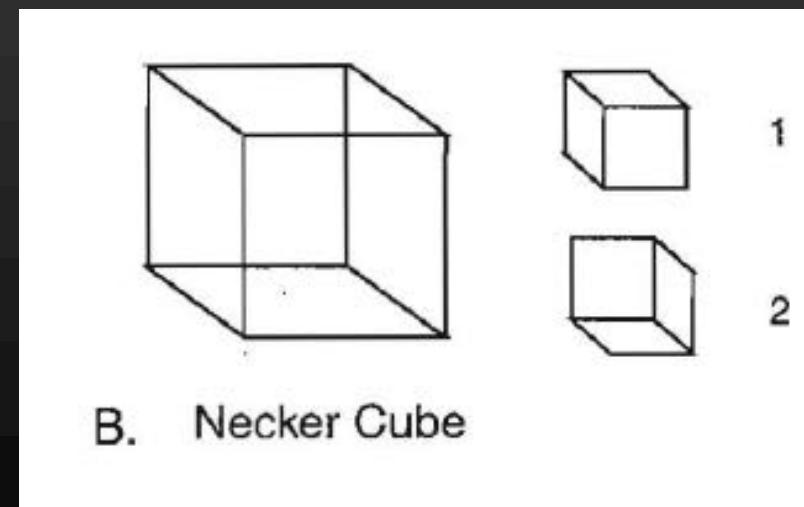
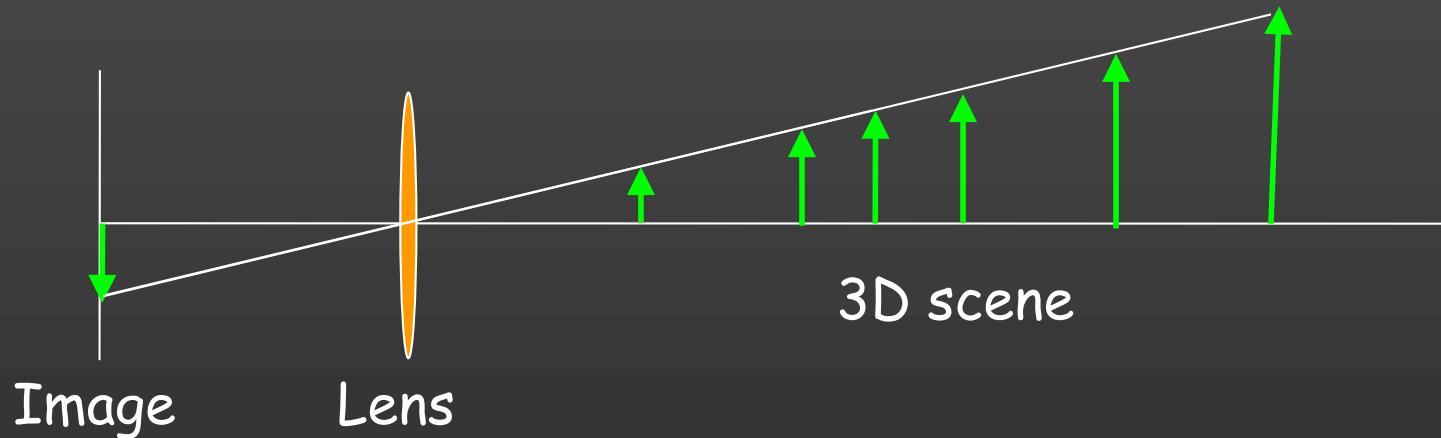
B. Large Aperture without Lens --> Fuzzy Image



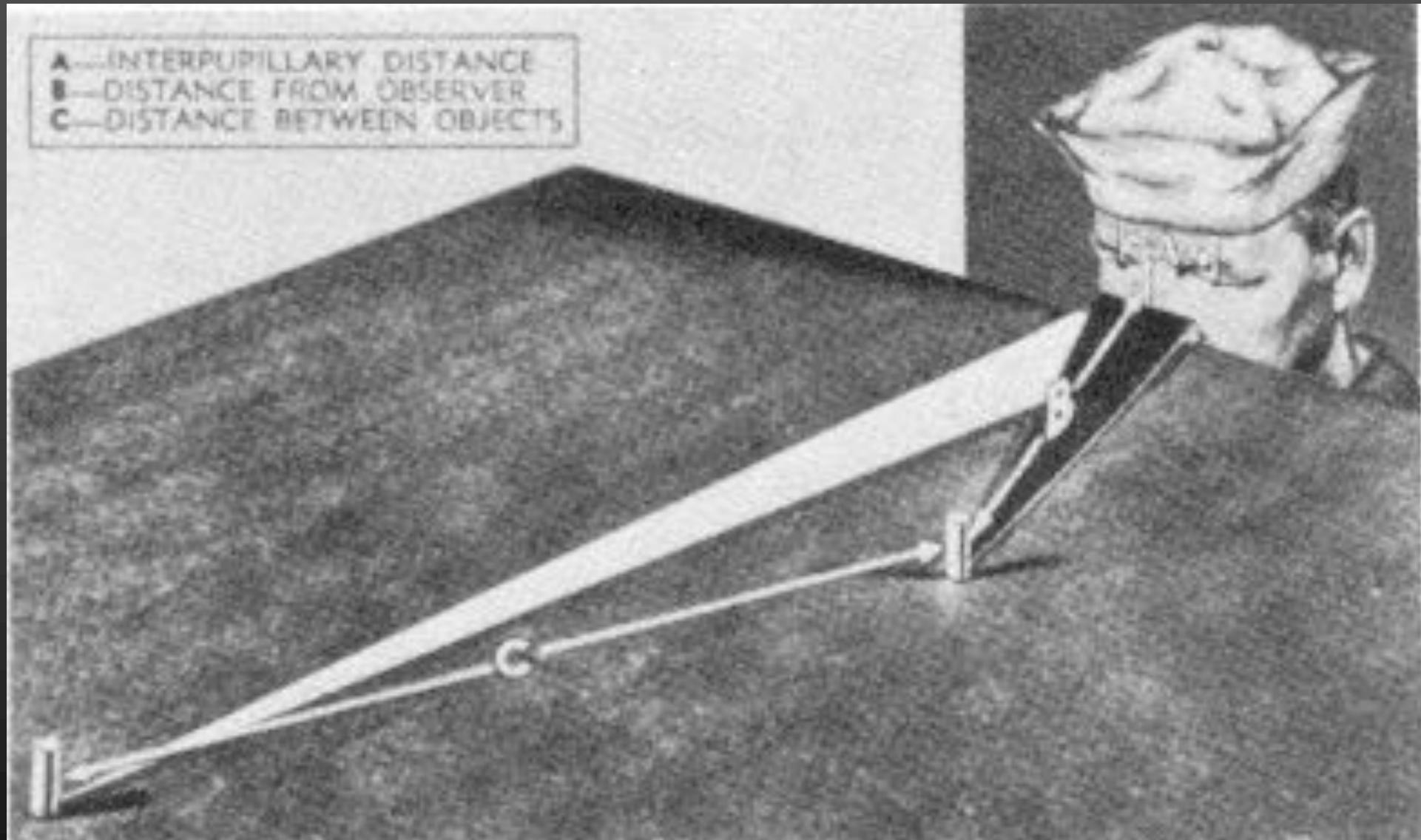
Camera Obscura in
Edinburgh

Fundamental problem I:
3D world is “flattened” to 2D images

→ Loss of information



Question : how do we see "in 3D" ?



(First-order) answer: with our two eyes.

Simulated 3D perception



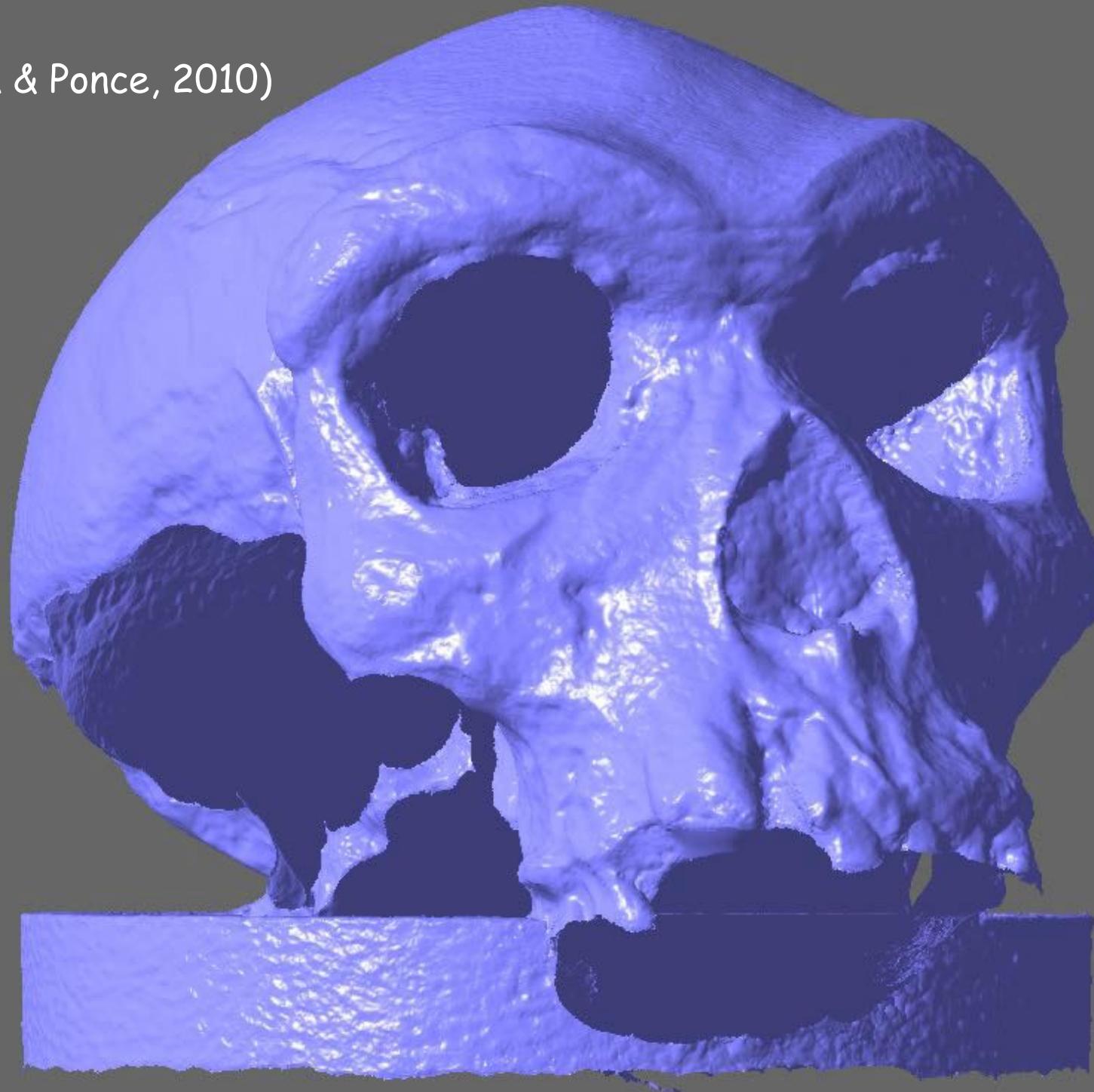
Disparity



© Getty Images

PMVS

(Furukawa & Ponce, 2010)



But there are other cues..





© 2002 National Geographic Society. All rights reserved.

NATIONAL GEOGRAPHIC.COM

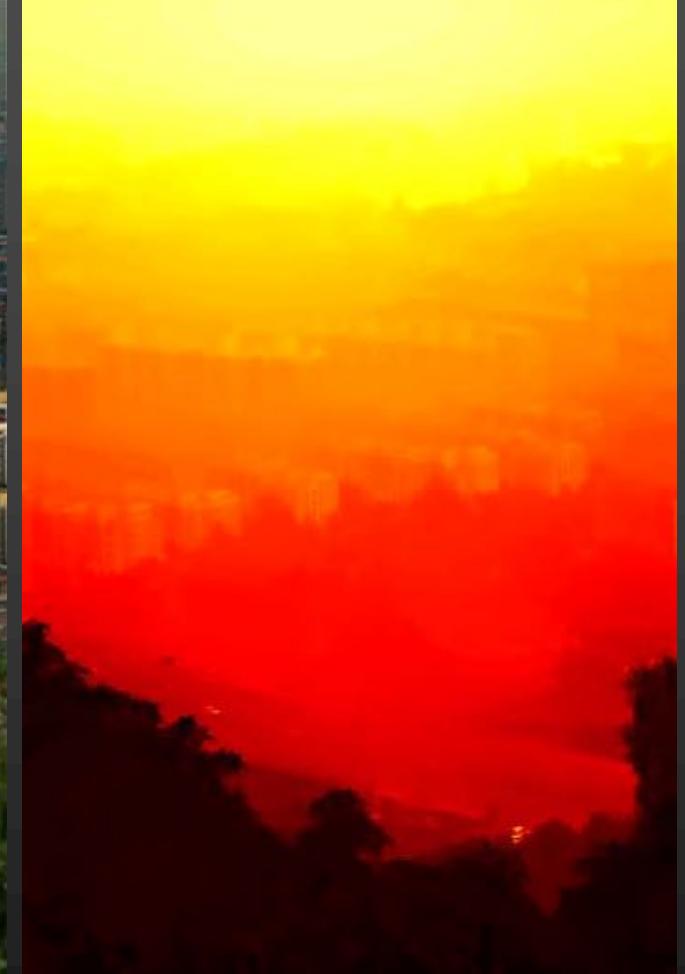
Depth from haze



Input haze image

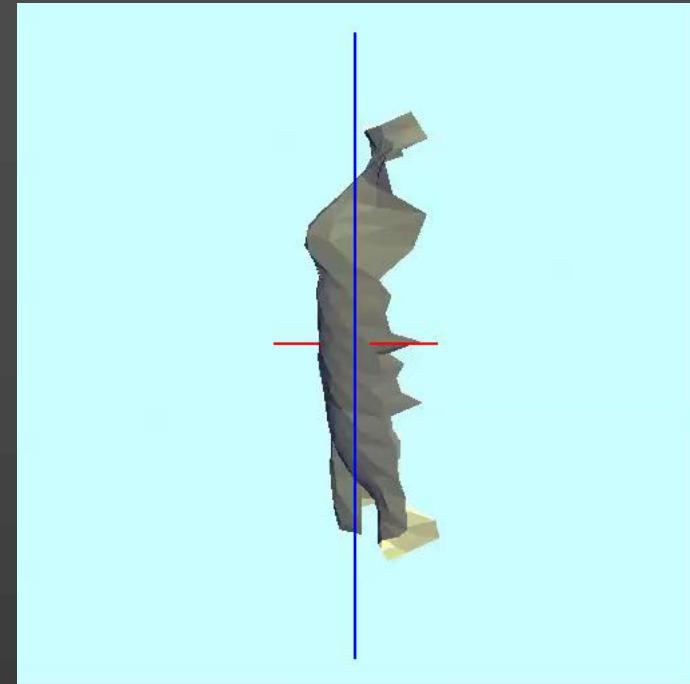
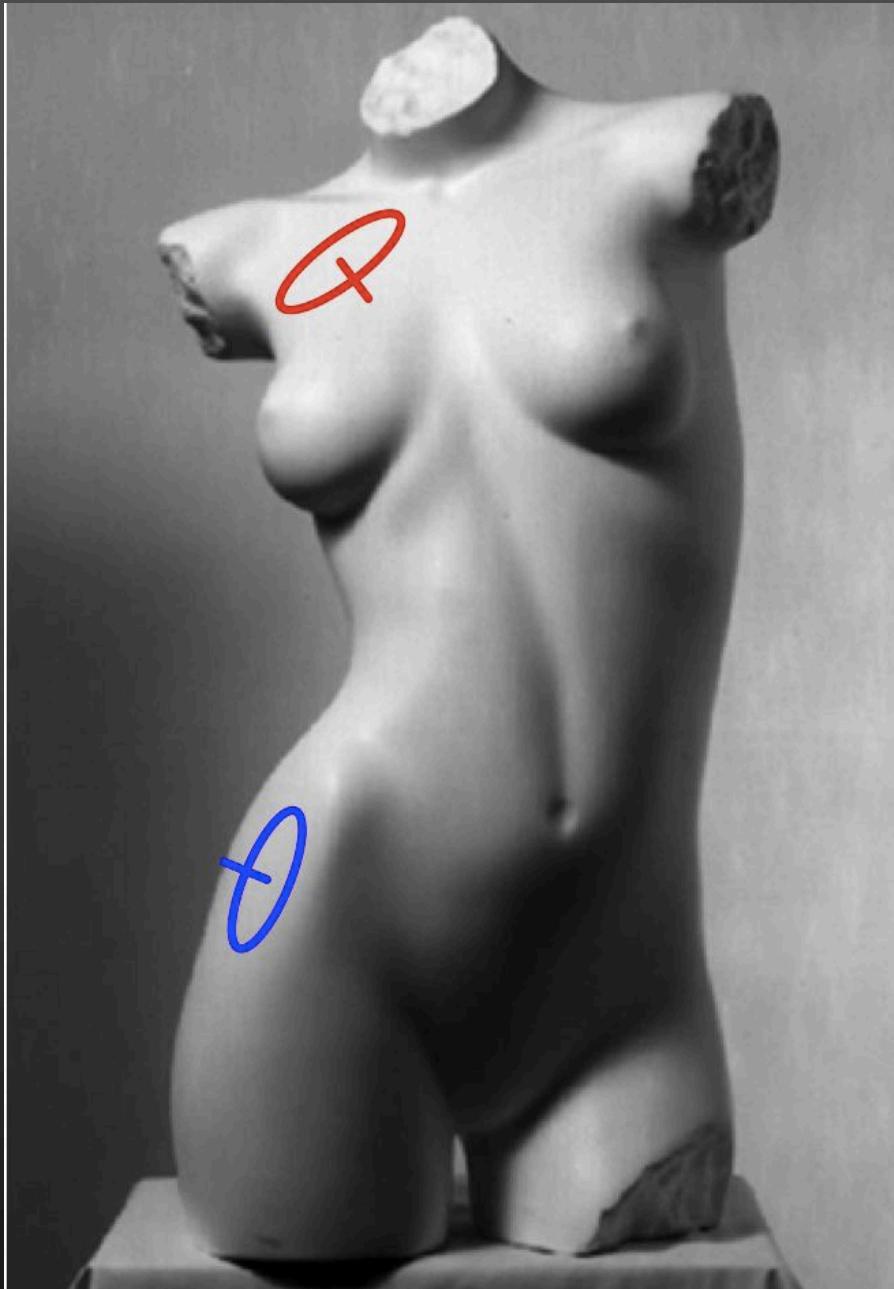


Reconstructed images



Recovered depth map

[K. HE, J. Sun and X. Tang, CVPR 2009]



Depth perception without disparity (Koenderink & Van Doorn, 1995)



<http://go.funpic.hu>

Source: J. Koenderink

Challenges or opportunities?



Image source: J. Koenderink

- Images are confusing, but they also reveal the structure of the world through numerous cues.
- Our job is to interpret the cues!

Vision is more than 3D,
e.g. scene analysis



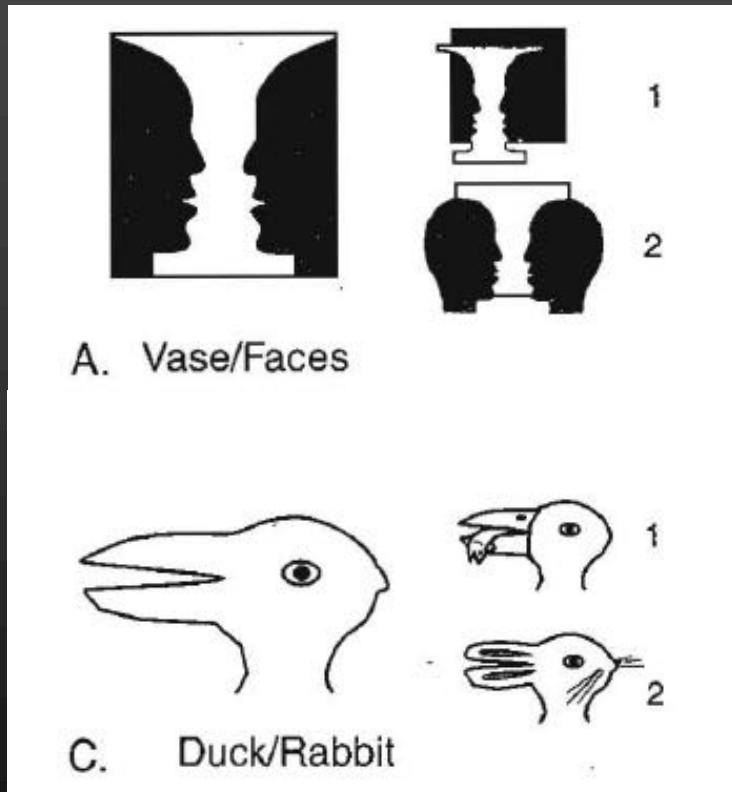
The true output of a digital camera

08122635252121314133210507102023222326333642453122202526231614141320343524131413131282317140918192125272127444233263323191924333640424044352
08133242231010244837260708101827272283487485436312729282318171614406153521815151417261108071121242930253250473431383218192631263139394940322
09112942230507164739281118182026262630364853544234293027211816161533595448171415151414191409070815252529253754473536434024222531243135324540372
0909264139404649503929212928272727283440505557301820262722191716151727292014151515151415201308091225262822345146343640321720253222335334436311
0910234354575245483930141212152828283537475255291518252727322621202018181818191817181717181409101323262822366444273545362223253228332385341401
07091839451831223840321404081328282831333946523516182529263330262523232323262829313128241313120822727223648402735392918222432334043434748441
0508153842142722303731170509122830313034894548711714222524303330292725262726283032353329221217151726302420364738264245312025263124323445453280
0812143338171411263632210509102730312938404240311012192425171319262425293025262728322923161316162331292318324033223740271623253025331465448340
22251628314170112035312005060926282824434036301608101719171204172724223524100813112835140612171823303739303337251843442619262729243331394142270
23261923312528091531292205050922191318272929181413172221191603182425162517050912153538150506192321215771272936342938402217232828213131535042180
18191819281603071123201506051021110318232319161314161618191504111823093114000615122831100404233129222628232933293546412120252927172927424747190
0518241418160609122316141117202215031513151312120912121216120706112214291504021007233310040421193525242221222324334340261717273342933473833100
05132513161614171918161507131619180406131607091515161721201711071031132617020410092734110506111024272927252122203450393741261719364142504837100
09102615151211141515161506111413140710181918182318101728181921111332081410040709093429070508030112231907251826263148403638303228181736363630070
1108241714130912131314141215171915110910101012150804132220191710183511050507080508331304072223080413150621183030333830353252326313340463821050
5722211709141722162221162523202019201413201910050000307140602051313040304060605153614091114171904020205181215042824212632312624233635364022050
48241320091319351719191524272425262922182212030201030306090505081306020303040405121412141004050802090804071111081711142330282929263739363219301
0804102209131235232821090710141006050912110904070809090706080506070709080605050407070306060705060410181004040788510307213433272213237383617363
050404130607083422190803010002020505081311071319141108100712090204081312101111008040708082500050804151905021185670303254644331118263134118362
020508030304071009110400020100000605081722381611040201010203040506091111110110802010207101001070905171303040320120504312923312122425272118372
0805050304050304050400000004040002030711190305030102020303050607080909111011103010102131503091304141102040307200503322131712142725241616262
050201030415140306081312080611140401010001040508040202030405060607070809131011117140500021507011220041410020203061804042912081511202522241416232
04040304041616132019131317251214150302030306050504040404050606060707101918121219261101010901011623070802030203071303052207050506162308151021242
0506081010151705060201030918131319070303040606060504050506071016263020141215382606030200001517020203051417121103051014120706202204101124262
14171714091405021401020305060503121201020406060606050605071223363823111219373316090503060905020406051619150702020304090205231603061424333
10141224091503041004040506030201040302020305060606060605060718102034472917172335320100606070709050505071217060300010101000006210305121335623
13090707081006040303020203080400010303030505060508120508060702022337301817254750250906050612150706060911070003020102020100002060308180870996
1514090704030201010101010804000203040504040506070505040306060203112910201917254443190906060712141208071617080101010202010000000108131162976
2109030201010101000000000001010203040506060606060911110304070903040203071719227231813110909081012110825351606000002020100000000005061525331
08021207000000000000000102030304050505060707080808070817322304082908000406081118263541342517111009060709092942191104000202010102020100000610050
01020503010101010203040406060607080808090907040708180104131100021534242146444437272015110907050707284424150802030100010102020100010101071
0206020403050812070503040506060708090910090909060403020202010516010517130910444945363124191311080506051843232614070401000000000010101000101010
0002091107040711130904061006060707080809101010100905030302020206161513021210044757494135282218151306050513392128241009040302000000000000101000
000003050506040402030205160707070808091010100806020201010318042535020400195453514437292318121609040509293326261209070202010100000001010000
000001040506030000020216140506060607070809100906060302020101270903040101012753494847393123181210150705041533362208130704020101000000010201000
000000010204010000010014090203030404040405060704040202010101130601000000010936373535292116130906080703040916301604130504030101000000000501000

(Courtesy of Ivan Laptev)

Fundamental problem II: Cameras do not measure semantics

→ We need lots of prior knowledge to make meaningful interpretations of an image



Outline

- What computer vision is about
- What this class is about
- A brief history of visual recognition
- A brief recap on geometry

Course outline

1. Instance-level recognition

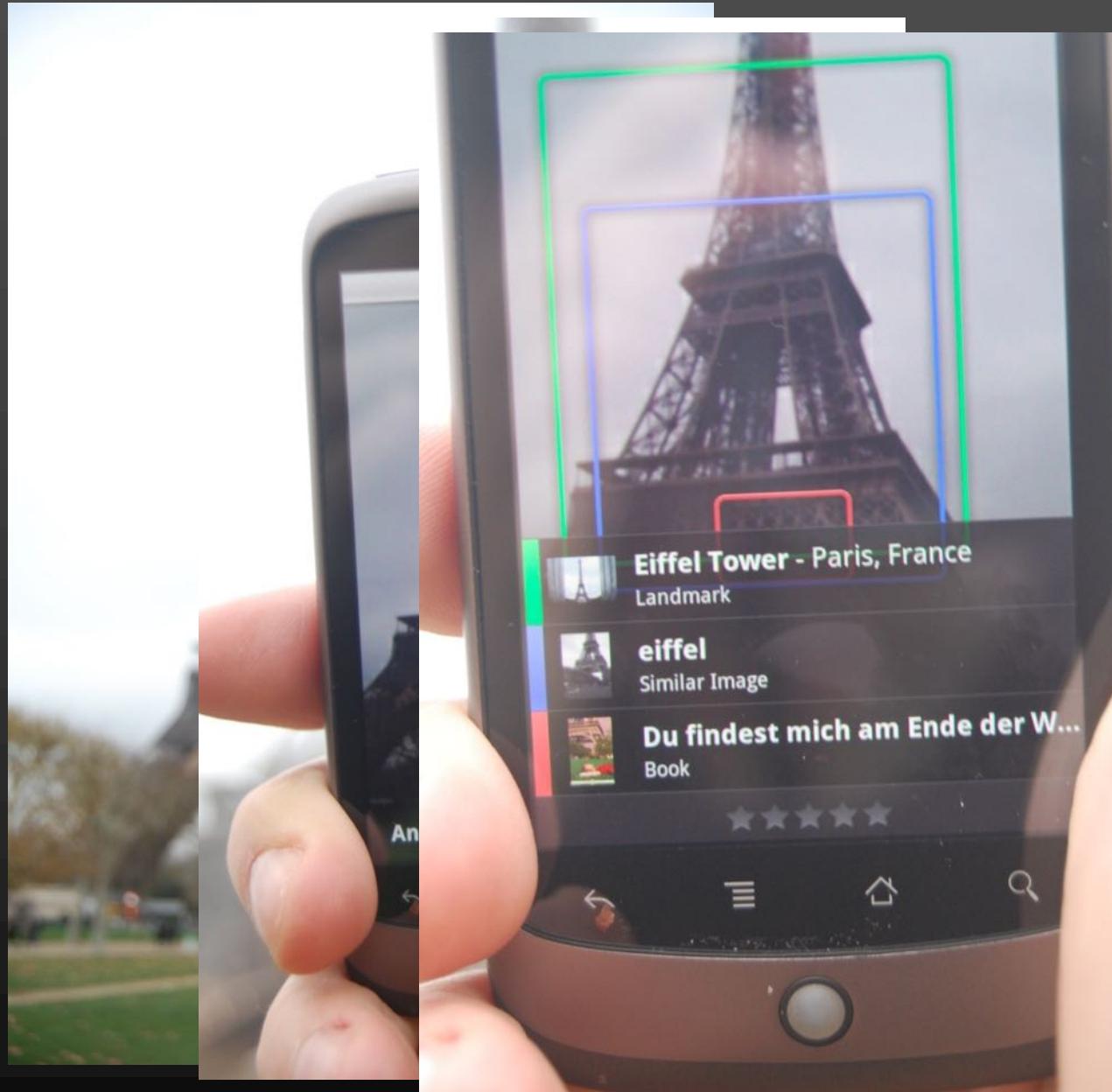
Camera geometry and image alignment
Local invariant features
Large-scale visual search

2. Category-level recognition

Bags-of-features
Sparse coding and dictionary learning
Convolutional neural networks

3. Advanced topics

Motion and human actions
3D object recognition
Weakly-supervised learning



Specific object detection



(Lowe, 2004)

Lecture	Date	Topic and reading materials.	Slides
1	Oct 3	Introduction; Camera geometry (3hrs, J. Sivic)	
2	Oct 10	Instance-level recognition I. - Local invariant features, correspondence, image matching (3hrs, J. Sivic);	
3	Oct 17	Instance-level recognition II. - Efficient visual search (1.5hrs, C. Schmid) Bag-of-feature models for category-level recognition (1.5hrs, C. Schmid)	
4	Oct 24	ICCV 2017. No lecture. Assignments: Assignment 1 due.	
5	Oct 31	Sparse coding and dictionary learning for image analysis (3hrs, J. Ponce)	
6	Nov 7	Neural networks; Optimization methods (3hrs, A. Joulin) Assignments: Assignment 2 due.	
7	Nov 14	Convolutional neural networks for visual recognition I. (I. Laptev) Final project topics are out. Due date for project proposals: Nov 28.	
8	Nov 21	Convolutional neural networks for visual recognition II. (J. Sivic) Assignments: Assignment 3 due.	
9	Nov 28	Motion and human actions I. (C. Schmid) Assignments: Final project proposal due.	
10	Dec 5	Human pose estimation; Weakly-supervised learning I (I. Laptev)	
11	Dec 12	3D object recognition and Convolutional neural networks (M. Aubry) Weakly-supervised learning II (I. Laptev)	
12	Jan 15 Jan 16 Jan 17	Final project presentations and evaluation (I. Laptev, J. Sivic) Jan 15: 13:00-17:00 Jan 16: 13:00-17:00 Jan 17: 13:00-17:00 The presentations will take place at Salle Alan Turing - 1st floor at Inria Paris research center, 2 Rue Simone Iff, 75012, Paris. Directions are here. When you enter the building tell the receptionist you are going for the presentation and go directly to the first floor (no special access card is needed).	

Course outline

1. Instance-level recognition

Camera geometry and image alignment
Local invariant features
Large-scale visual search

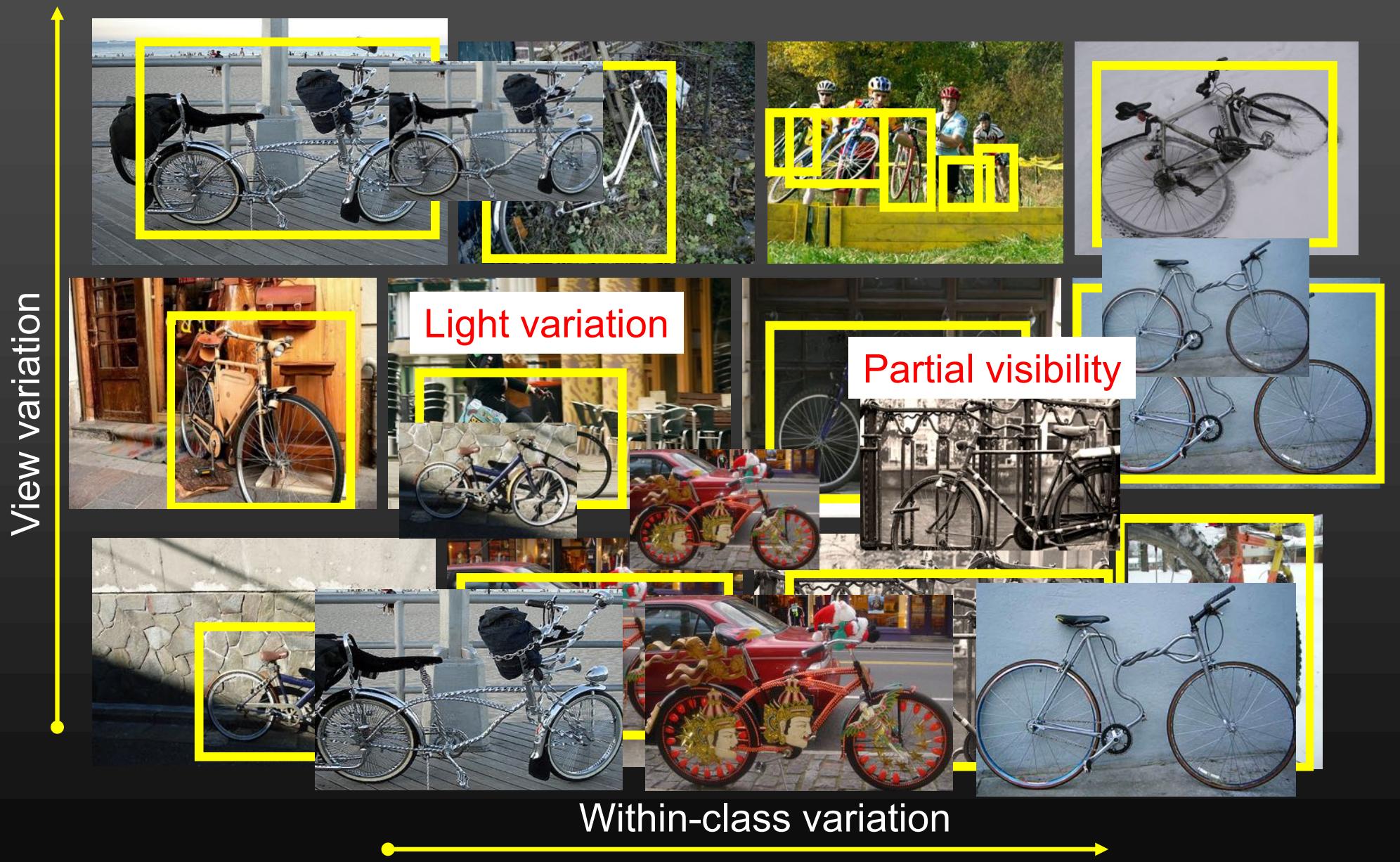
2. Category-level recognition

Bags-of-features
Sparse coding and dictionary learning
Convolutional neural networks

3. Advanced topics

Motion and human actions
3D object recognition
Weakly-supervised learning

Category-level recognition

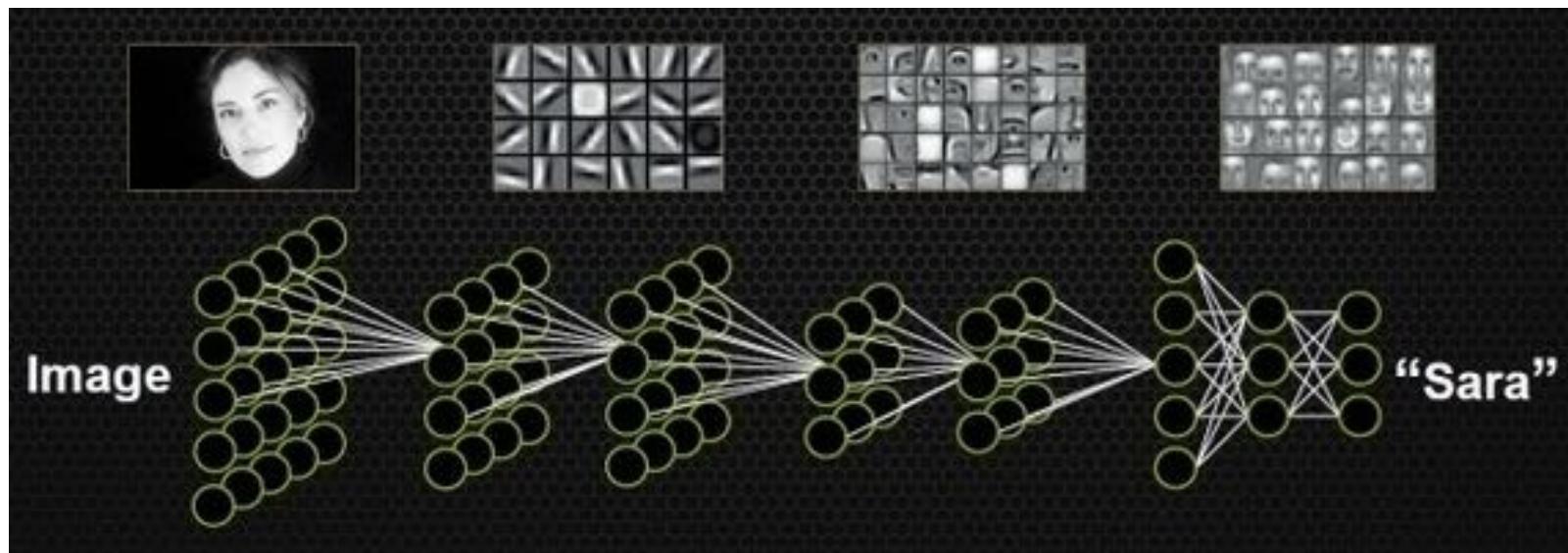


Convolutional neural networks (CNNs)

Multi-layer nested image representation

$$\text{Output representation } z = f^n(\dots f^2(f^1(x))\dots)$$

"Layer" n "Layer" 2 "Layer" 1



Source:
A. Shivkumar

Guest lecture by A. Joulin on neural networks and optimization.

[Rosenblatt'57], [Hubel&Wiesel'59], [Fukushima'80], [Rumelhart'86], [LeCun et al.'89], [LeCun et al.'98], [Hinton&Salakhutdinov'06], [Krizhevsky'12], ...

Lecture	Date	Topic and reading materials.	Slides
1	Oct 3	Introduction; Camera geometry (3hrs, J. Sivic)	
2	Oct 10	Instance-level recognition I. - Local invariant features, correspondence, image matching (3hrs, J. Sivic);	
3	Oct 17	Instance-level recognition II. - Efficient visual search (1.5hrs, C. Schmid) Bag-of-feature models for category-level recognition (1.5hrs, C. Schmid)	
4	Oct 24	ICCV 2017. No lecture. Assignments: Assignment 1 due.	
5	Oct 31	Sparse coding and dictionary learning for image analysis (3hrs, J. Ponce)	
6	Nov 7	Neural networks; Optimization methods (3hrs, A. Joulin) Assignments: Assignment 2 due.	
7	Nov 14	Convolutional neural networks for visual recognition I. (I. Laptev) Final project topics are out. Due date for project proposals: Nov 28.	
8	Nov 21	Convolutional neural networks for visual recognition II. (J. Sivic) Assignments: Assignment 3 due.	
9	Nov 28	Motion and human actions I. (C. Schmid) Assignments: Final project proposal due.	
10	Dec 5	Human pose estimation; Weakly-supervised learning I (I. Laptev)	
11	Dec 12	3D object recognition and Convolutional neural networks (M. Aubry) Weakly-supervised learning II (I. Laptev)	
12	Jan 15 Jan 16 Jan 17	Final project presentations and evaluation (I. Laptev, J. Sivic) Jan 15: 13:00-17:00 Jan 16: 13:00-17:00 Jan 17: 13:00-17:00 The presentations will take place at Salle Alan Turing - 1st floor at Inria Paris research center, 2 Rue Simone Iff, 75012, Paris. Directions are here. When you enter the building tell the receptionist you are going for the presentation and go	

Course outline

1. Instance-level recognition

- Camera geometry and image alignment
- Local invariant features
- Large-scale visual search

2. Category-level recognition

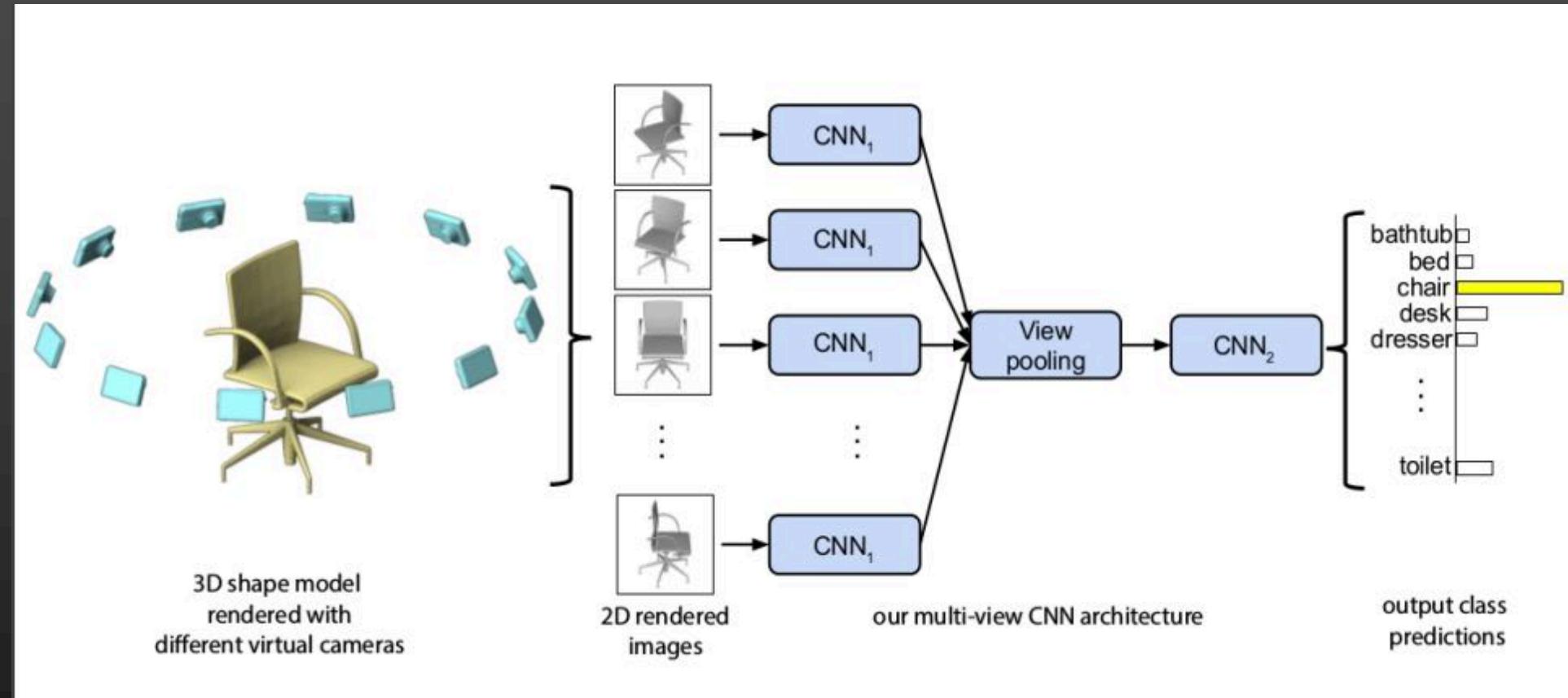
- Bags-of-features
- Sparse coding and dictionary learning
- Convolutional neural networks

3. Advanced topics

- Motion and human actions
- 3D object recognition
- Weakly-supervised learning

Learning from Video and Text via Large-Scale Discriminative Clustering
ICCV2017 submission #1353
Supplementary Material

3D object recognition



[Su et al., ICCV'15]

Weakly supervised learning: learning from **incomplete** and **noisy** meta-data



“Public bikes in Warsaw during night”

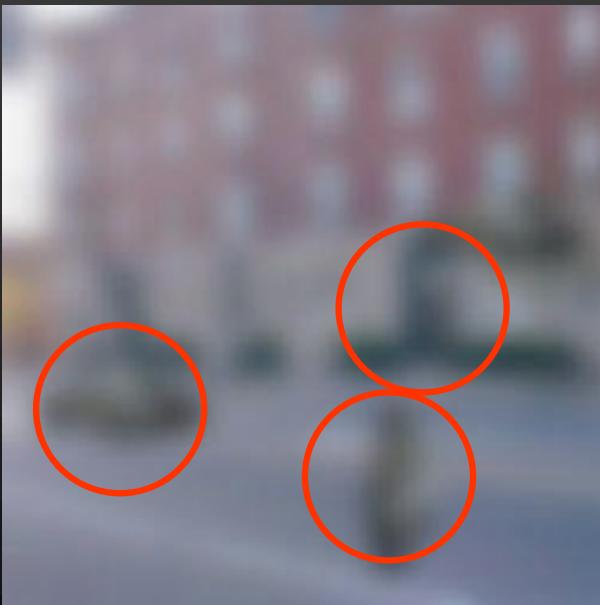
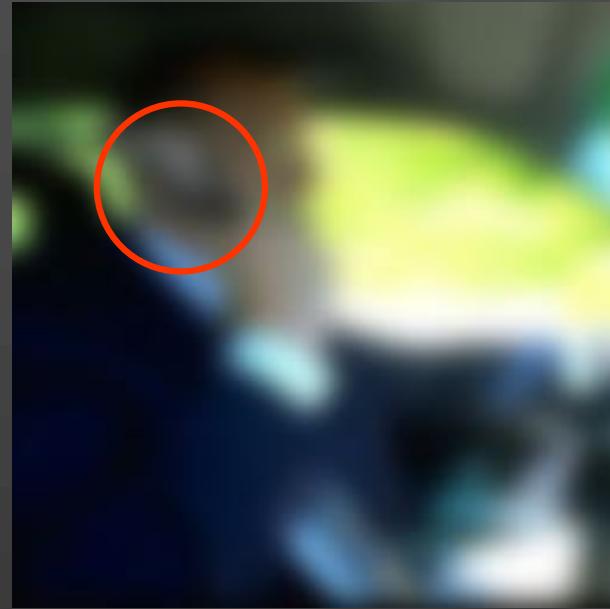
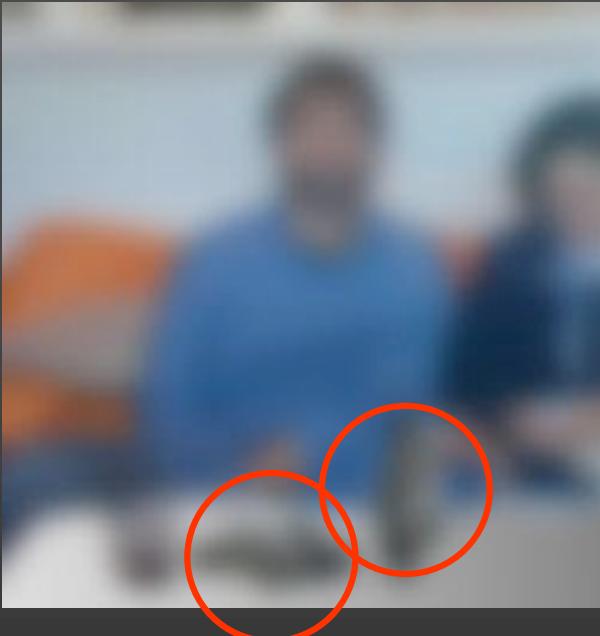
[Barnard et al.'03], [Berg et al.'04], [Gupta and Davis'08], [Ordonez et al.'11], [Kulkarni et al.'11], [Karpathy et al.'14], ...

Towards Scene understanding



Photo courtesy A. Efros.

Local ambiguity and global scene interpretation



slide credit: Fei-Fei, Fergus & Torralba

Lecture	Date	Topic and reading materials.	Slides
1	Oct 3	Introduction; Camera geometry (3hrs, J. Sivic)	
2	Oct 10	Instance-level recognition I. - Local invariant features, correspondence, image matching (3hrs, J. Sivic);	
3	Oct 17	Instance-level recognition II. - Efficient visual search (1.5hrs, C. Schmid) Bag-of-feature models for category-level recognition (1.5hrs, C. Schmid)	
4	Oct 24	ICCV 2017. No lecture. Assignments: Assignment 1 due.	
5	Oct 31	Sparse coding and dictionary learning for image analysis (3hrs, J. Ponce)	
6	Nov 7	Neural networks; Optimization methods (3hrs, A. Joulin) Assignments: Assignment 2 due.	
7	Nov 14	Convolutional neural networks for visual recognition I. (I. Laptev) Final project topics are out. Due date for project proposals: Nov 28.	
8	Nov 21	Convolutional neural networks for visual recognition II. (J. Sivic) Assignments: Assignment 3 due.	
9	Nov 28	Motion and human actions I. (C. Schmid) Assignments: Final project proposal due.	
10	Dec 5	Human pose estimation; Weakly-supervised learning I (I. Laptev)	
11	Dec 12	3D object recognition and Convolutional neural networks (M. Aubry) Weakly-supervised learning II (I. Laptev)	
12	Jan 15 Jan 16 Jan 17	Final project presentations and evaluation (I. Laptev, J. Sivic) Jan 15: 13:00-17:00 Jan 16: 13:00-17:00 Jan 17: 13:00-17:00 The presentations will take place at Salle Alan Turing - 1st floor at Inria Paris research center, 2 Rue Simone Iff, 75012, Paris. Directions are here. When you enter the building tell the receptionist you are going for the presentation and go	

Books for this class

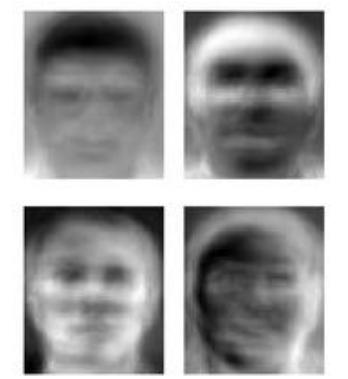
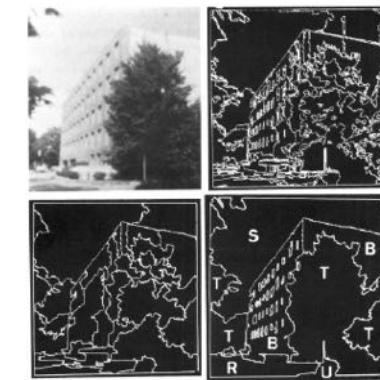
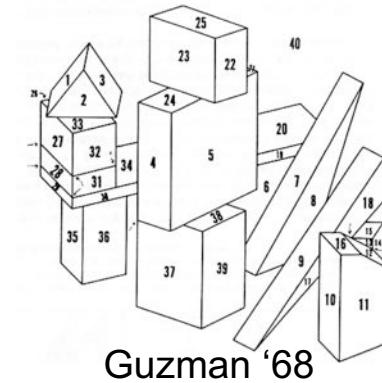
- D.A. Forsyth and J. Ponce, "Computer Vision: A Modern Approach", Prentice-Hall/Pearson, 2nd edition, 2011.
- J. Ponce, M. Hebert, C. Schmid, and A. Zisserman, "Toward category-level object recognition", Springer LNCS, 2007.
- R. Szeliski, "Computer Vision: Algorithms and Applications", Springer, 2010.
- C. Bishop: Pattern Recognition and Machine Learning, 2006
- R. Hartley and A. Zisserman, "Multiple View Geometry in Computer Vision", Cambridge University Press, 2004.
- J.J. Koenderink, "Solid Shape", MIT Press, 1990.
- J.J. Koenderink,
<http://www.gestaltrevision.be/en/resources/clootcrans-press>

Outline

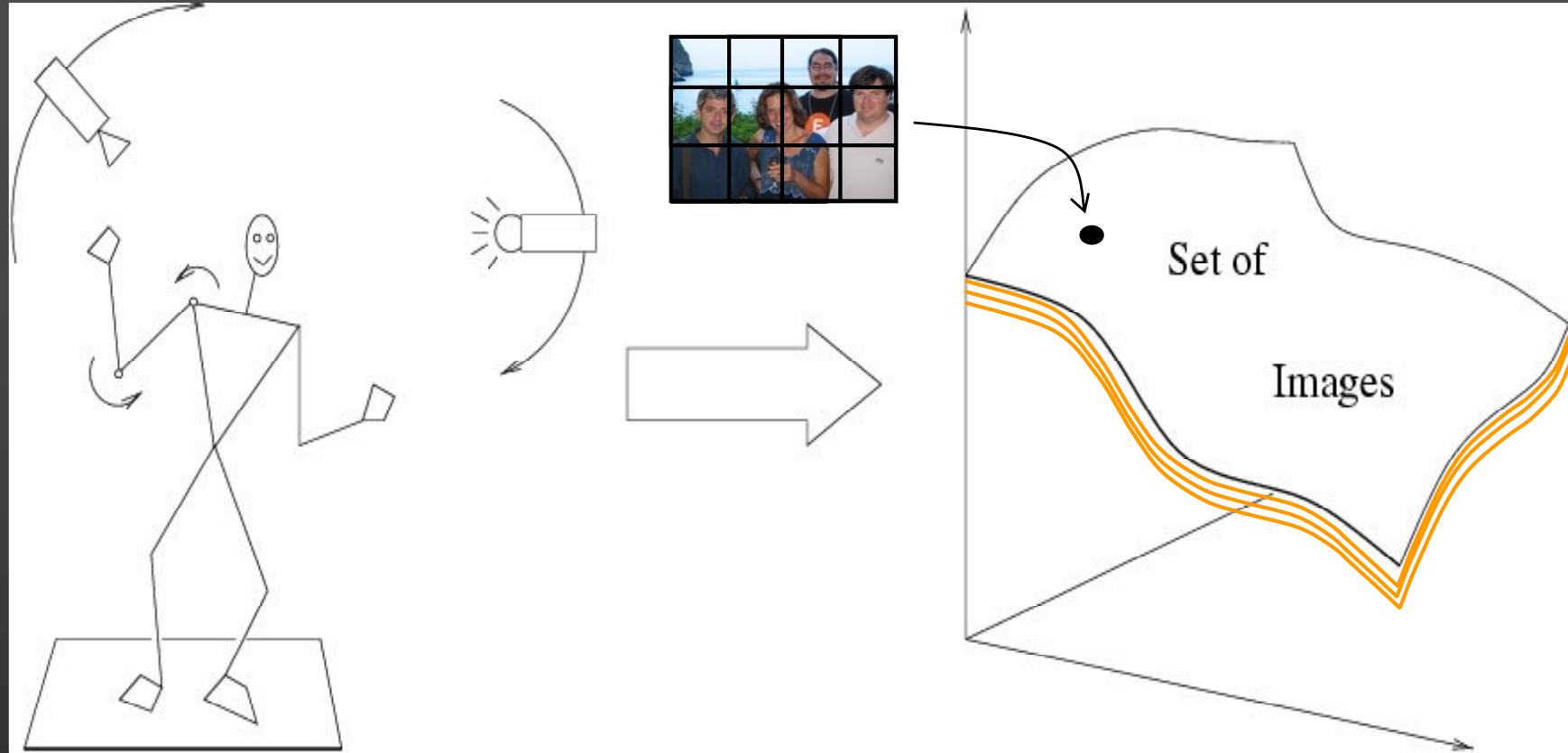
- What computer vision is about
- What this class is about
- A brief history of visual recognition
- A brief recap on geometry

A very brief history of computer vision

- 1966: Minsky assigns computer vision as an undergrad summer project
- 1960's: interpretation of synthetic worlds
- 1970's: some progress on interpreting selected images
- 1980's: ANNs come and go; shift toward geometry and increased mathematical rigor
- 1990's: face recognition; statistical analysis
- 2000's: broader recognition; large annotated datasets available; video processing starts
- 2010's: Deep learning with ConvNets
- 2030's: ...

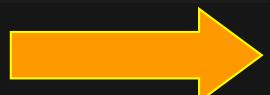


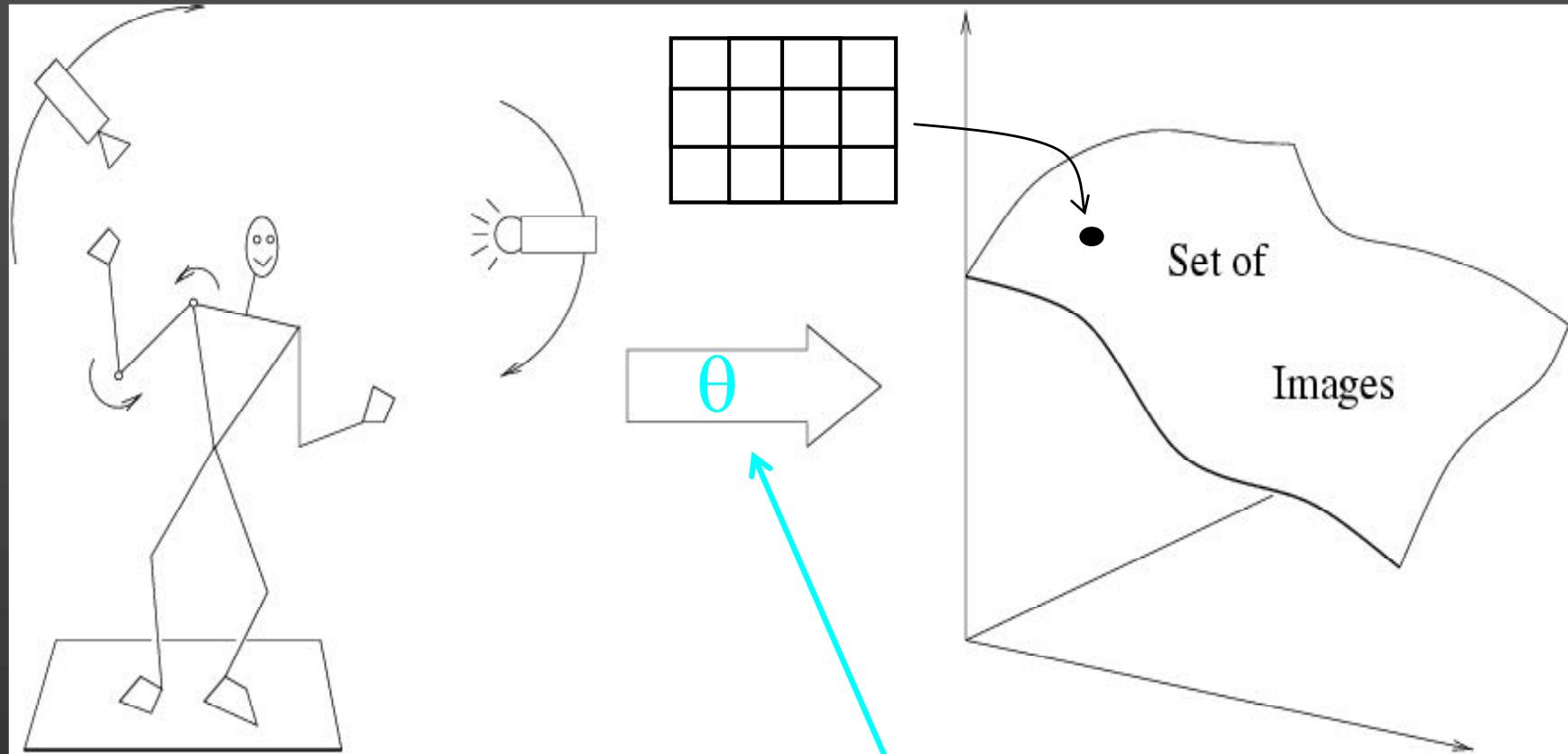
Turk and Pentland '91



Variability:

Camera position
Illumination
Internal parameters
Within-class variations



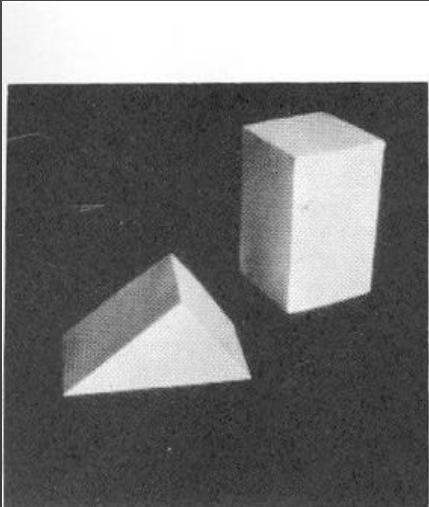


Variability:

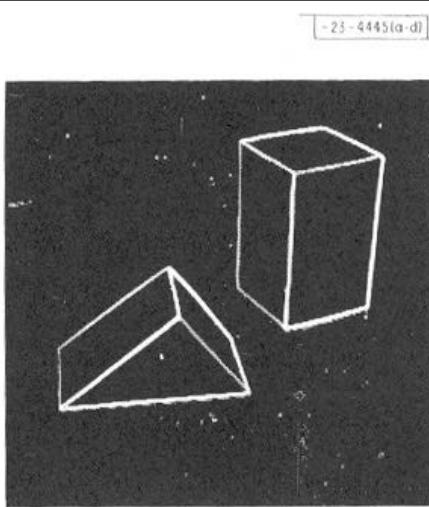
Camera position
Illumination
Internal parameters

Roberts (1963); Lowe (1987); Faugeras & Hebert (1986); Grimson & Lozano-Perez (1986); Huttenlocher & Ullman (1987)

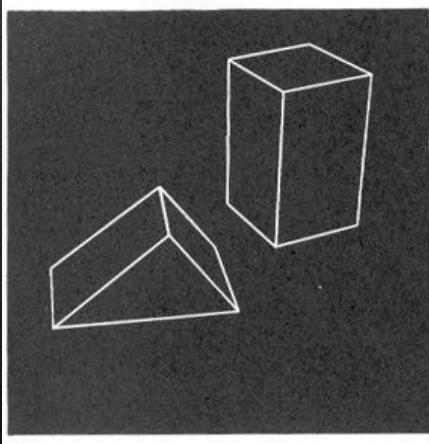
Origins of computer vision



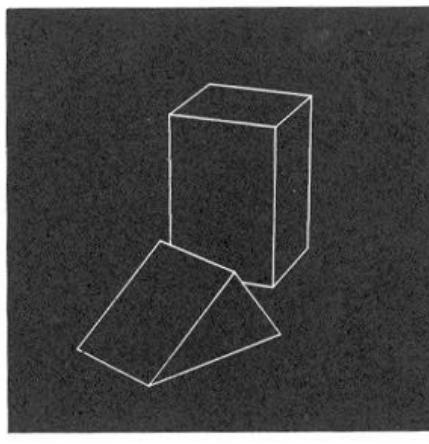
(a) Original picture.



(b) Differentiated picture.



(c) Line drawing.



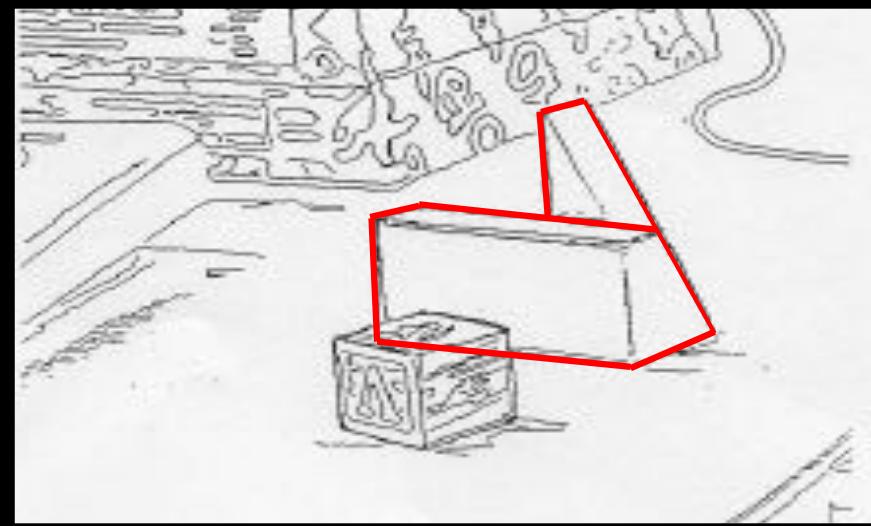
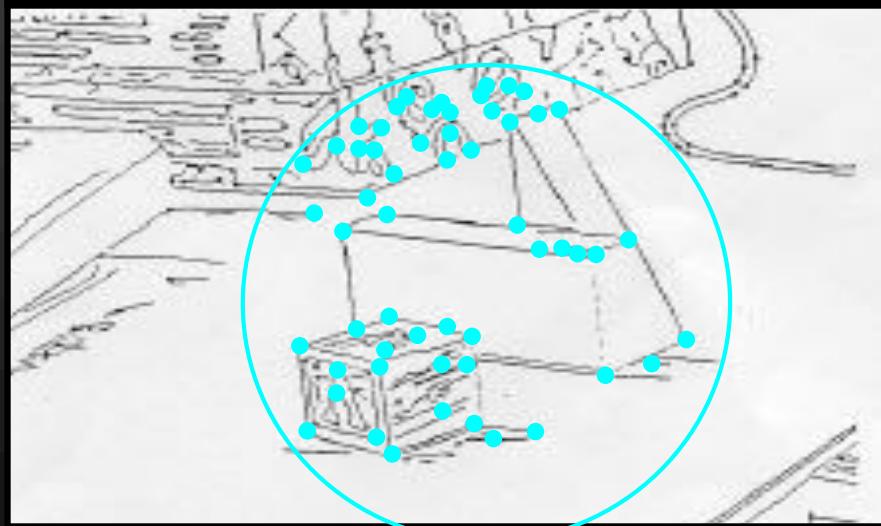
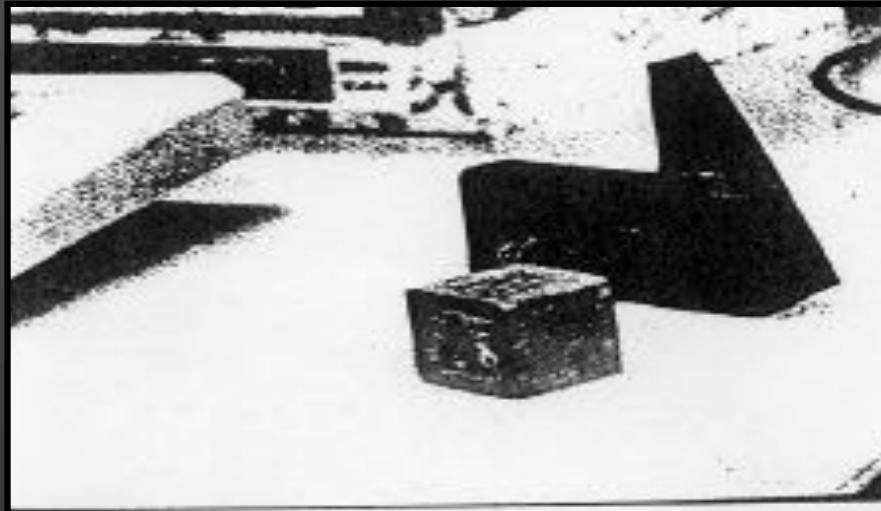
(d) Rotated view.

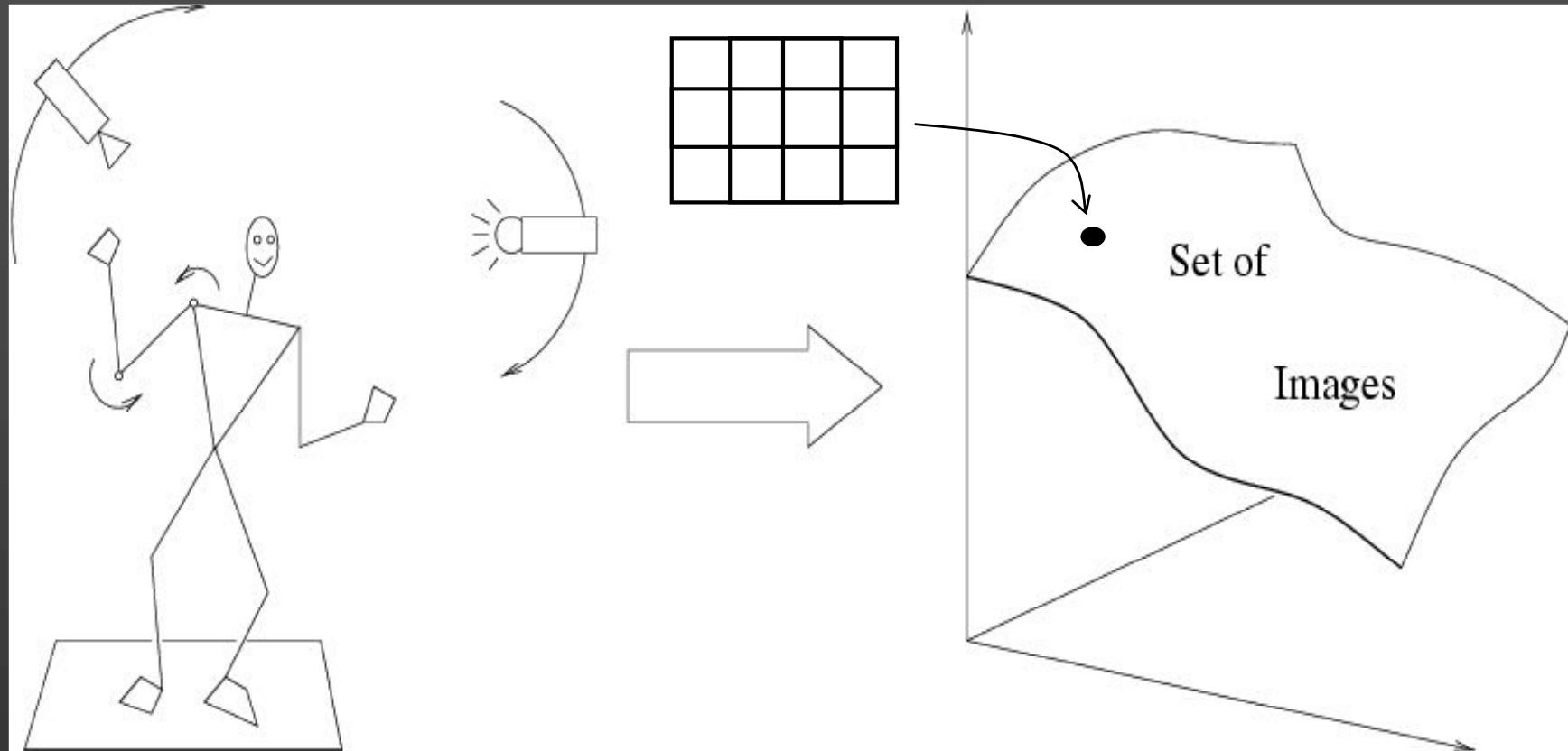


L. G. Roberts, *Machine Perception of Three Dimensional Solids*, Ph.D. thesis, MIT Department of Electrical Engineering, 1963.

photo credit: Joe Mundy

Huttenlocher & Ullman (1987)





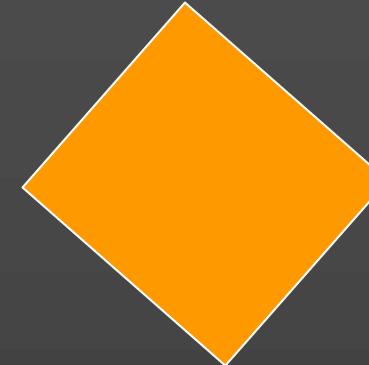
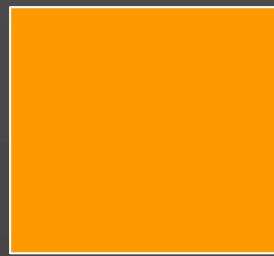
~~Variability~~

Invariance to:

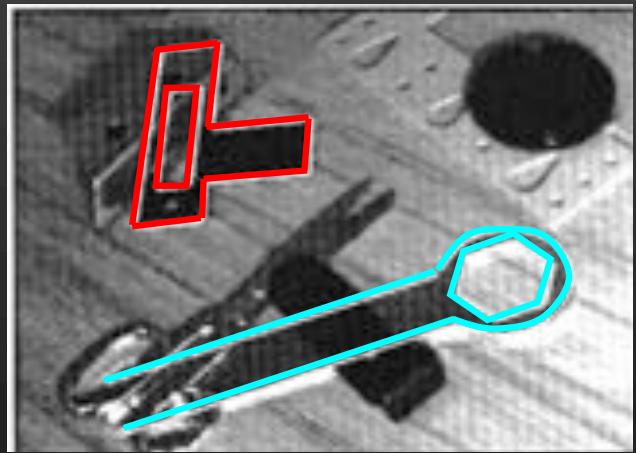
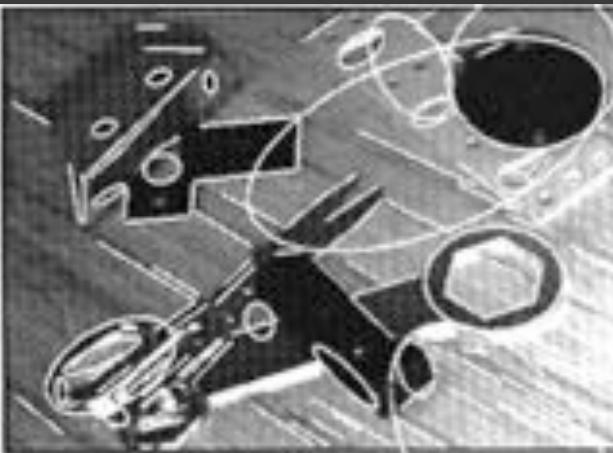
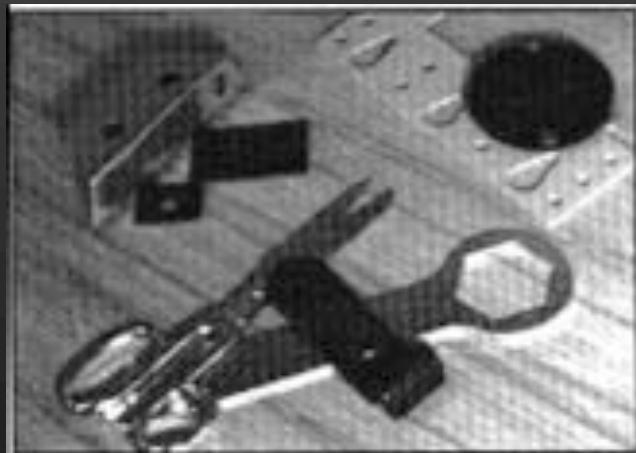
Camera position
Illumination
Internal parameters

Duda & Hart (1972); Weiss (1987); Mundy et al. (1992-94);
Rothwell et al. (1992); Burns et al. (1993)

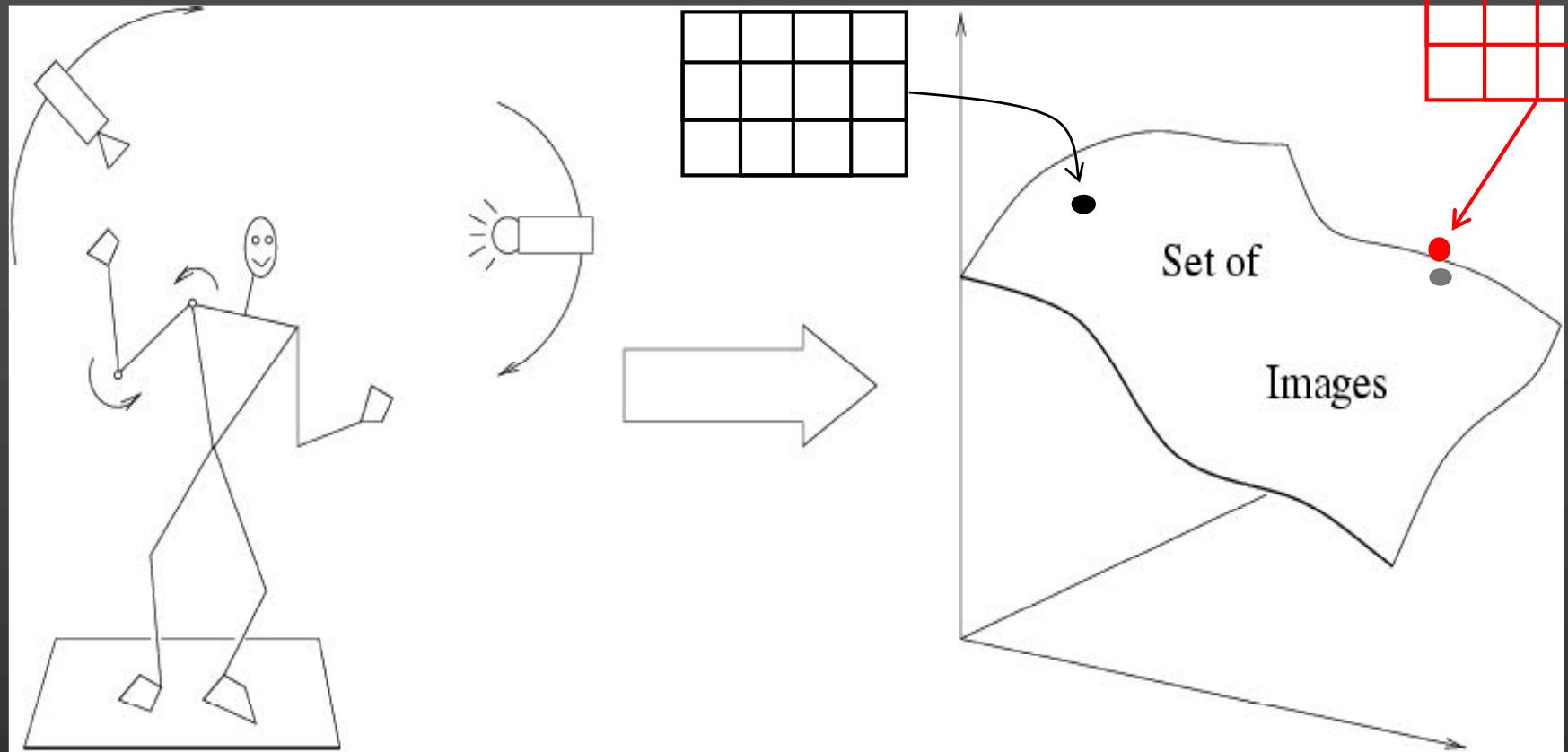
Example: rotation invariants



Projective invariants (Rothwell et al., 1992):



BUT: True 3D objects do not admit monocular viewpoint invariants (Burns et al., 1993) !!



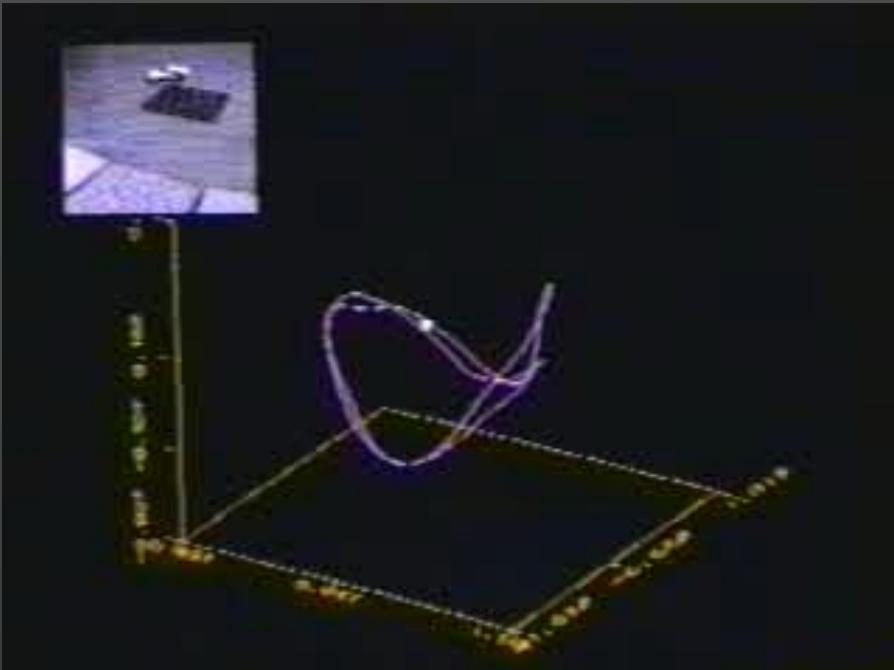
Empirical models of image variability: Appearance-based techniques

Turk & Pentland (1991); Murase & Nayar (1995); etc.

Eigenfaces (Turk & Pentland, 1991)



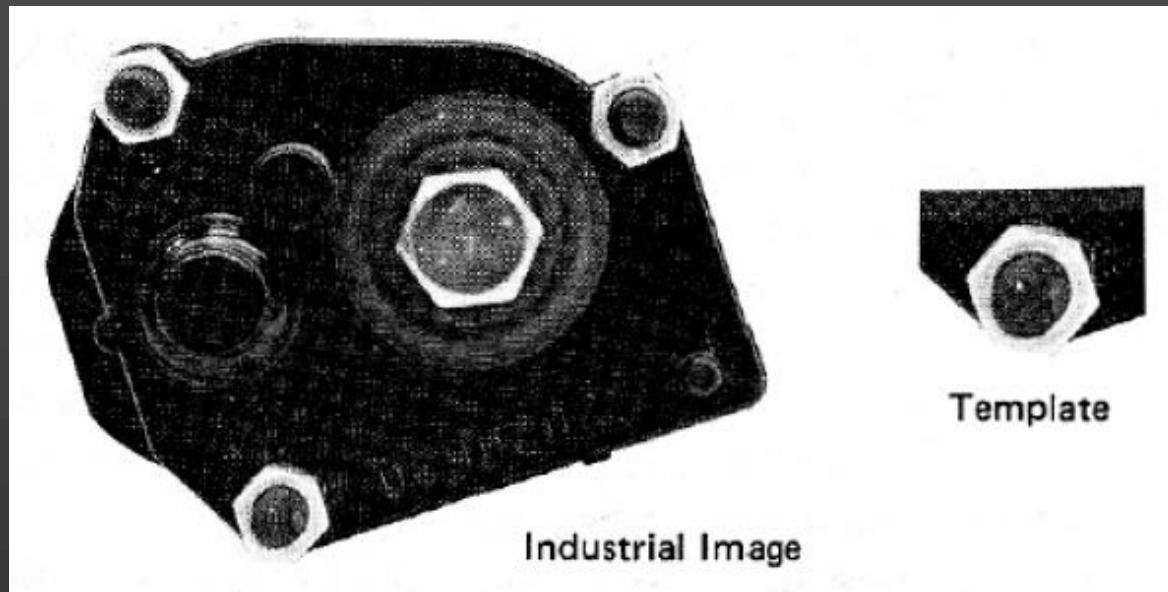
Experimental Condition	Correct/Unknown Recognition Percentage		
Condition	Lighting	Orientation	Scale
Forced classification	96/0	85/0	64/0
Forced 100% accuracy	100/19	100/39	100/60
Forced 20% unknown rate	100/20	94/20	74/20



Appearance manifolds
(Murase & Nayar, 1995)



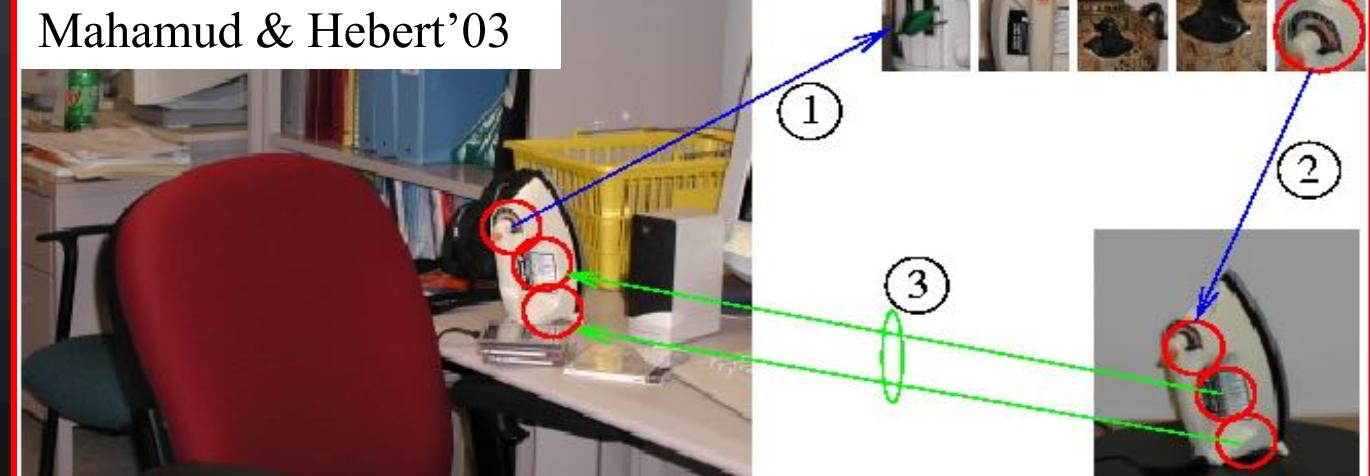
Correlation-based template matching (60s)



Ballard & Brown (1980, Fig. 3.3). Courtesy Bob Fisher and Ballard & Brown on-line.

- Automated target recognition
- Industrial inspection
- Optical character recognition
- Stereo matching
- Pattern recognition

In the late 1990s, a new approach emerges:
Combining local appearance, spatial constraints, invariants,
and classification techniques from machine learning.



Late 1990s: Local appearance models



(Image courtesy of C. Schmid)

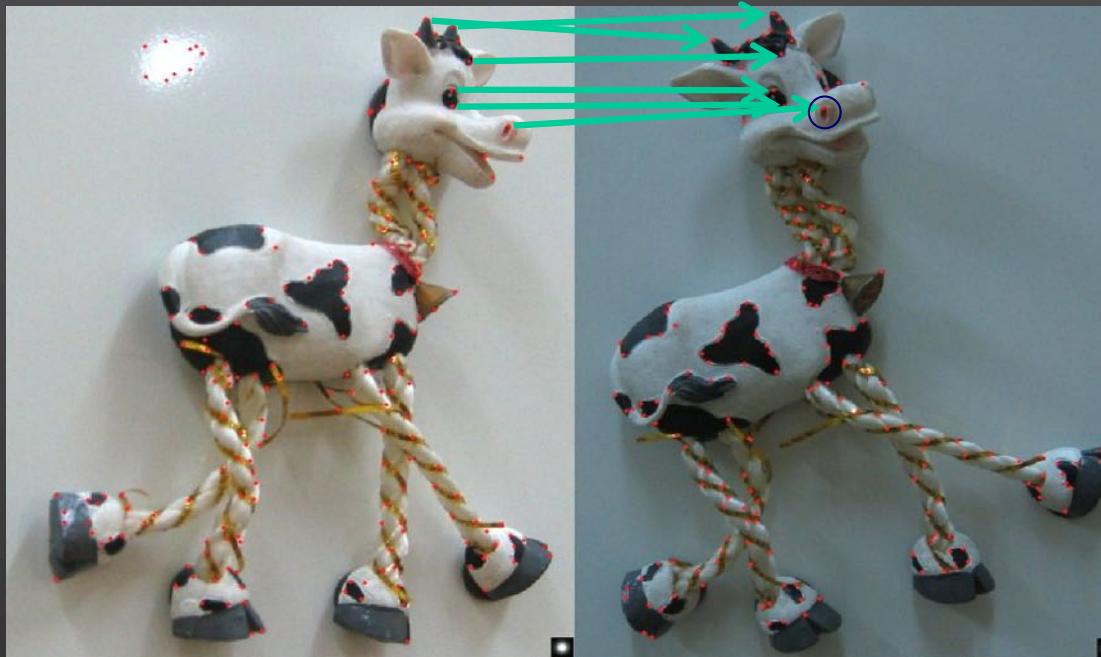
Late 1990s: Local appearance models



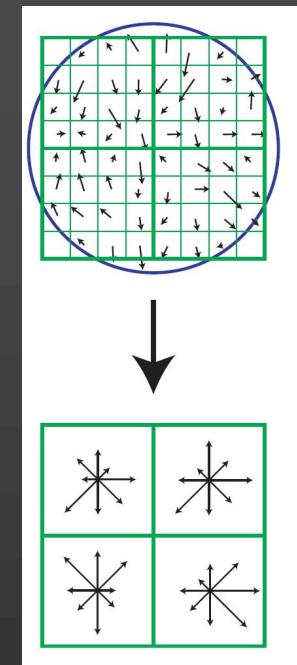
(Image courtesy of C. Schmid)

- Find features (interest points).

Late 1990s: Local appearance models



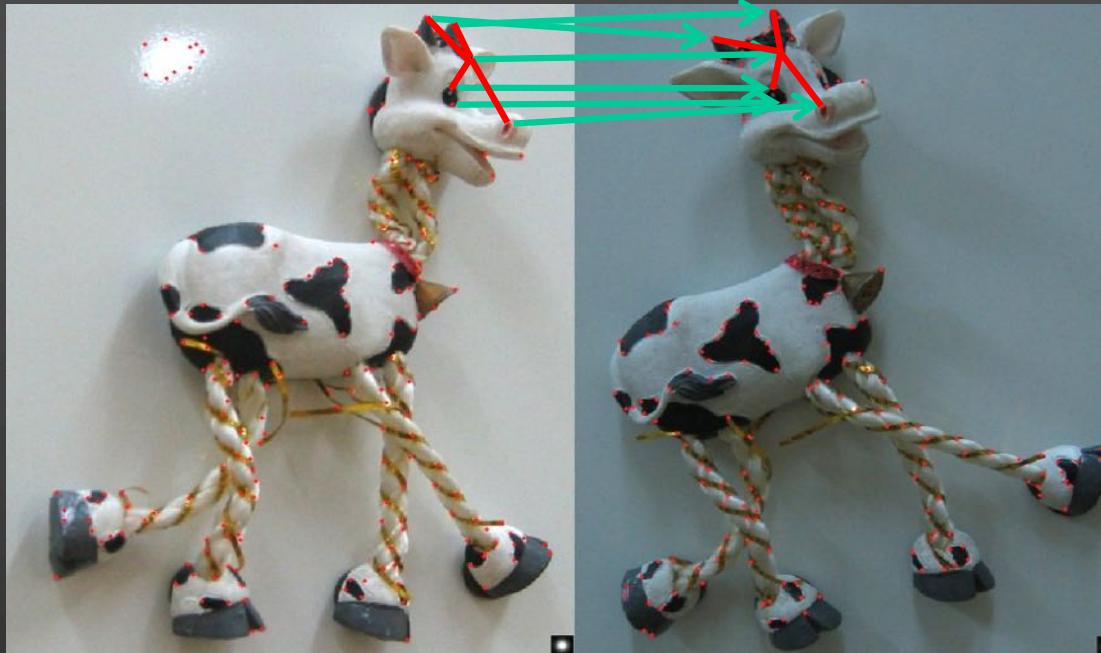
(Image courtesy of C. Schmid)



(Lowe 2004)

- Find features (interest points).
- Match them using local invariant descriptors (jets, SIFT).

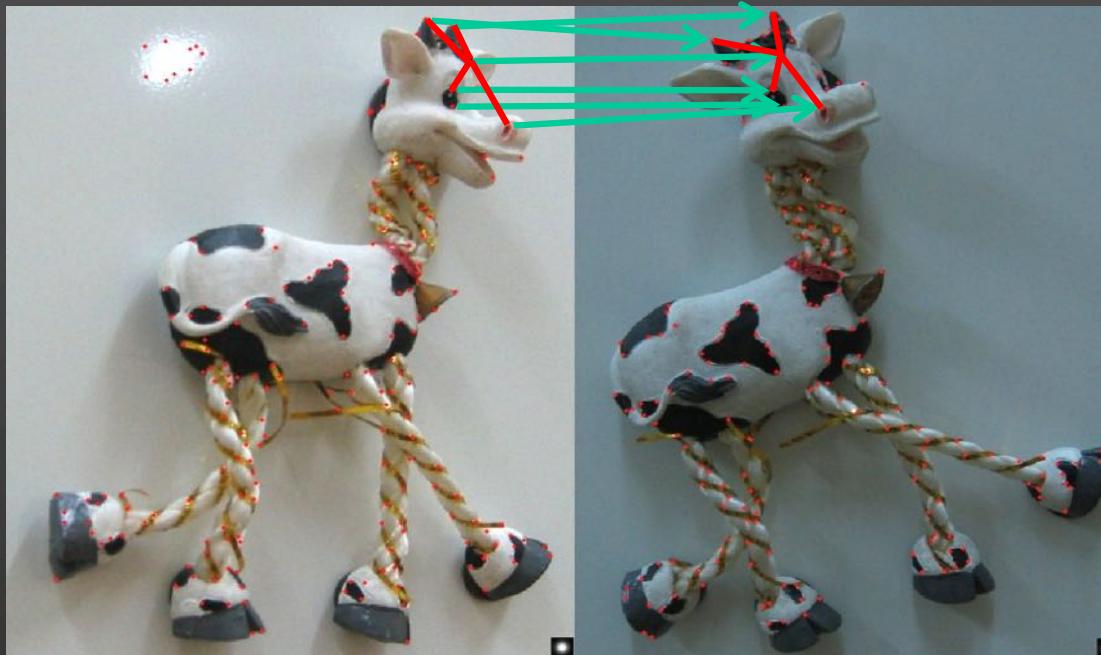
Late 1990s: Local appearance models



(Image courtesy of C. Schmid)

- Find features (interest points).
- Match them using local invariant descriptors (jets, SIFT).
- Optional: Filter out outliers using geometric consistency.

Late 1990s: Local appearance models



(Image courtesy of C. Schmid)

- Find features (interest points).
- Match them using local invariant descriptors (jets, SIFT).
- Optional: Filter out outliers using geometric consistency.
- Vote.

See, for example, Schmid & Mohr (1996); Lowe (1999); Tuytelaars & Van Gool, (2002); Rothganger et al. (2003); Ferrari et al., (2004).

Bags of words: Visual “Google” (Sivic & Zisserman, ICCV’ 03)

“Visual word” clusters

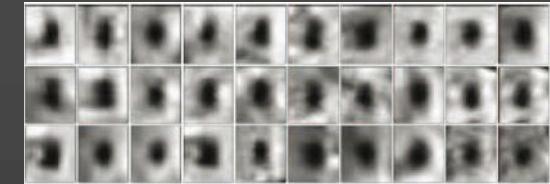
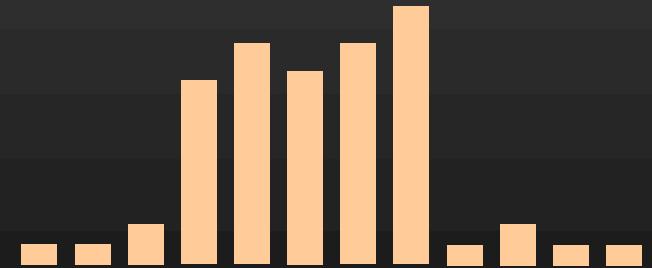


Image retrieval in videos



Vector quantization into histogram
(the “bag of words”)

Bags of words: Visual “Google”

(Sivic & Zisserman, ICCV' 03)

Retrieved shots



Select a region

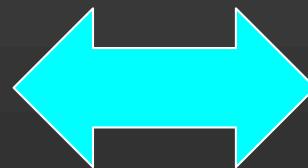
Shots Keyframes

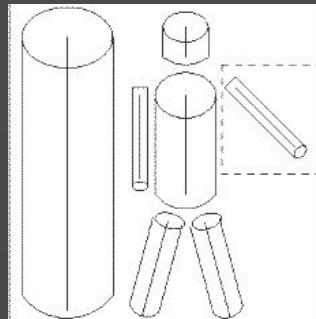
Select a region and click on Search to search for an object:

A screenshot of a web-based visual search application. It shows a frame from a movie scene with a man eating and a woman looking on. A yellow rectangular box highlights a decorative plate hanging on the wall behind them. Below the image is a green header bar with the text "Select a region and click on Search to search for an object:" and two buttons, "Delete" and "Submit".



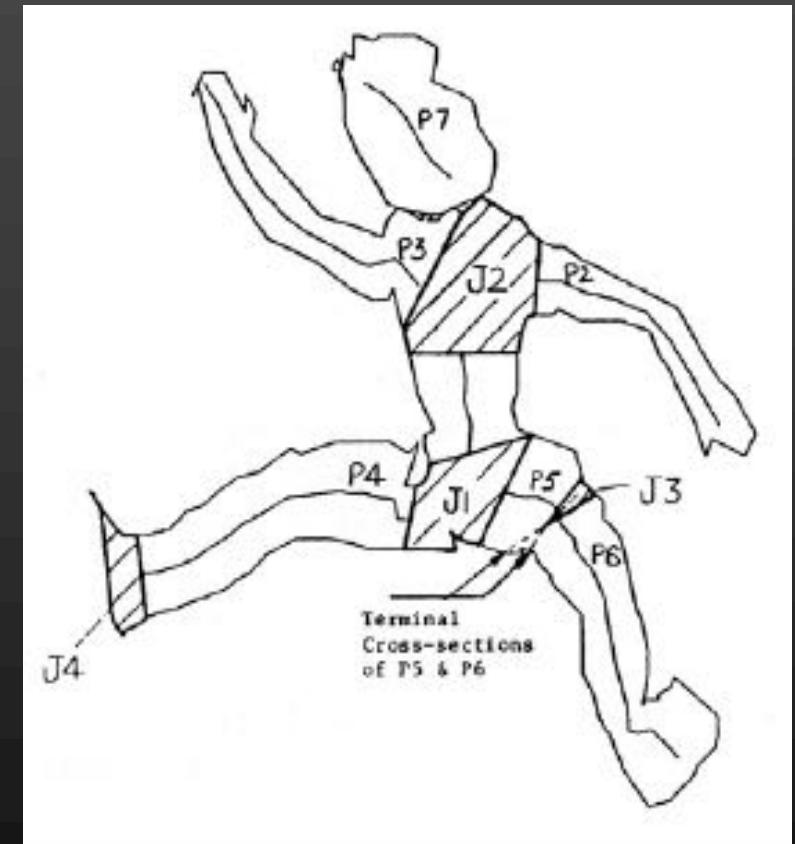
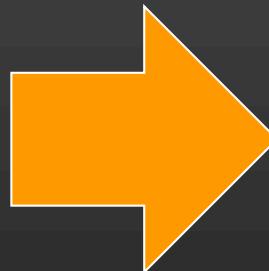
Image categorization is harder





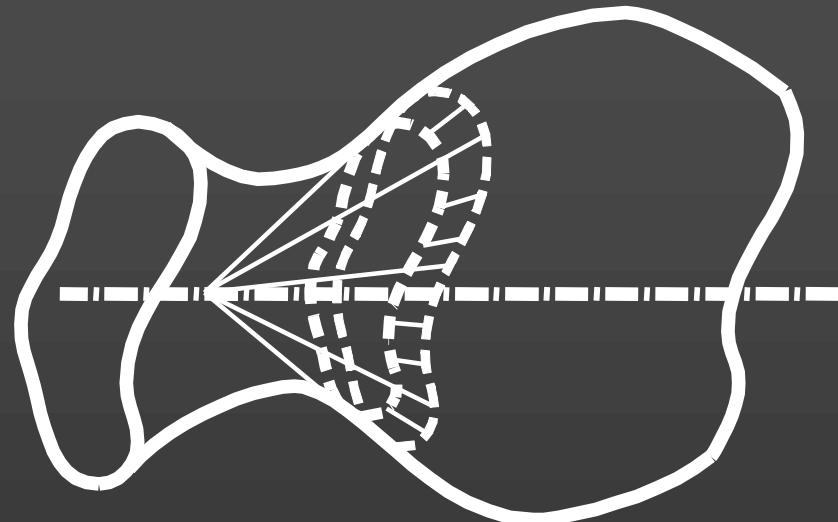
Structural part-based models

(Binford, 1971; Marr & Nishihara, 1978)

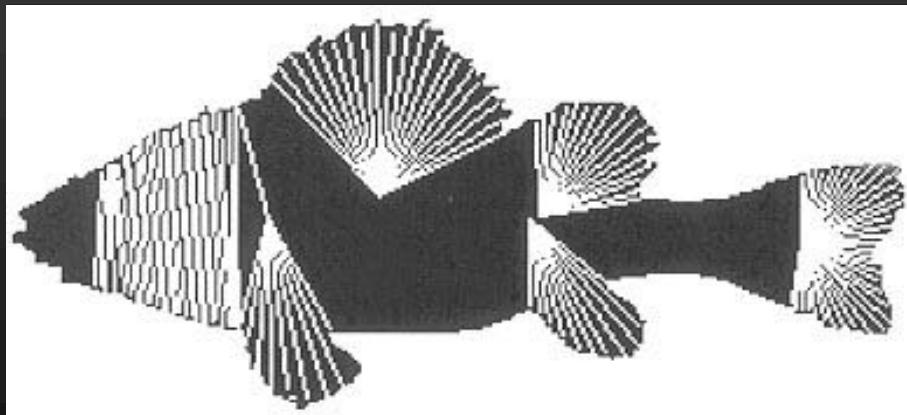


(Nevatia & Binford, 1972)

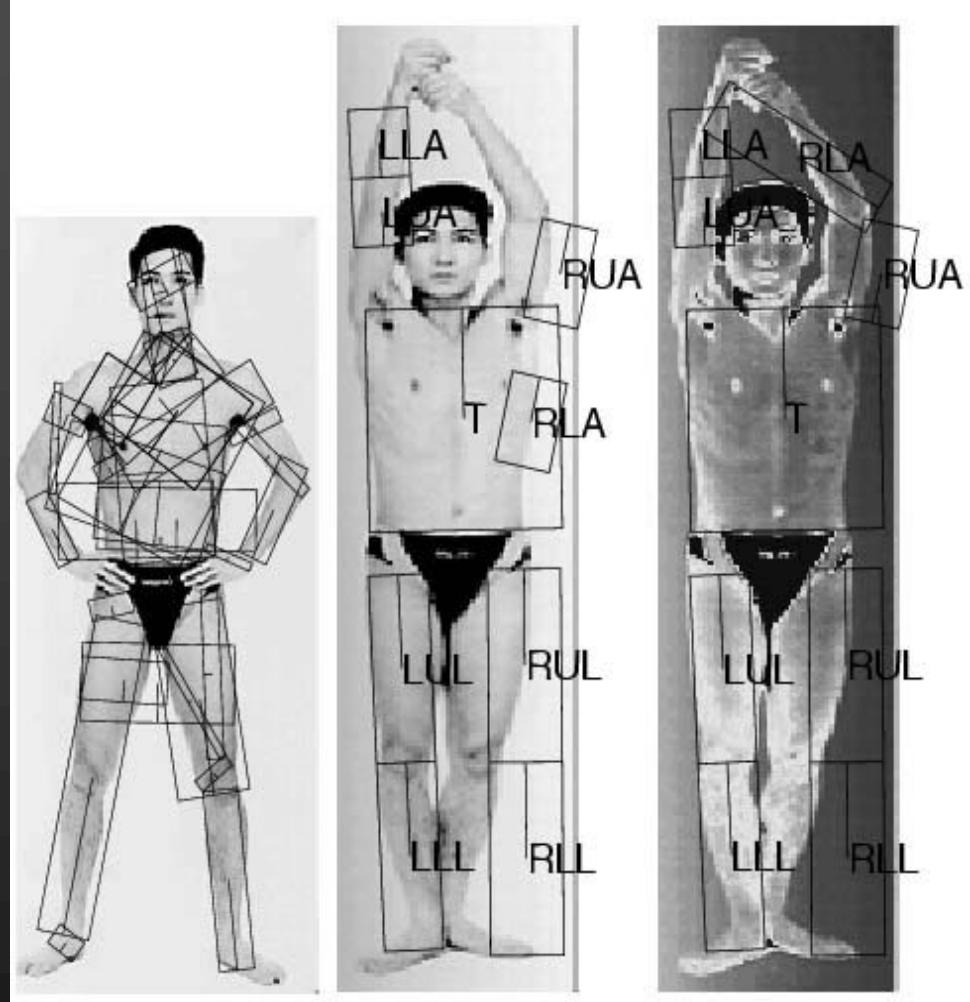
This is hard to operationalize



Ponce et al. (1989)

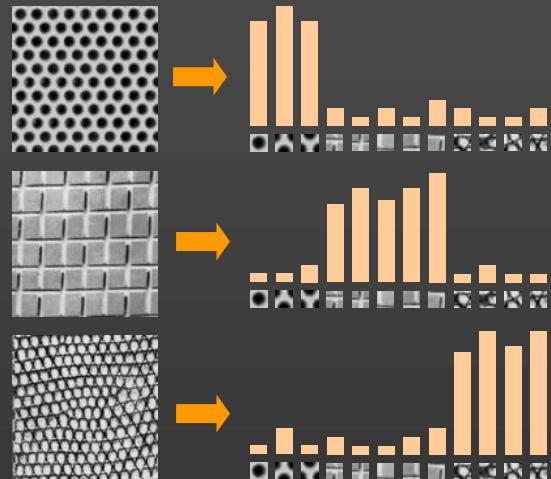


Zhu and Yuille (1996)



Ioffe and Forsyth (2000)

Bags of words and their variants have become the dominant model for image categorization



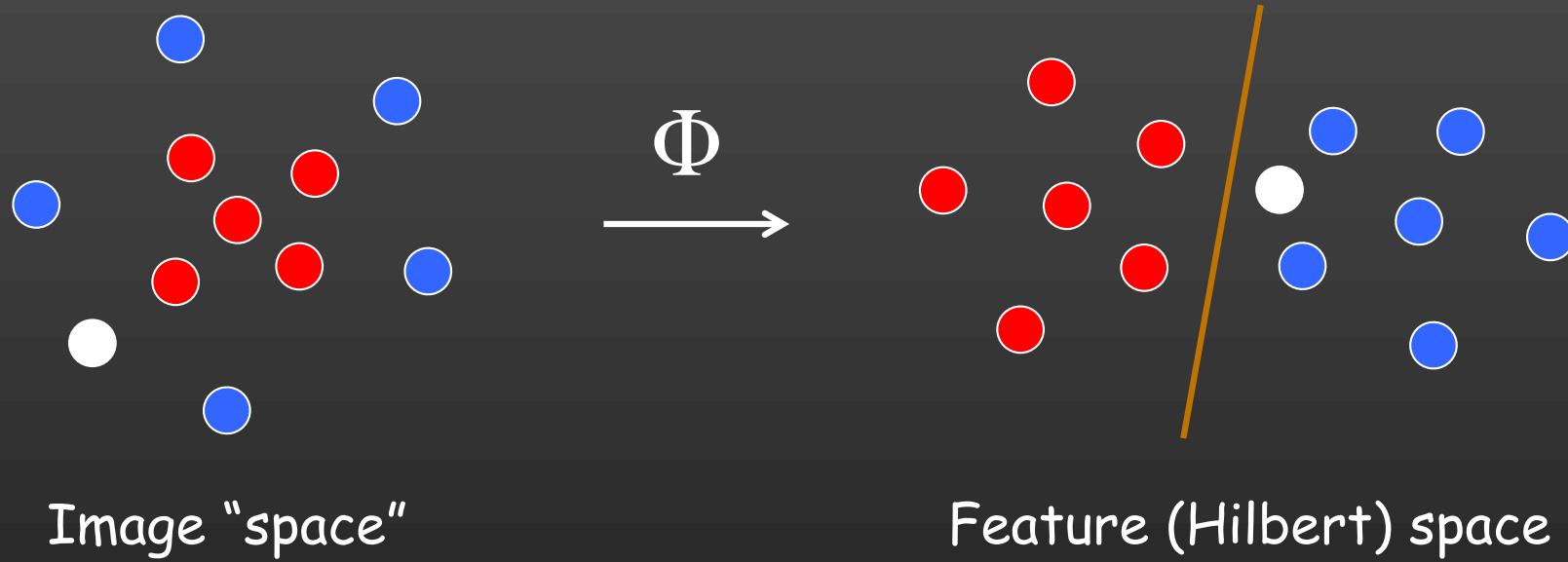
(Swain & Ballard'91; Lazebnik, Schmid, Ponce'03; Sivic & Zisserman,'03; Csurka et al.'04; Zhang et al.'06)

Locally orderless
image models



(Koenderink & Van Doorn'99; Dalal & Triggs'05; Lazebnik, Schmid, Ponce'06; Chum & Zisserman'07)

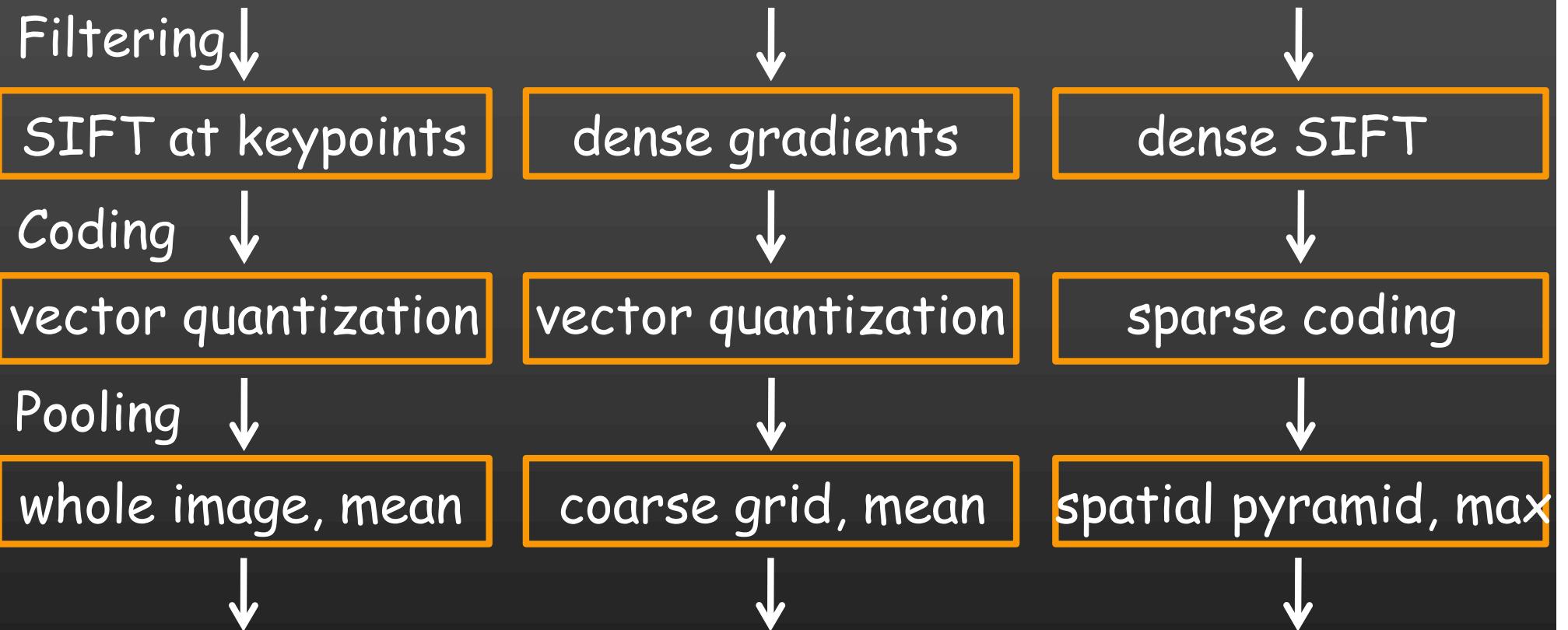
Image categorization as supervised classification



$$\min_{f \in \mathcal{F}} \frac{1}{N} \sum_n \ell(z_n, f(\phi(x_n))) + \Omega(f)$$

affine

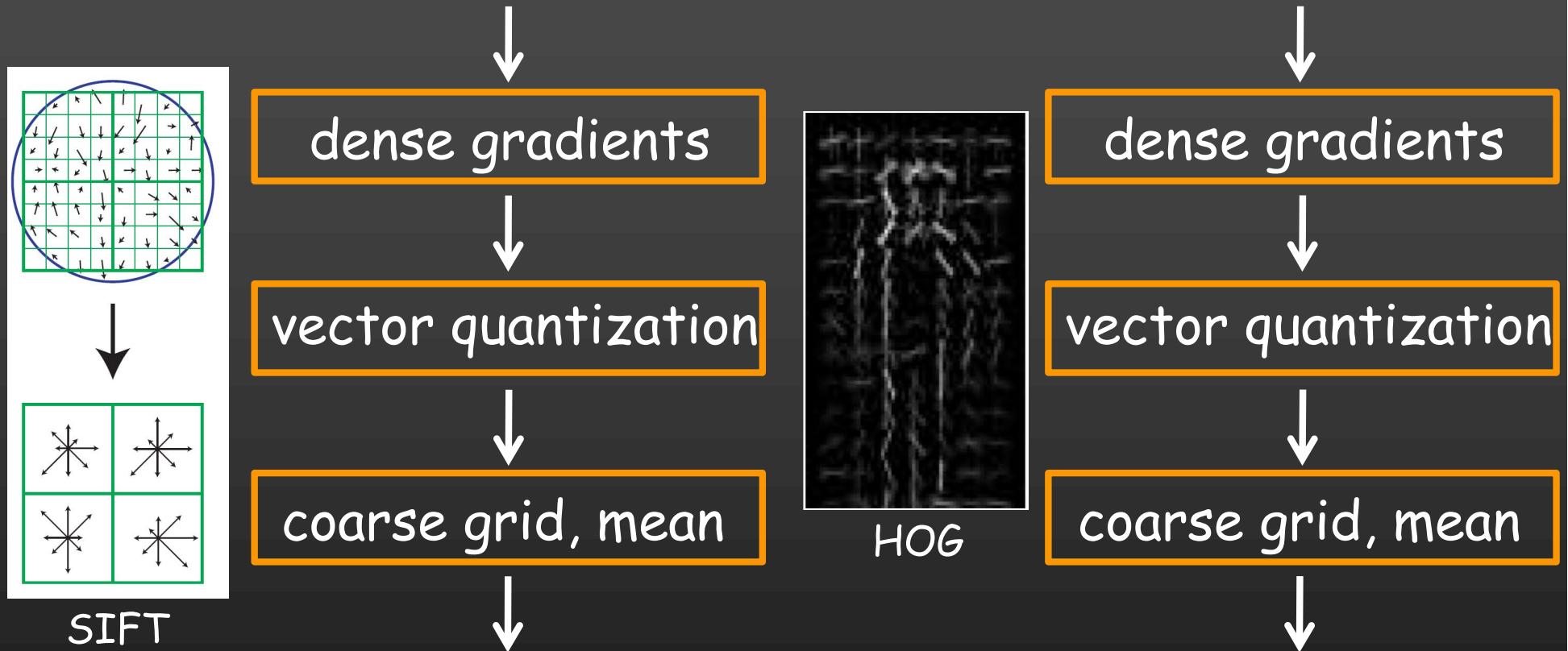
A common architecture for image classification



(Lowe'04, Csurka et al.'04, Dalal & Triggs'05)

(Yang et al.'09-10, Boureau et al.'10, Mallat'11)

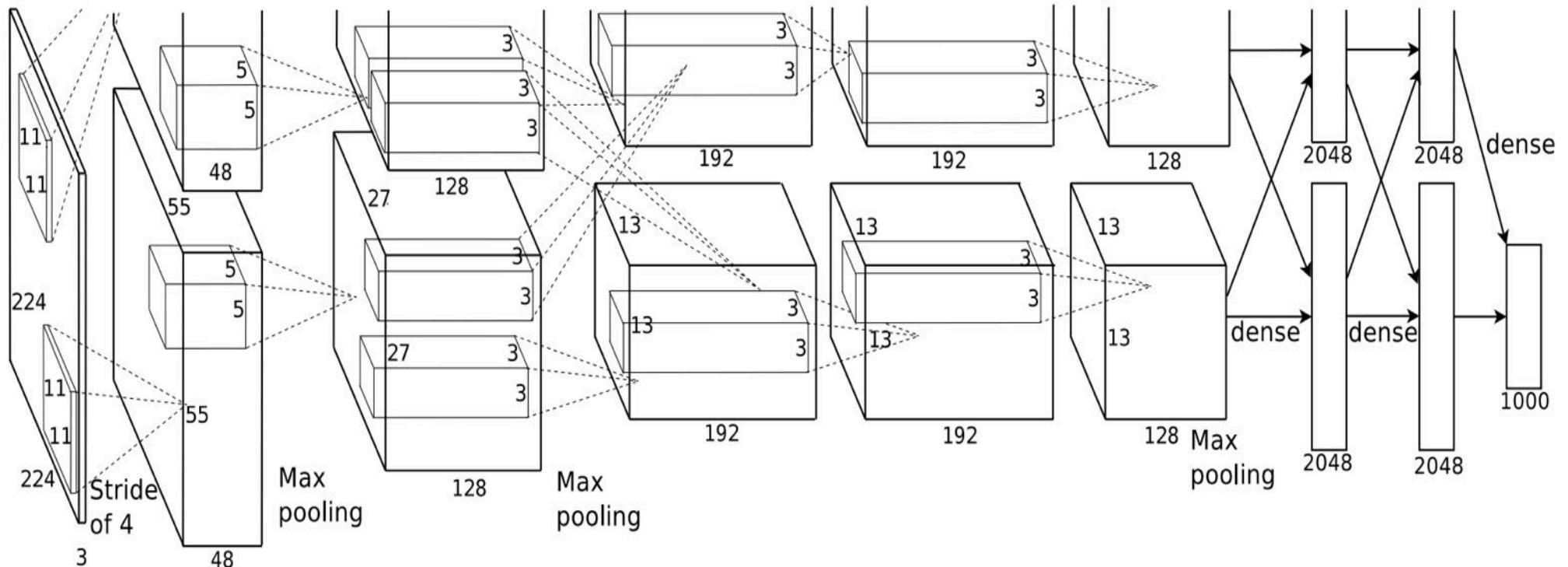
A common architecture for image classification



(Lowe'04, Csurka et al.'04, Dalal & Triggs'05)

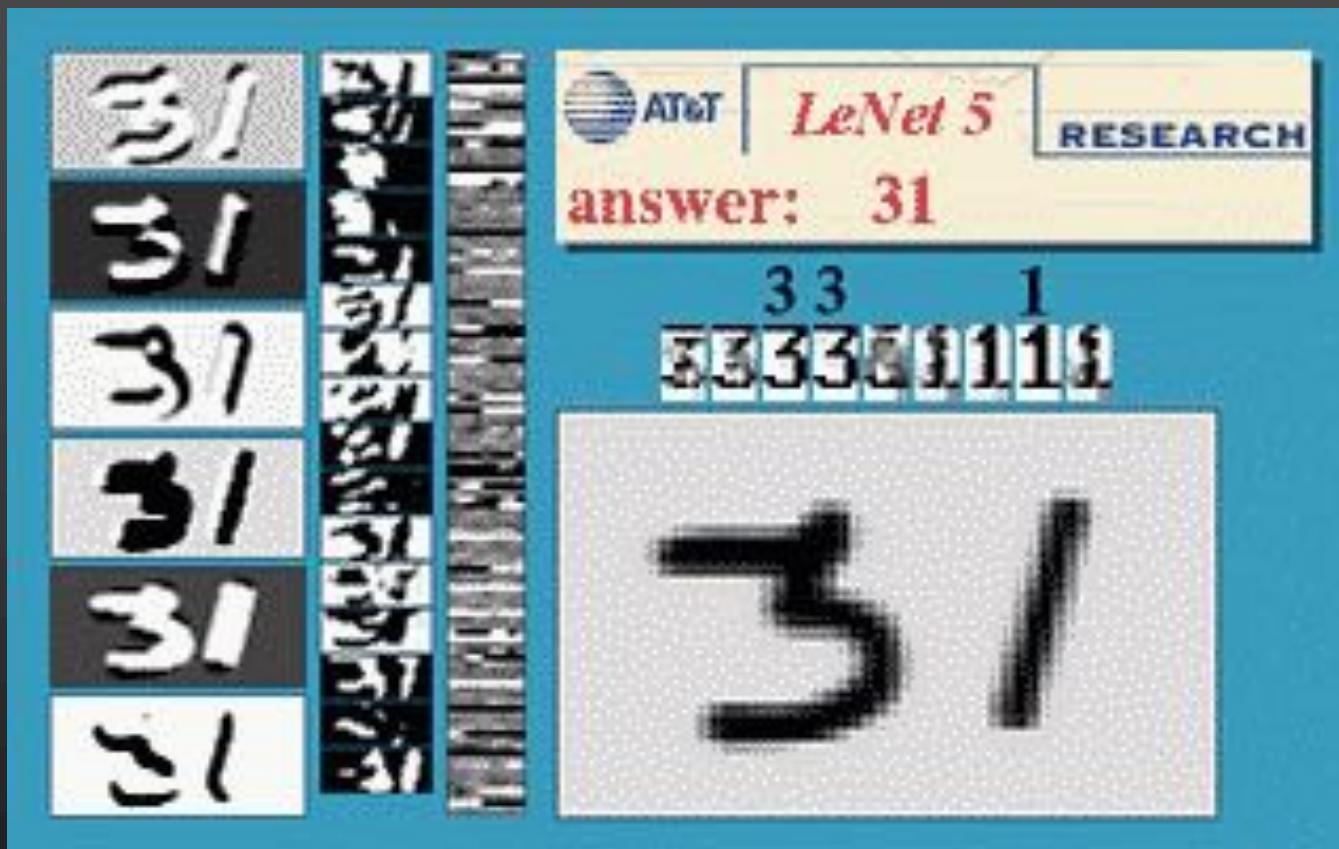
(Yang et al.'09-10, Boureau et al.'10, Mallat'11)

A common architecture for image classification



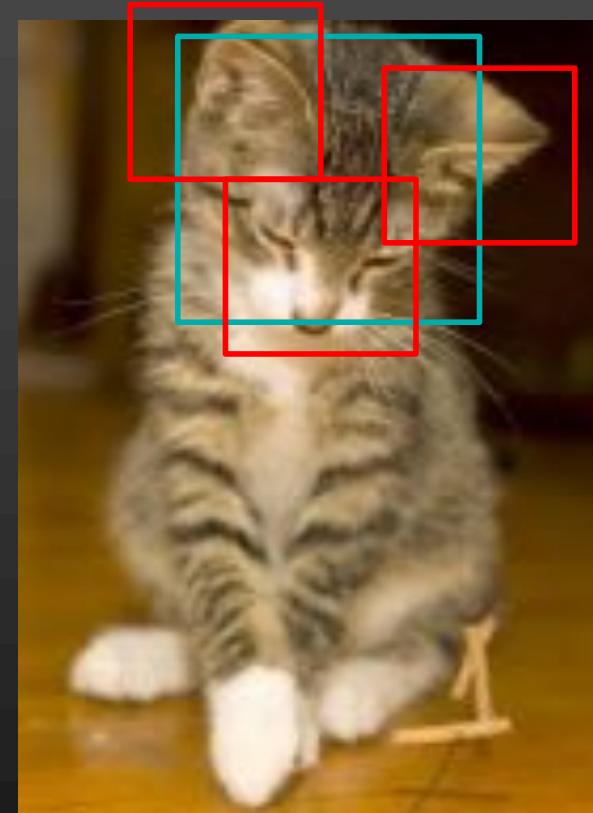
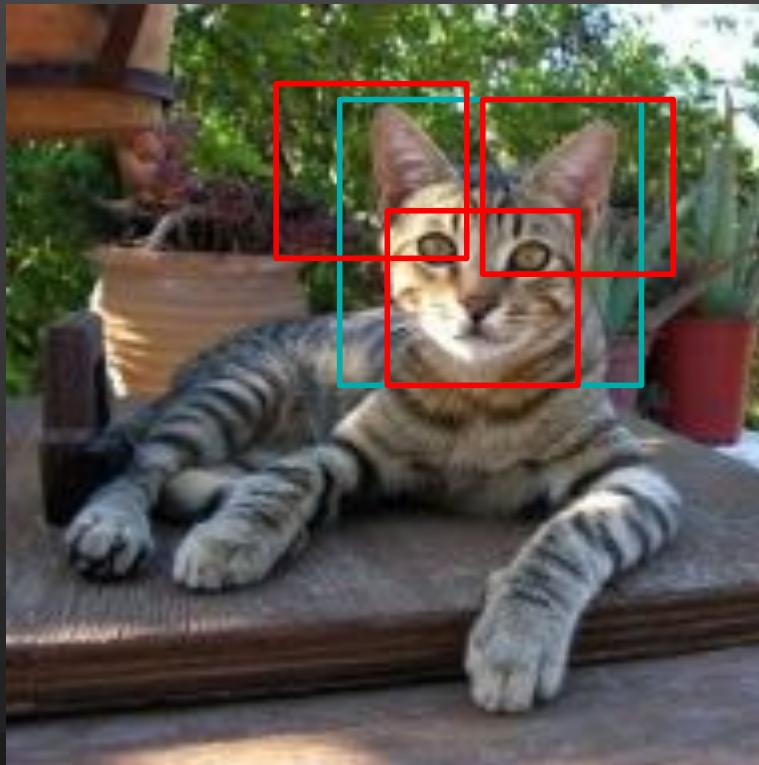
(Deep learning: Krizhevsky, Sutskever, Hinton, 2012)

A common architecture for image classification



(Convolutional neural networks, LeCun, 1998)

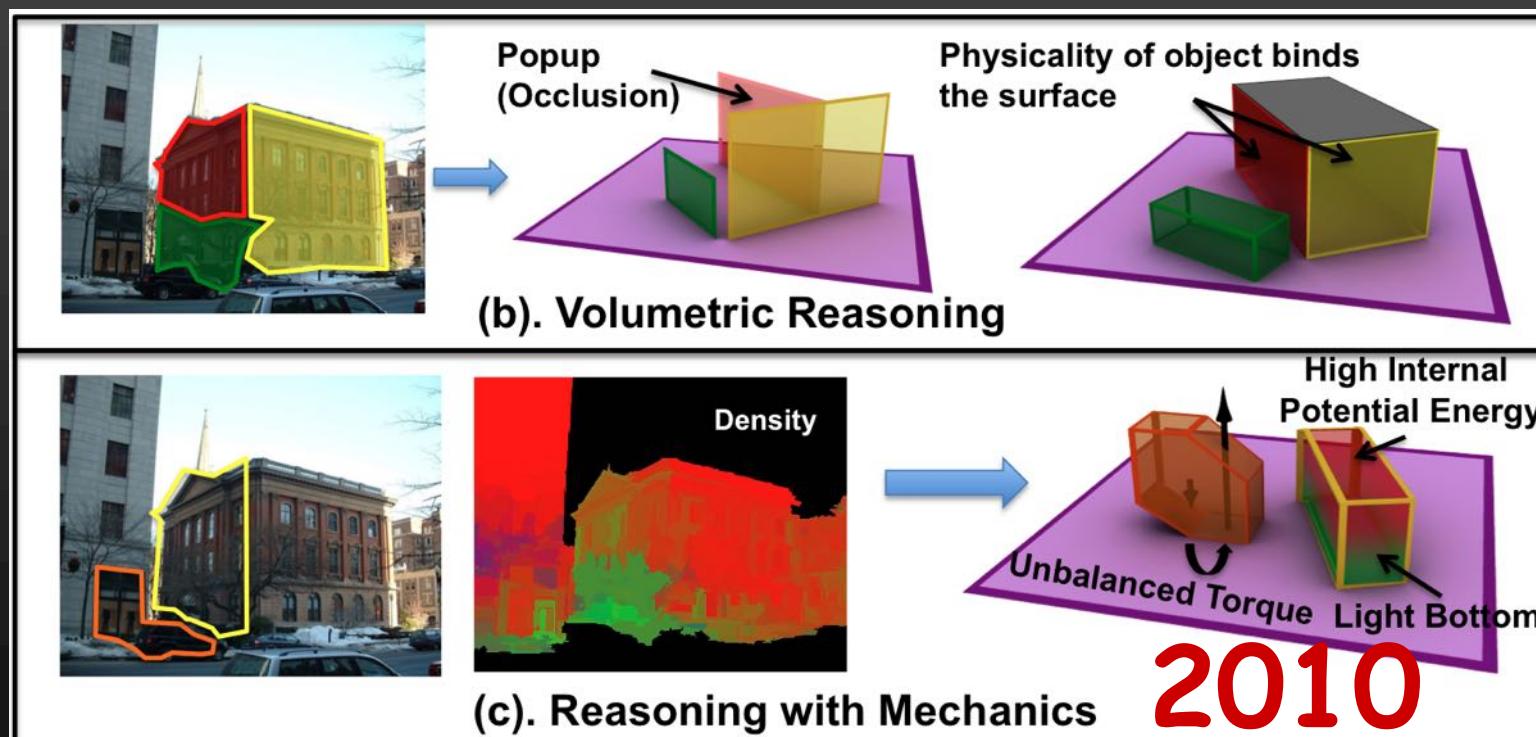
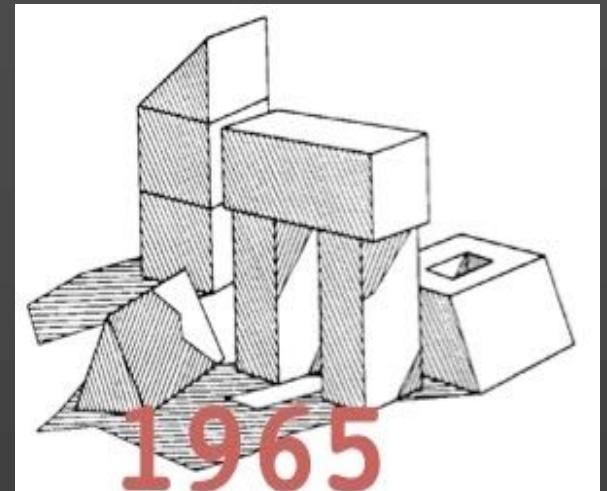
Object detection (vue "d'artiste")



(A la Felzenszwalb, McAllester, Ramanan, 2008)

What about scene understanding?

The blocks world revisited

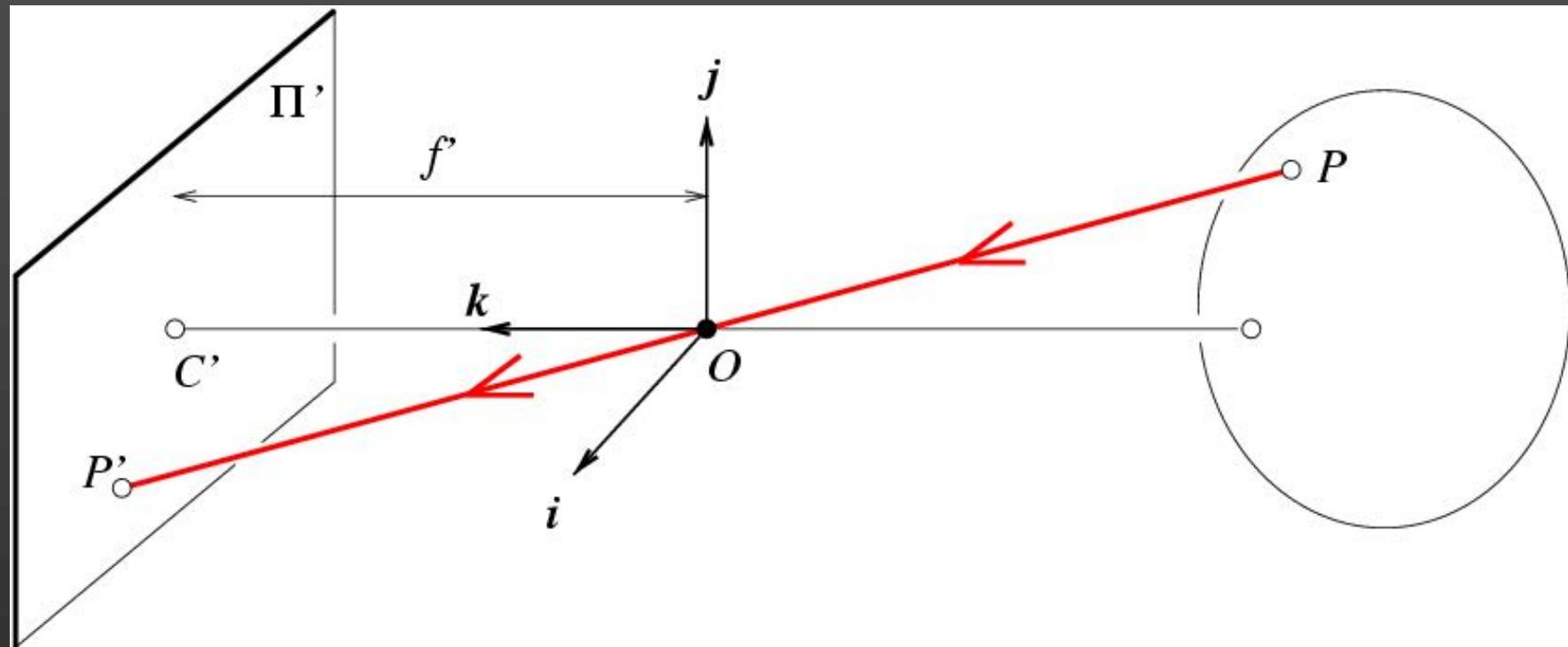


(Gupta, Efros, Hebert, ECCV'10)

Outline

- What computer vision is about
- What this class is about
- A brief history of visual recognition
- A brief recap on geometry

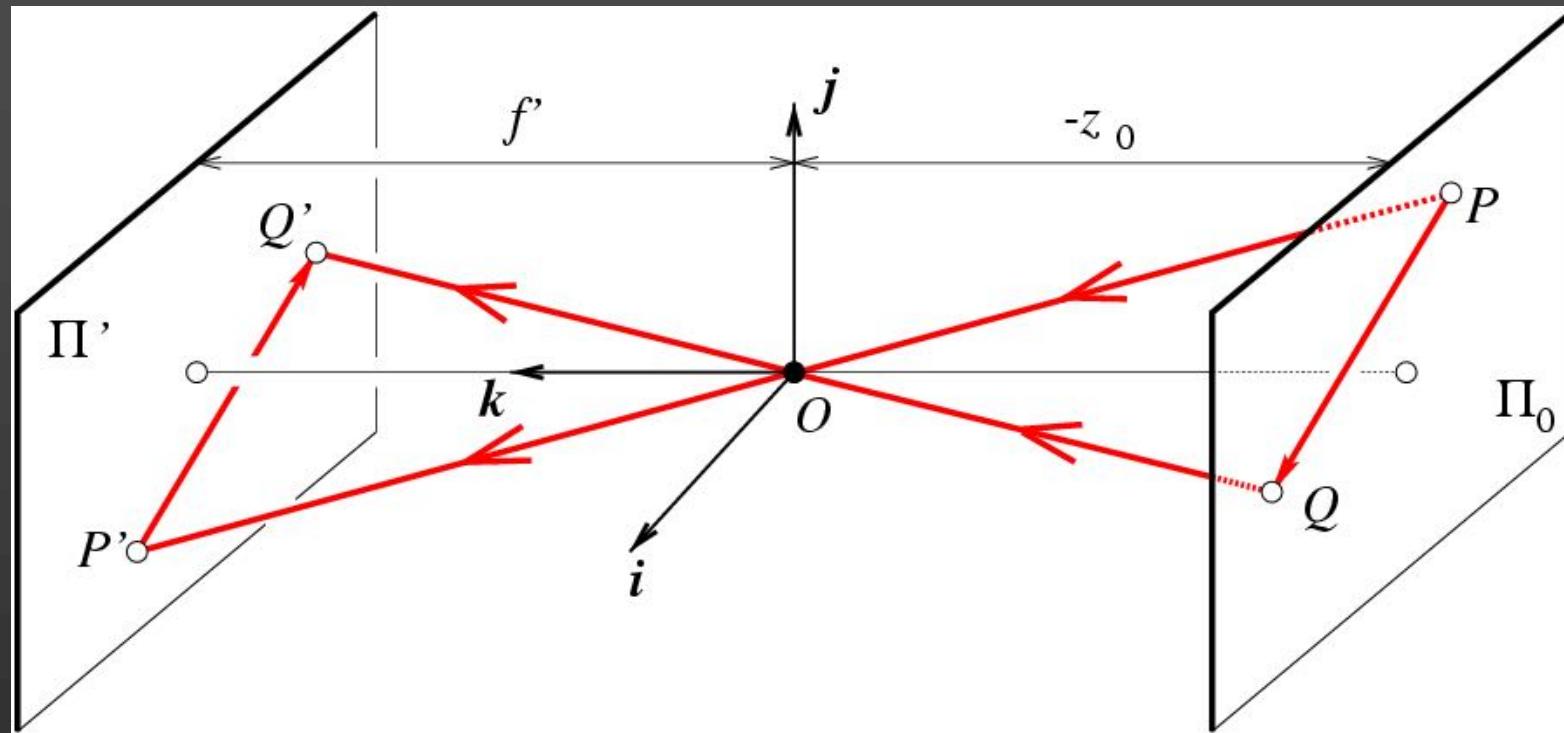
Pinhole perspective equation



$$\begin{cases} x' = f' \frac{x}{z} \\ y' = f' \frac{y}{z} \end{cases}$$

NOTE: z is always negative..

Affine models: Weak perspective projection

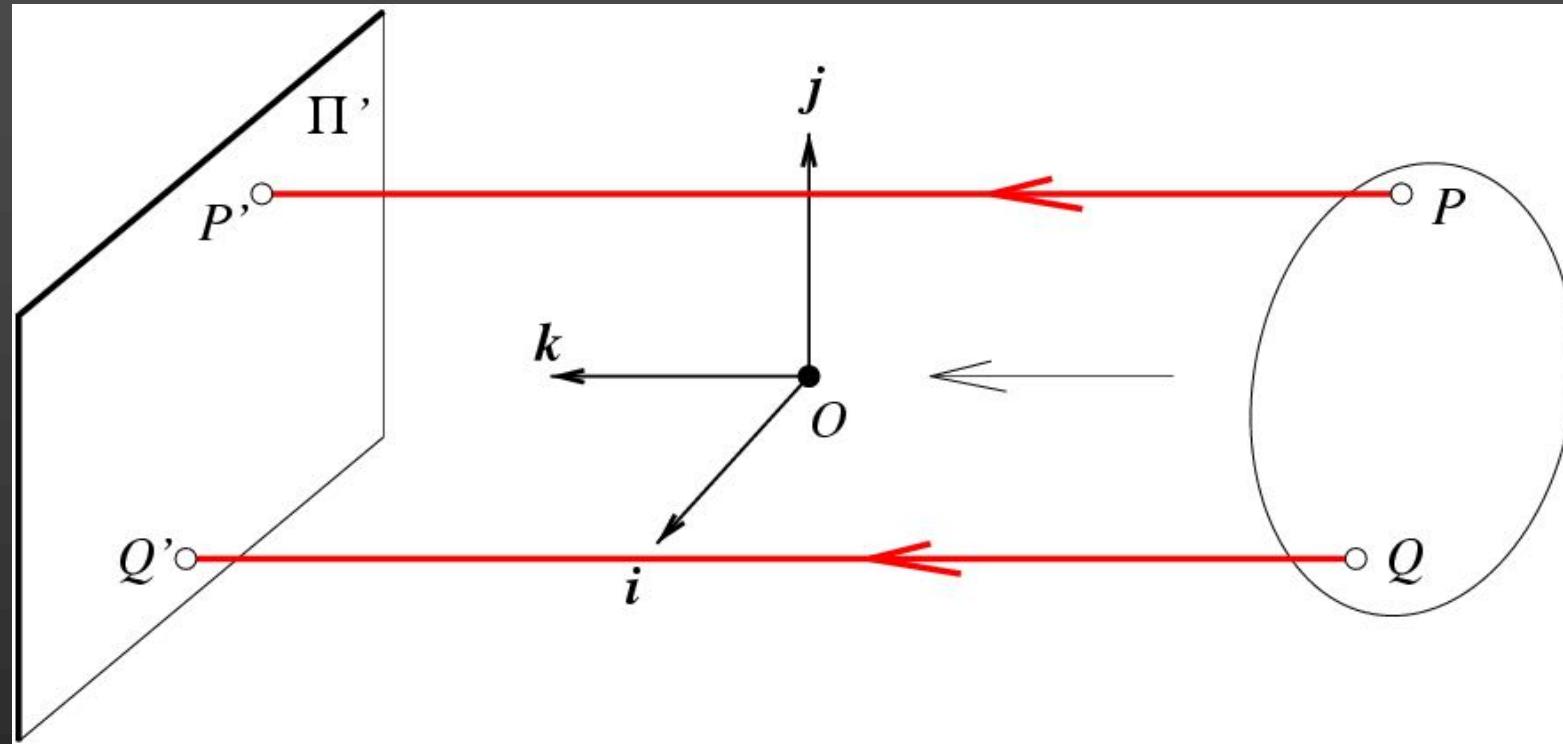


$$\begin{cases} x' = -mx \\ y' = -my \end{cases}$$

where $m = -\frac{f'}{z_0}$ is the magnification.

When the scene relief is small compared its distance from the camera, m can be taken constant: weak perspective projection.

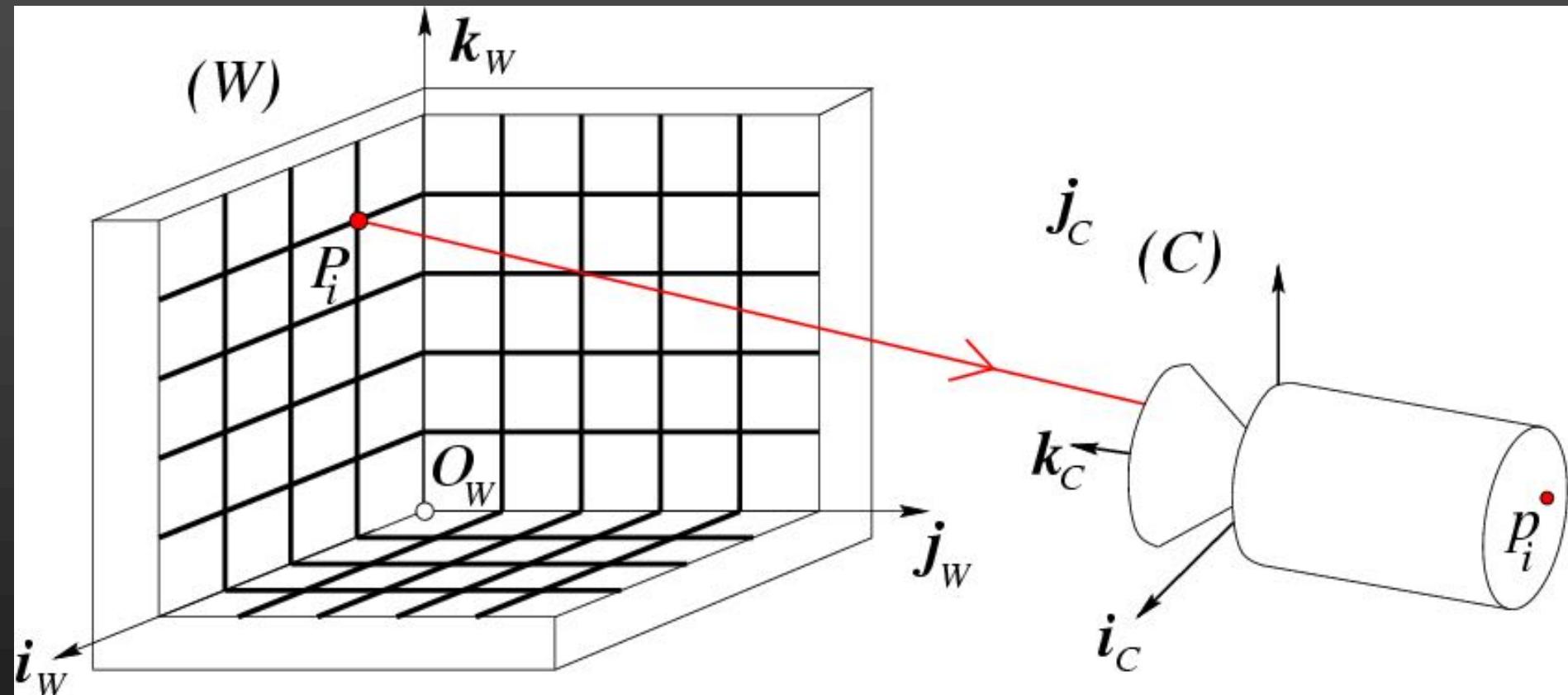
Affine models: Orthographic projection



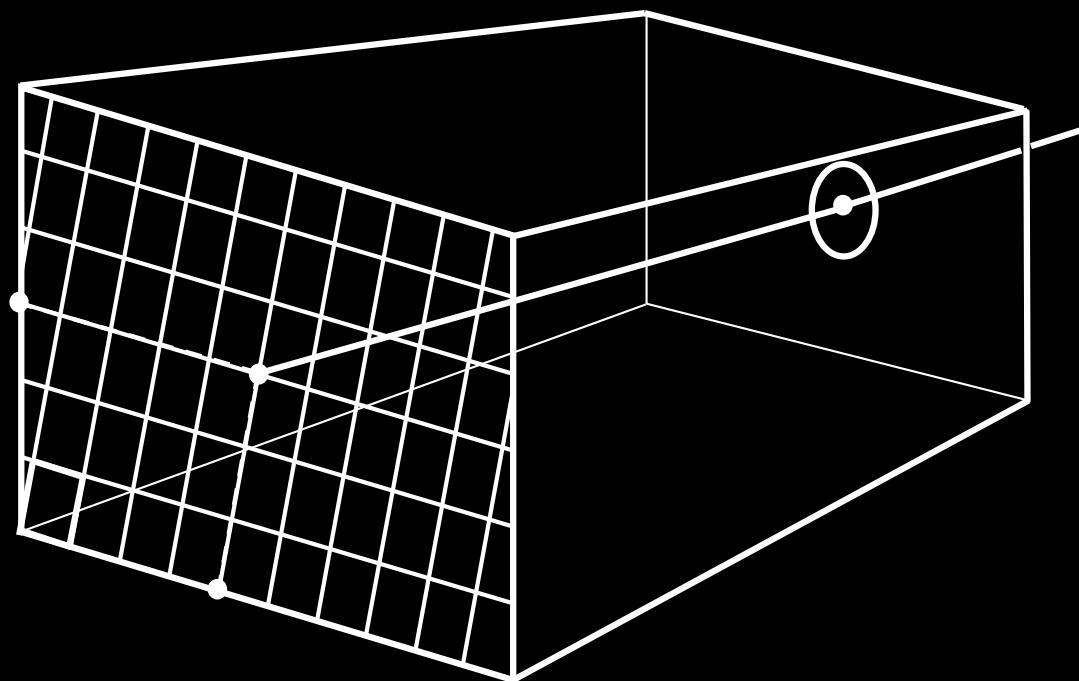
$$\begin{cases} x' = x \\ y' = y \end{cases}$$

When the camera is at a (roughly constant) distance from the scene, take $m=1$.

Analytical camera geometry



Cameras and their parameters



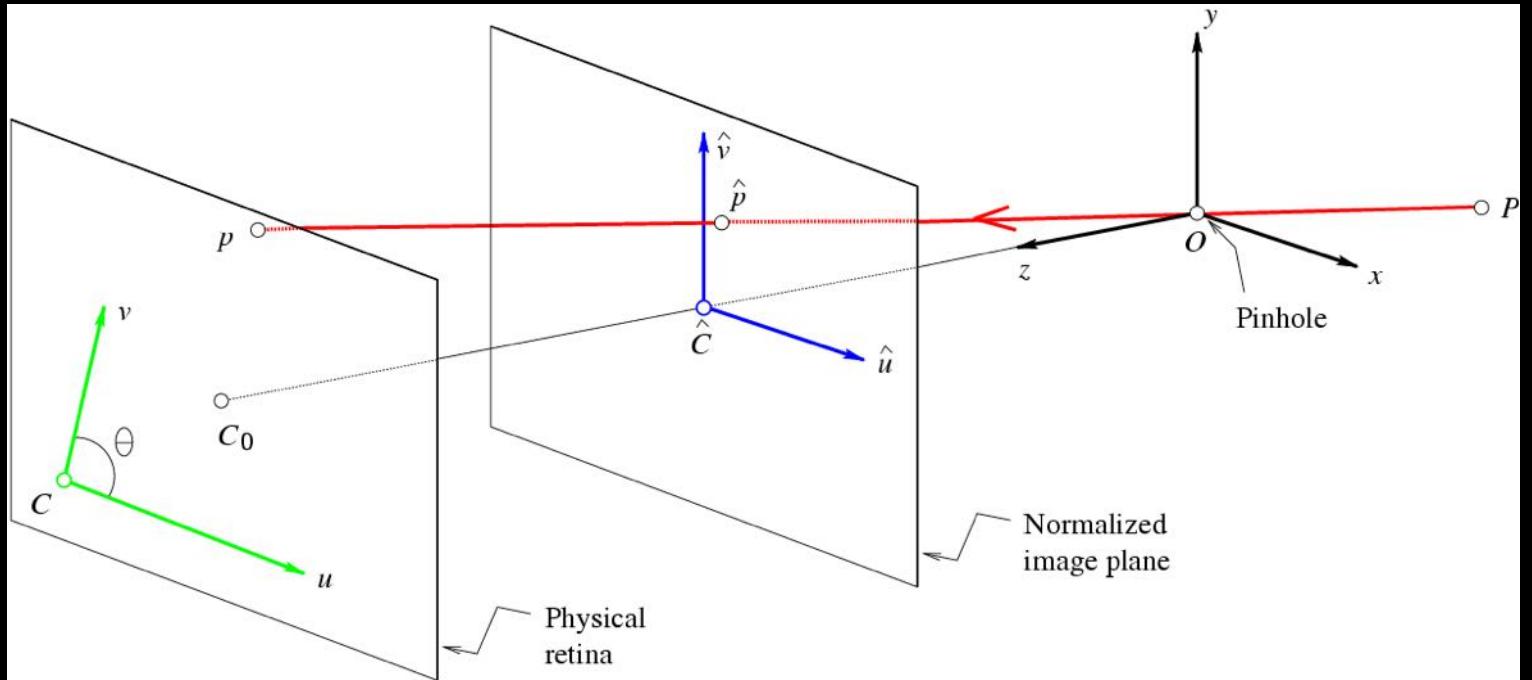
The intrinsic parameters of a camera

Units:

k, l : pixel/m

f : m

α, β : pixel



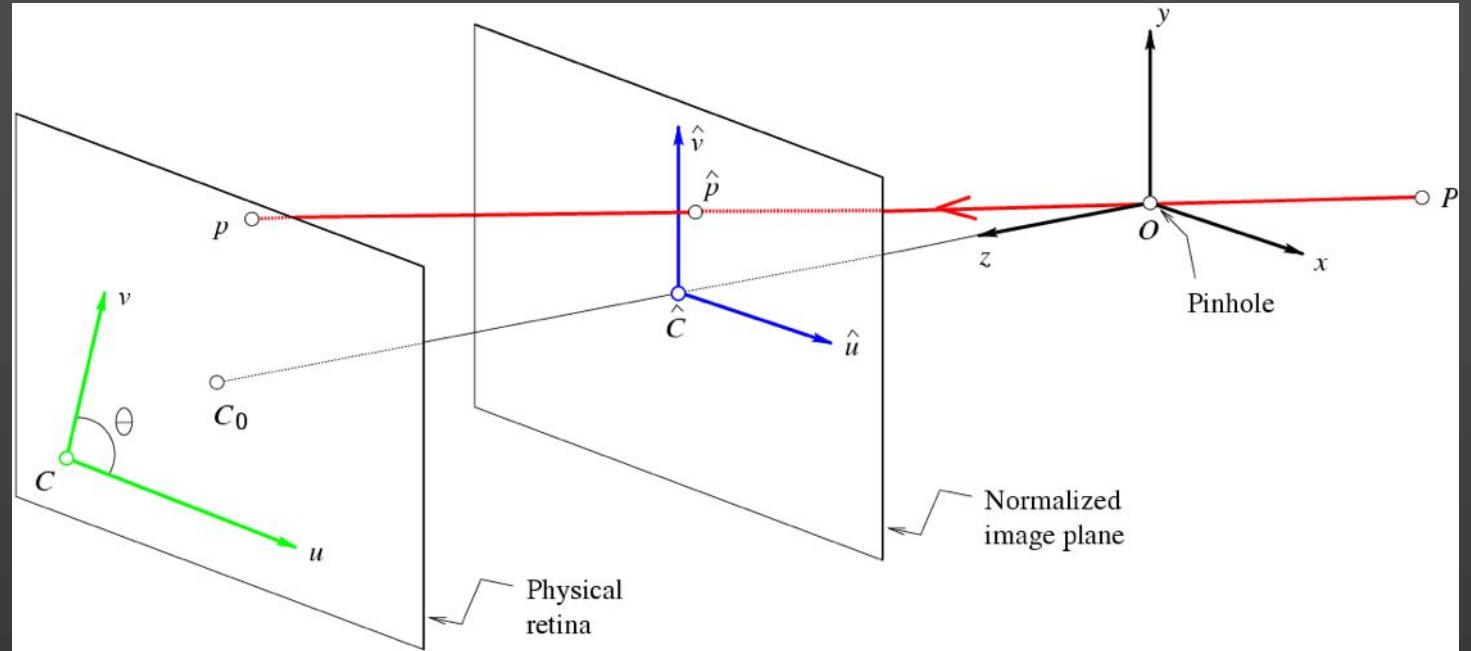
$$\begin{cases} \hat{u} = \frac{x}{z} \\ \hat{v} = \frac{y}{z} \end{cases} \iff \hat{\mathbf{p}} = \frac{1}{z} (\text{Id} \quad \mathbf{0}) \begin{pmatrix} \mathbf{P} \\ 1 \end{pmatrix}$$

Normalized image coordinates

Physical image coordinates

$$\begin{cases} u = kf \frac{x}{z} \\ v = lf \frac{y}{z} \end{cases} \rightarrow \begin{cases} u = \alpha \frac{x}{z} + u_0 \\ v = \beta \frac{y}{z} + v_0 \end{cases} \rightarrow \begin{cases} u = \alpha \frac{x}{z} - \alpha \cot \theta \frac{y}{z} + u_0 \\ v = \frac{\beta}{\sin \theta} \frac{y}{z} + v_0 \end{cases}$$

The intrinsic parameters of a camera



Calibration matrix

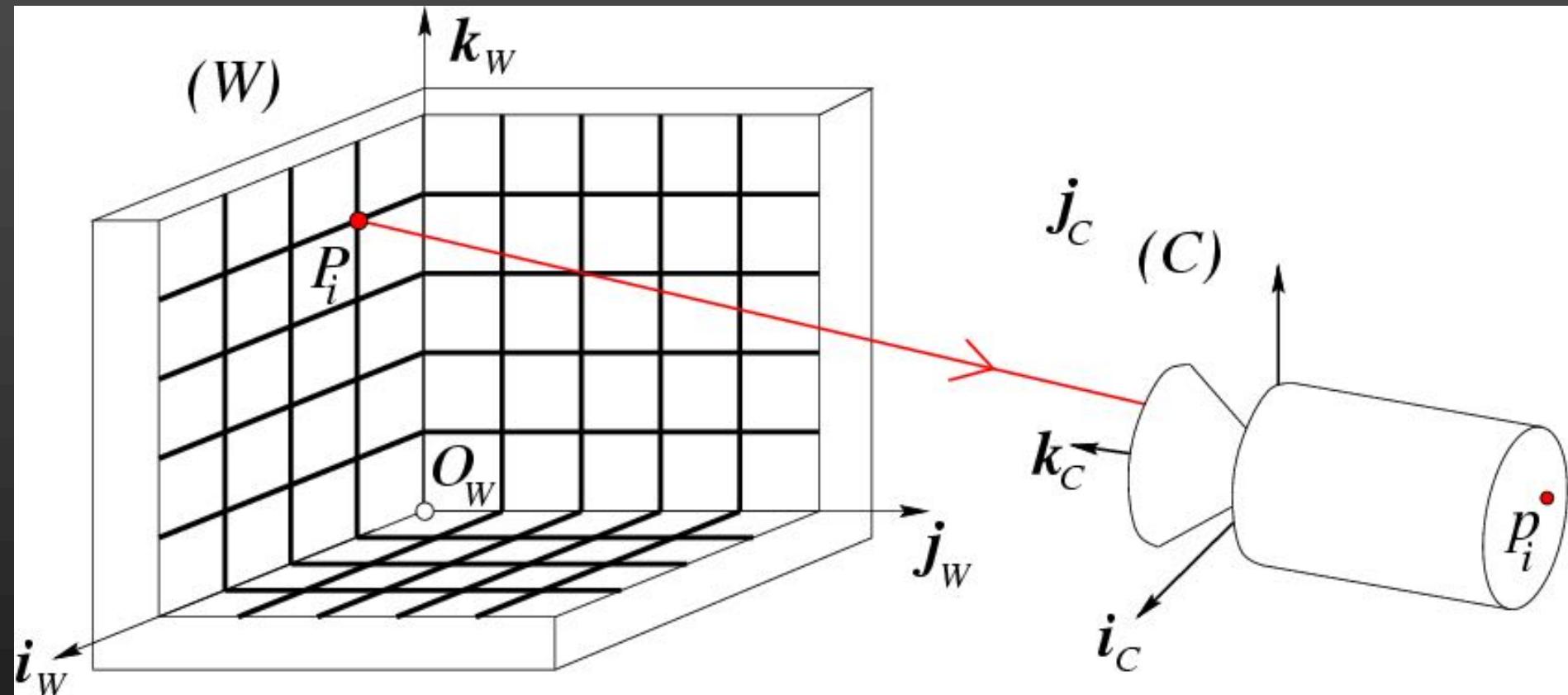
$$\mathbf{p} = \mathcal{K}\hat{\mathbf{p}}, \quad \text{where } \mathbf{p} = \begin{pmatrix} u \\ v \\ 1 \end{pmatrix} \quad \text{and} \quad \mathcal{K} \stackrel{\text{def}}{=} \begin{pmatrix} \alpha & -\alpha \cot \theta & u_0 \\ 0 & \frac{\beta}{\sin \theta} & v_0 \\ 0 & 0 & 1 \end{pmatrix}$$

→ Homogeneous coordinates

The perspective
projection equation

$$\mathbf{p} = \frac{1}{z} \mathcal{M} \mathbf{P}, \quad \text{where} \quad \mathcal{M} \stackrel{\text{def}}{=} (\mathcal{K} \quad \mathbf{0})$$

Analytical camera geometry



The extrinsic parameters of a camera

- When the camera frame (C) is different from the world frame (W),

$$\begin{pmatrix} {}^C P \\ 1 \end{pmatrix} = \begin{pmatrix} {}_W^C \mathcal{R} & {}^C O_W \\ \mathbf{0}^T & 1 \end{pmatrix} \begin{pmatrix} {}^W P \\ 1 \end{pmatrix}.$$

- Thus,

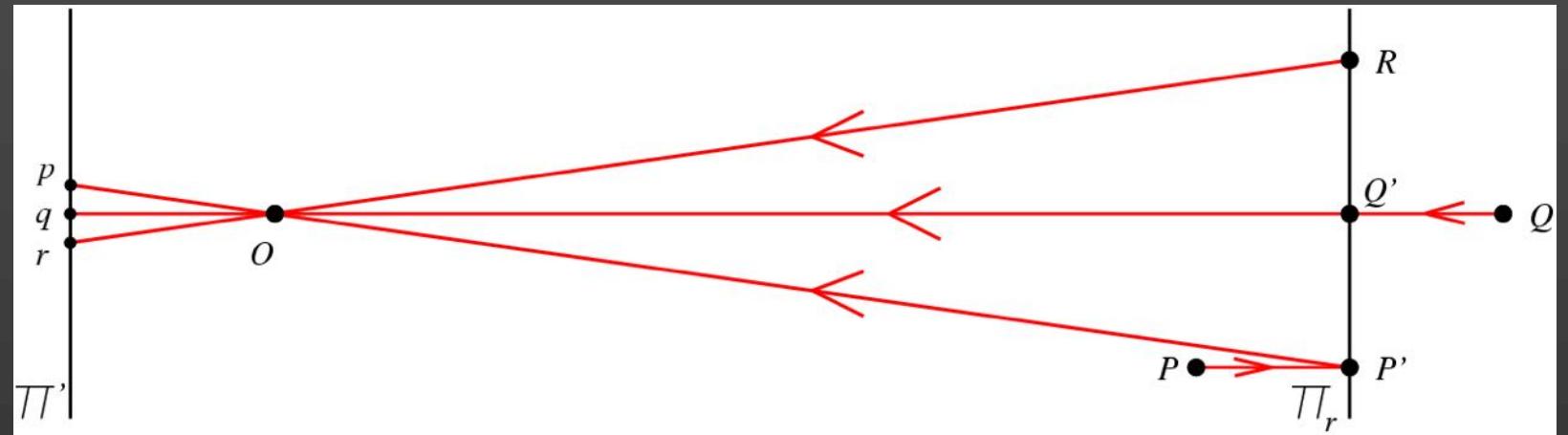
$$\boxed{\mathbf{p} = \frac{1}{z} \mathcal{M} \mathbf{P},} \quad \text{where} \quad \left\{ \begin{array}{l} \mathcal{M} = \mathcal{K}(\mathcal{R} \quad \mathbf{t}), \\ \mathcal{R} = {}_W^C \mathcal{R}, \\ \mathbf{t} = {}^C O_W, \\ \mathbf{P} = \begin{pmatrix} {}^W P \\ 1 \end{pmatrix}. \end{array} \right.$$

- Note: z is *not* independent of \mathcal{M} and \mathbf{P} :

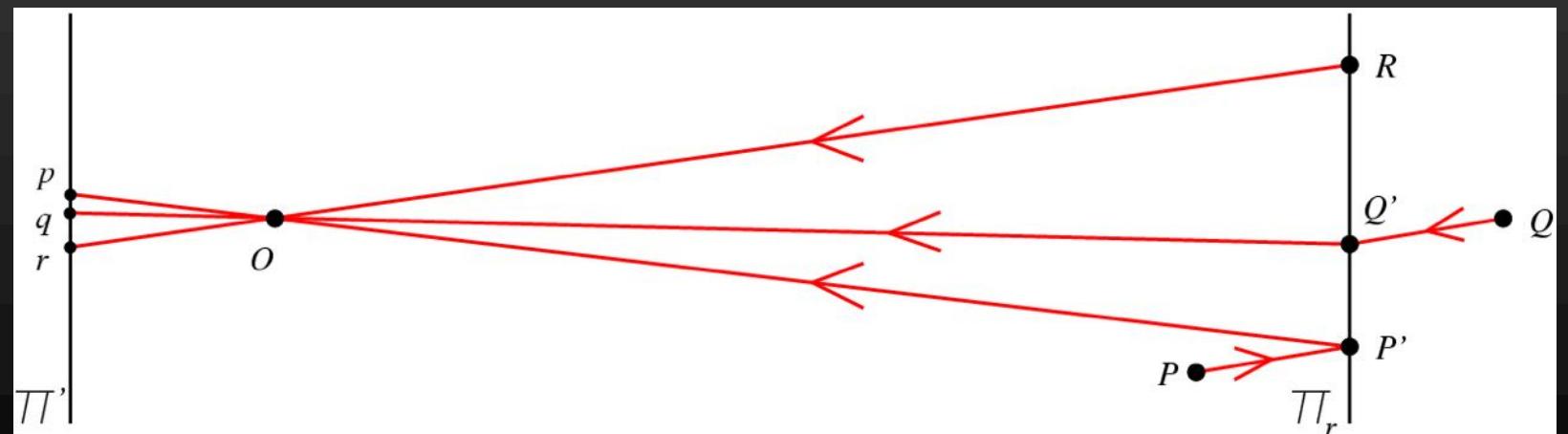
$$\mathcal{M} = \begin{pmatrix} \mathbf{m}_1^T \\ \mathbf{m}_2^T \\ \mathbf{m}_3^T \end{pmatrix} \implies z = \mathbf{m}_3 \cdot \mathbf{P}, \quad \text{or} \quad \left\{ \begin{array}{l} u = \frac{\mathbf{m}_1 \cdot \mathbf{P}}{\mathbf{m}_3 \cdot \mathbf{P}}, \\ v = \frac{\mathbf{m}_2 \cdot \mathbf{P}}{\mathbf{m}_3 \cdot \mathbf{P}}. \end{array} \right.$$

Affine cameras

Weak-perspective projection

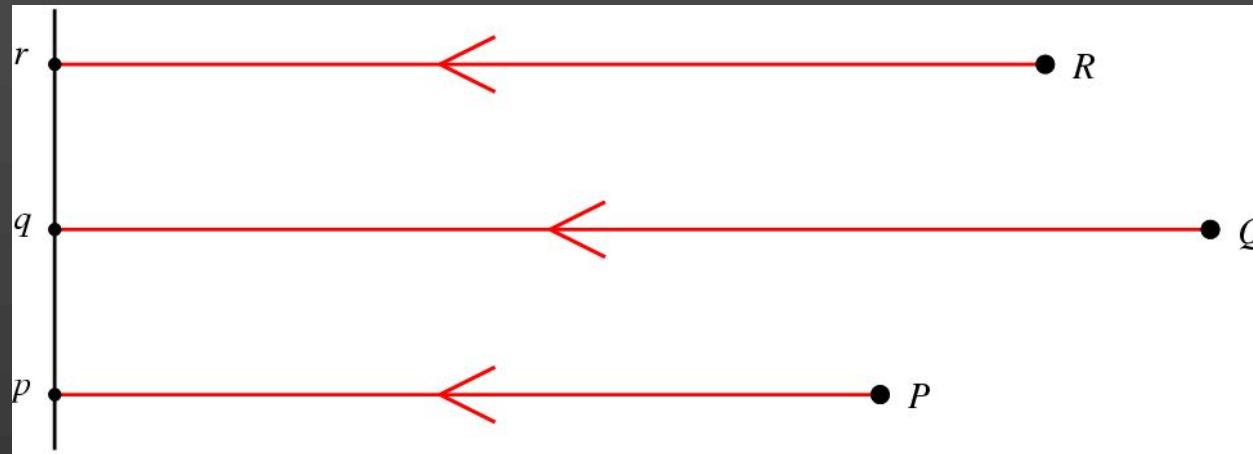


Paraperspective projection

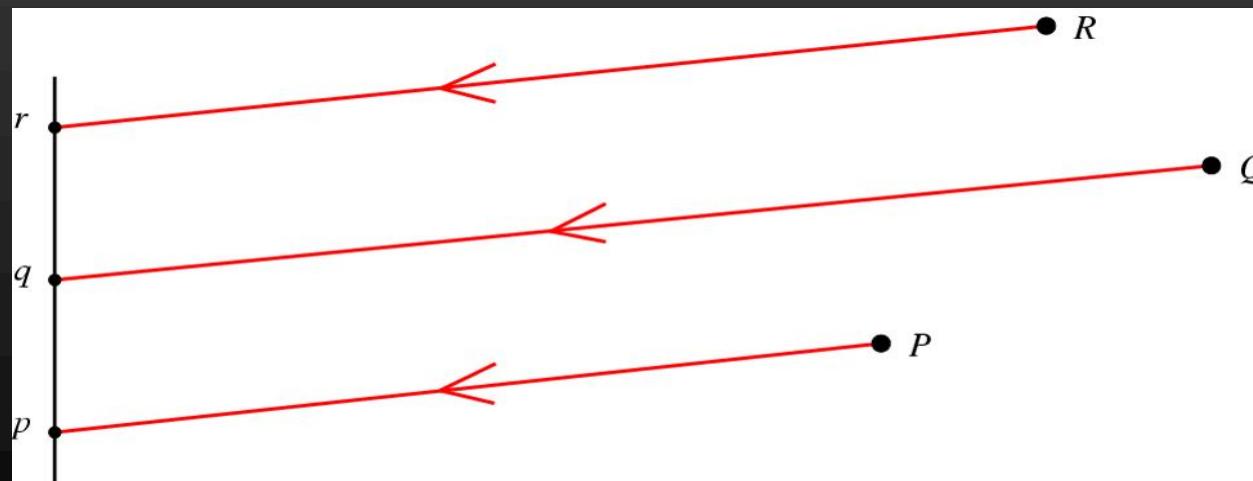


More affine cameras

Orthographic projection



Parallel projection



Weak-perspective projection model

$$\mathbf{p} = \frac{1}{z_r} \mathcal{M} \mathbf{P}$$

(\mathbf{p} and \mathbf{P} are in homogeneous coordinates)

$$\mathbf{p} = M \mathbf{P}$$

(\mathbf{P} is in homogeneous coordinates)

$$\mathbf{p} = A \mathbf{P} + \mathbf{b}$$

(neither \mathbf{p} nor \mathbf{P} is in hom. coordinates)



