Computer Vision I

Object Detection – 19.06.2013

TECHNISCHE UNIVERSITÄT DARMSTADT

vi visual inference

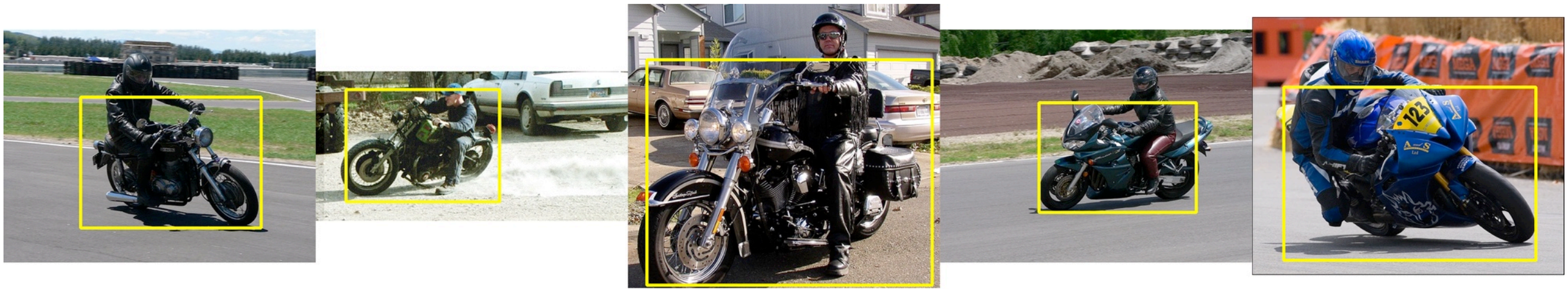
Stefan Roth, FB Informatik, TU Darmstadt – 19.06.2013

Announcements

TECHNISCHE UNIVERSITÄT DARMSTADT

- Class next week
  - Will most likely take place (contrary what was announced in the first lecture)
  - If there are any changes, we will post to the mailing list

From Categorization to Detection



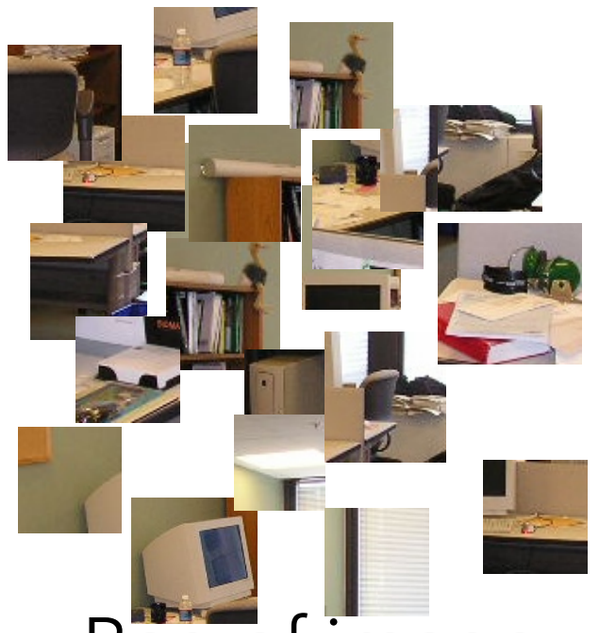
- We may not only want to recognize if an object is in an image, but rather find it in the image
  - Object detection / localization
- Next:
  - Sliding window approach
  - Histogram of Oriented Gradients (HOG)
    - global descriptor for object detection

Classifier-based methods

TECHNISCHE UNIVERSITÄT DARMSTADT

- Object detection and recognition is formulated as a classification problem.
- The image is partitioned into a set of overlapping windows
- … and a decision is taken at each window about if it contains a target object or not.

Where are the screens?



Bag of image patches

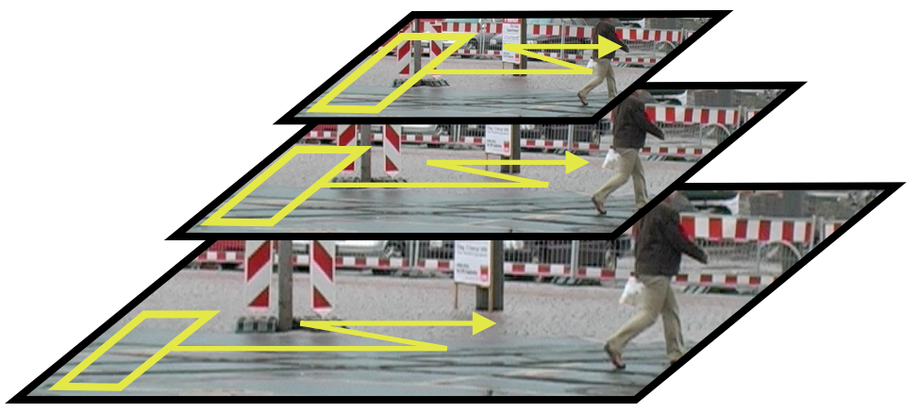
In some feature space

Sliding Window Methods - Overview

TECHNISCHE UNIVERSITÄT DARMSTADT

- Sliding window based object detection:

Scan Image → Extract Feature Vector → Classify Feature Vector → Non-Maxima Suppression



‘slide’ detection window
over all positions & scales

Two important questions:
1) which feature vector
2) which classifier

Histograms of Oriented Gradients (HOG)

TECHNISCHE UNIVERSITÄT DARMSTADT

- Original goal: Detect and localize people
- Applications:
  - Images, films & multi-media analysis
  - Pedestrian detection for autonomous cars
  - Visual surveillance, behavior analysis

[Dalal & Triggs]

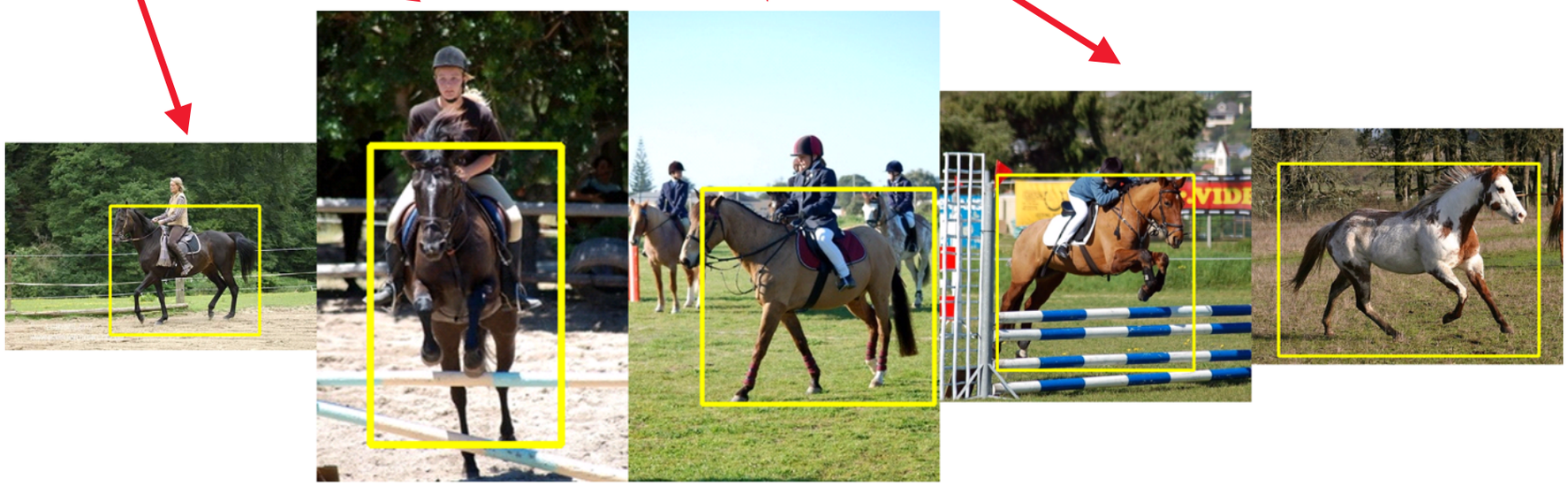
Challenges in Object Detection

TECHNISCHE UNIVERSITÄT DARMSTADT

Viewpoint variation

Articulation

Intra-class appearance variation



Challenges in Object Detection

TECHNISCHE UNIVERSITÄT DARMSTADT

- Some of the difficulties
  - Variable appearance (e.g. clothing)
  - Complex backgrounds
  - Unconstrained illumination
  - Occlusions, different scales
  - Wide variety of articulated poses

- Main assumption for HOG:
  - Limited amount of articulation and occlusion (e.g. upright fully visible people)

HOG: Static Feature Extraction

Input Image

Detection Window

Normalise gamma

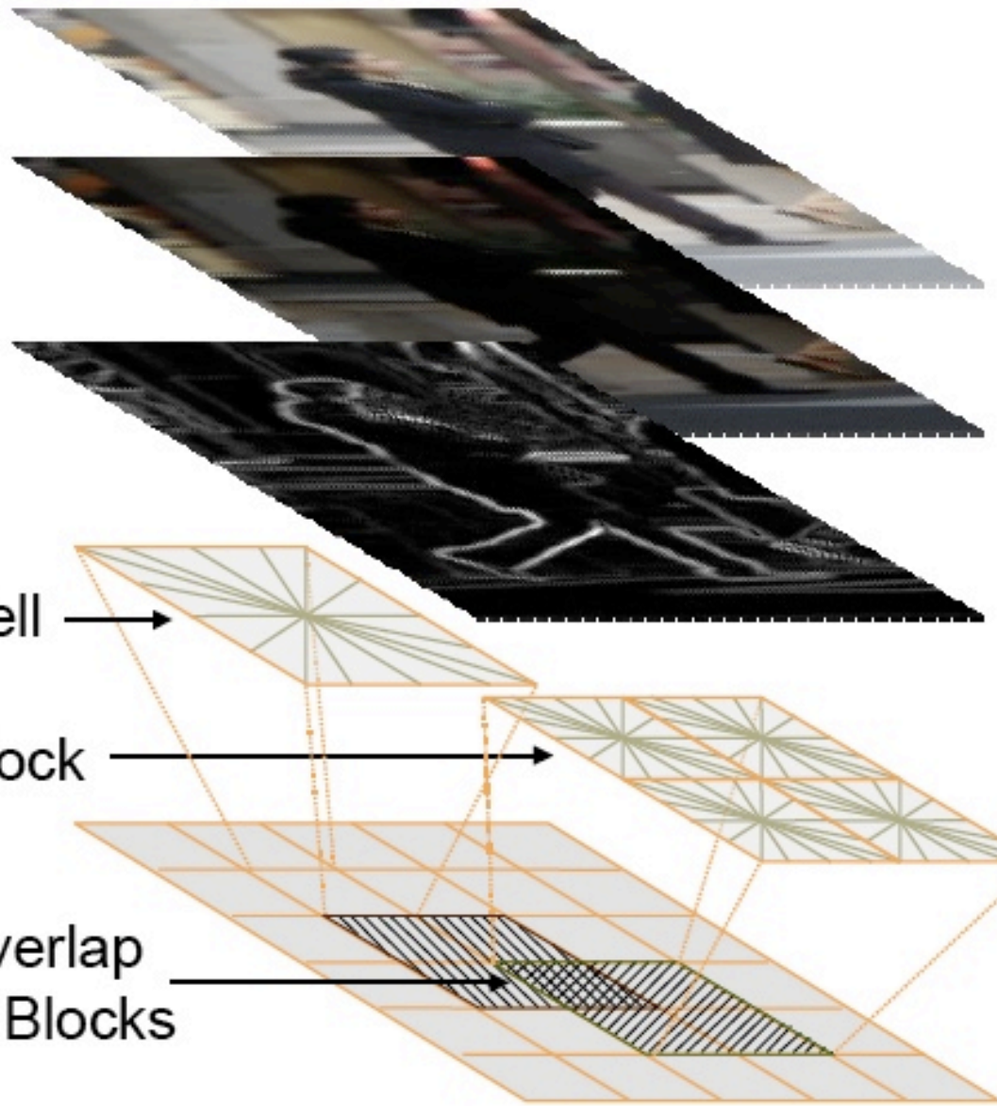
Compute gradients

Weighted vote in spatial & orientation cells

Contrast normalise over overlapping spatial cells

Collect HOGs over detection window

Linear SVM



Feature vector f = [..., ..., ...]

HOG-Descriptor & Variations

TECHNISCHE UNIVERSITÄT DARMSTADT

Parameters

- Gradient scale
- Orientation bins
- Percentage of block overlap

Schemes

- RGB or Lab, colour/gray-space
- Block normalisation

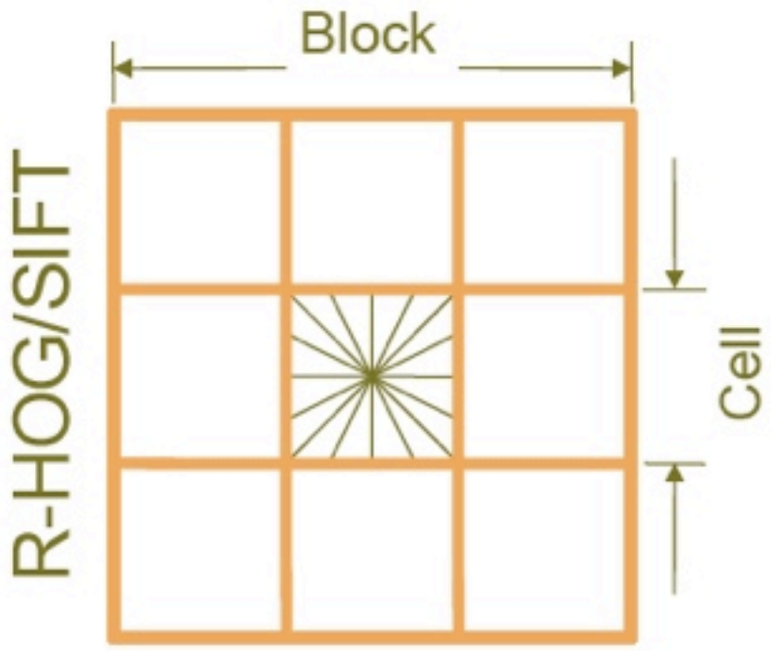
L2-norm,

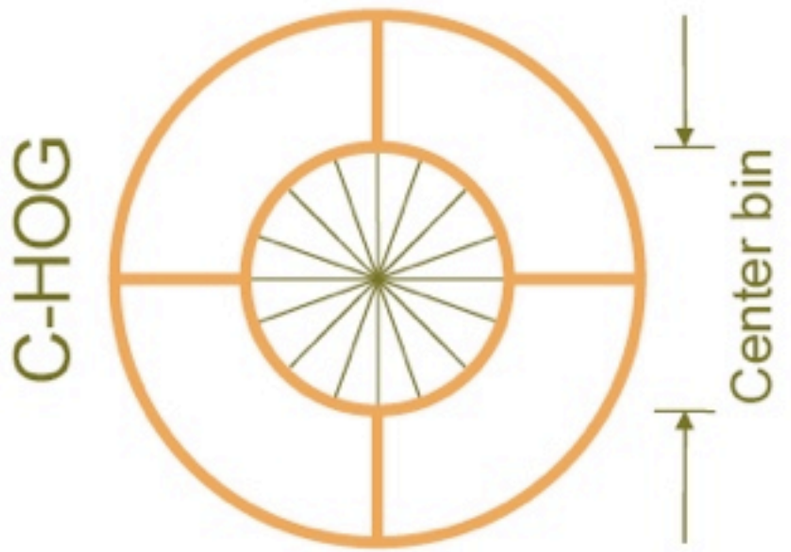
$$v \leftarrow v / \sqrt{\|v\|_2^2 + \varepsilon}$$

or

L1-norm,

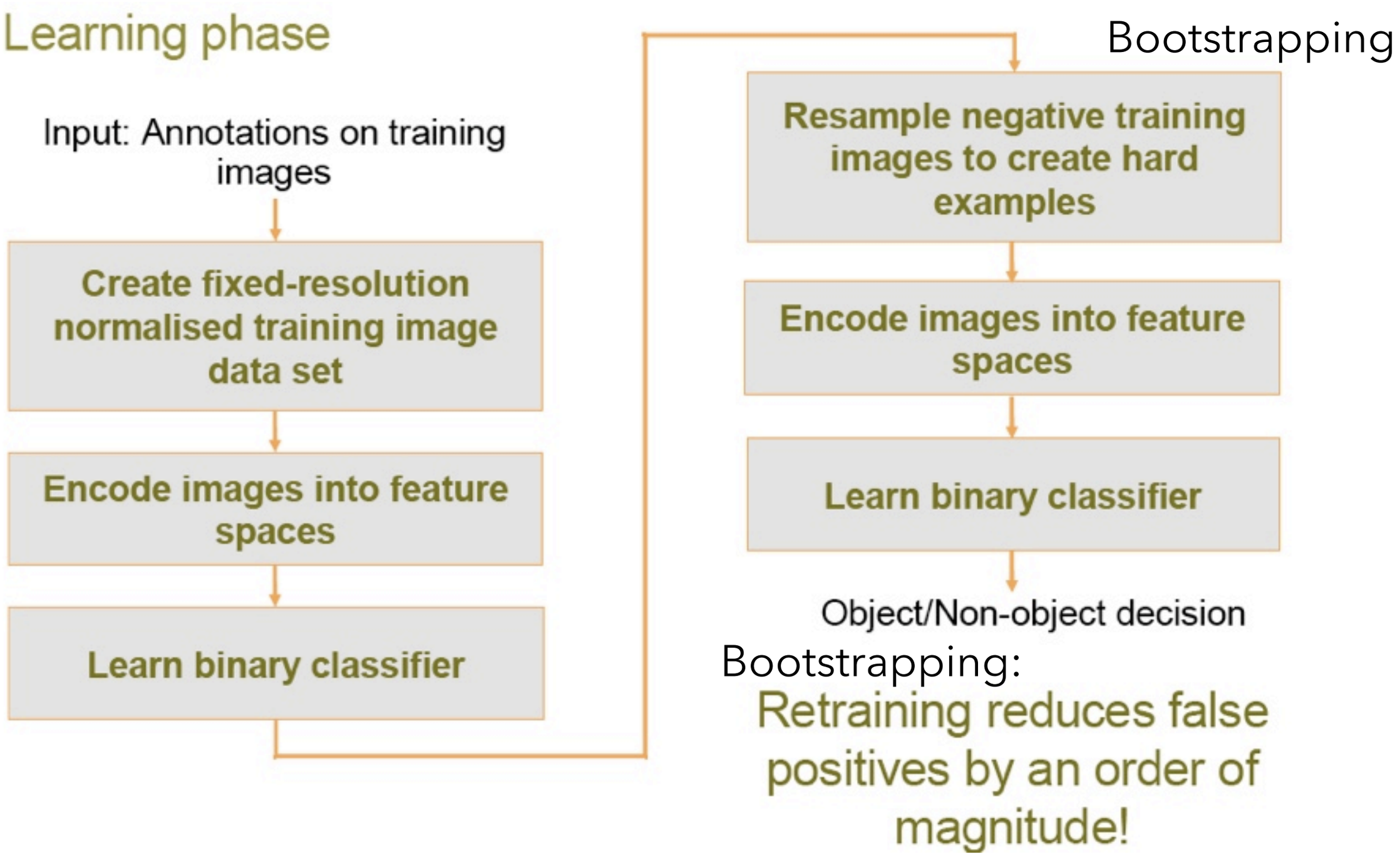
$$v \leftarrow \sqrt{v / (\|v\|_1 + \varepsilon)}$$





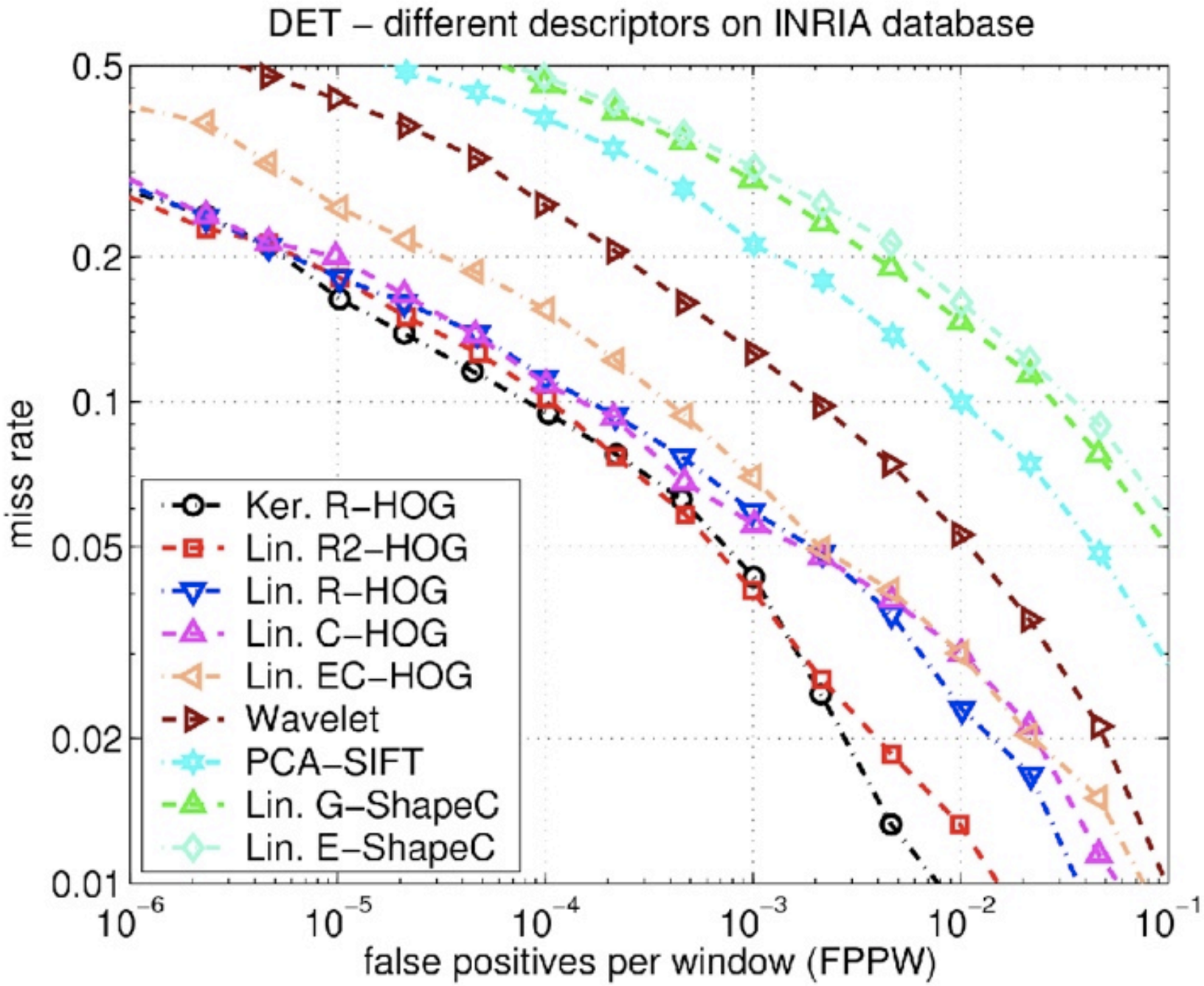
Overview of Learning

TECHNISCHE UNIVERSITÄT DARMSTADT



Performance on INRIA Dataset

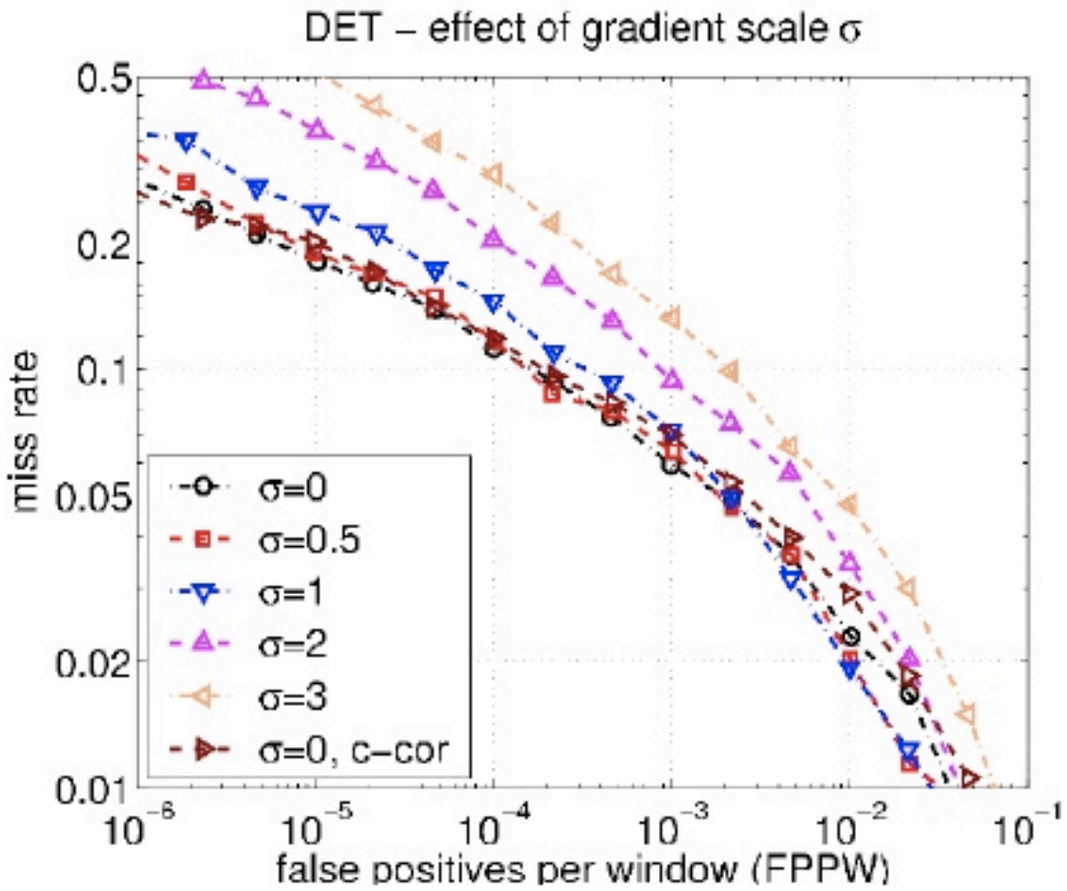
TECHNISCHE UNIVERSITÄT DARMSTADT



Effects of Parameters

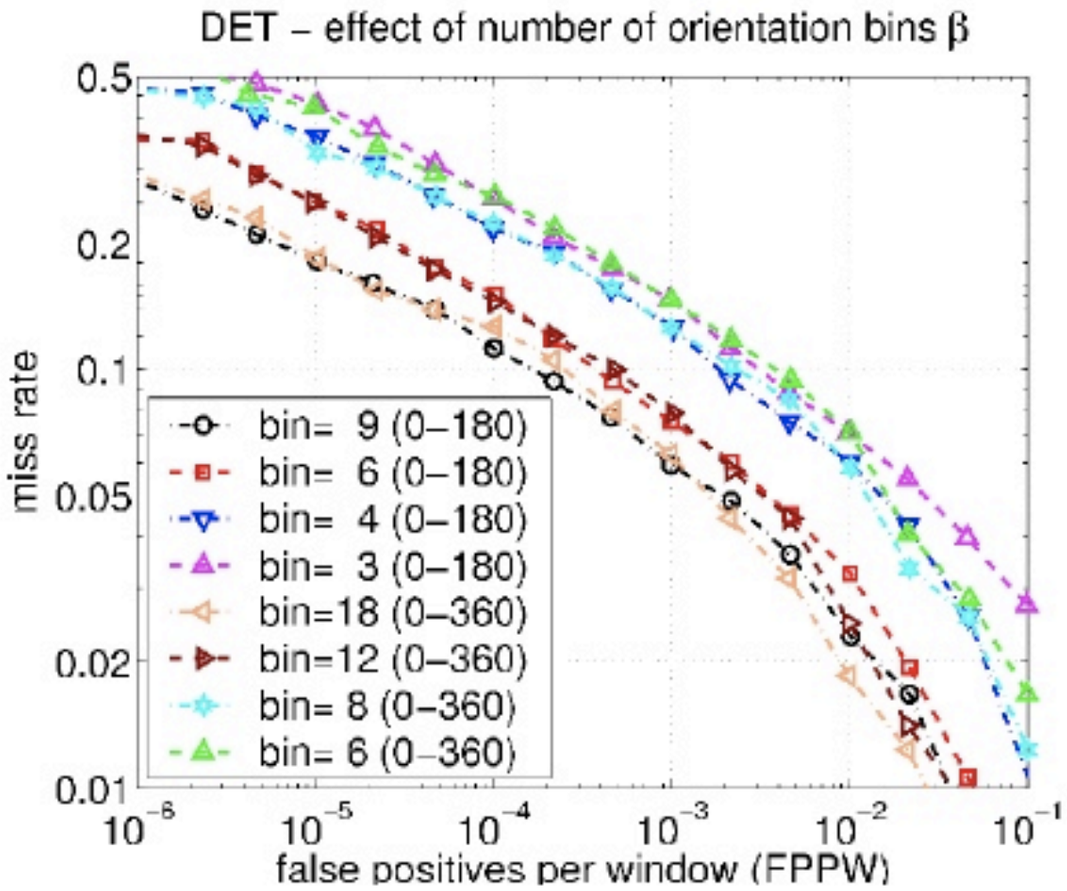
TECHNISCHE UNIVERSITÄT DARMSTADT

Gradient smoothing, σ



Reducing gradient scale from 3 to 0 decreases false positives by 10 times

Orientation bins, β

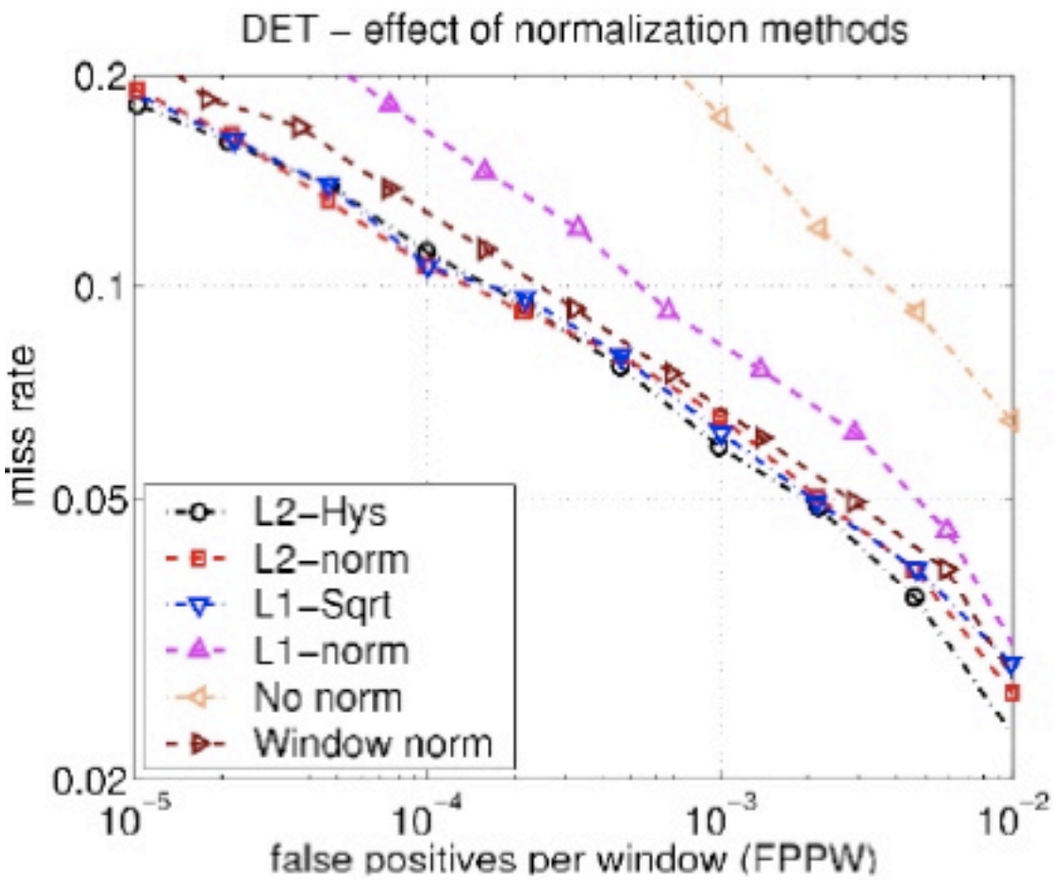


Increasing orientation bins from 4 to 9 decreases false positives by 10 times

Normalization Method & Block Overlap

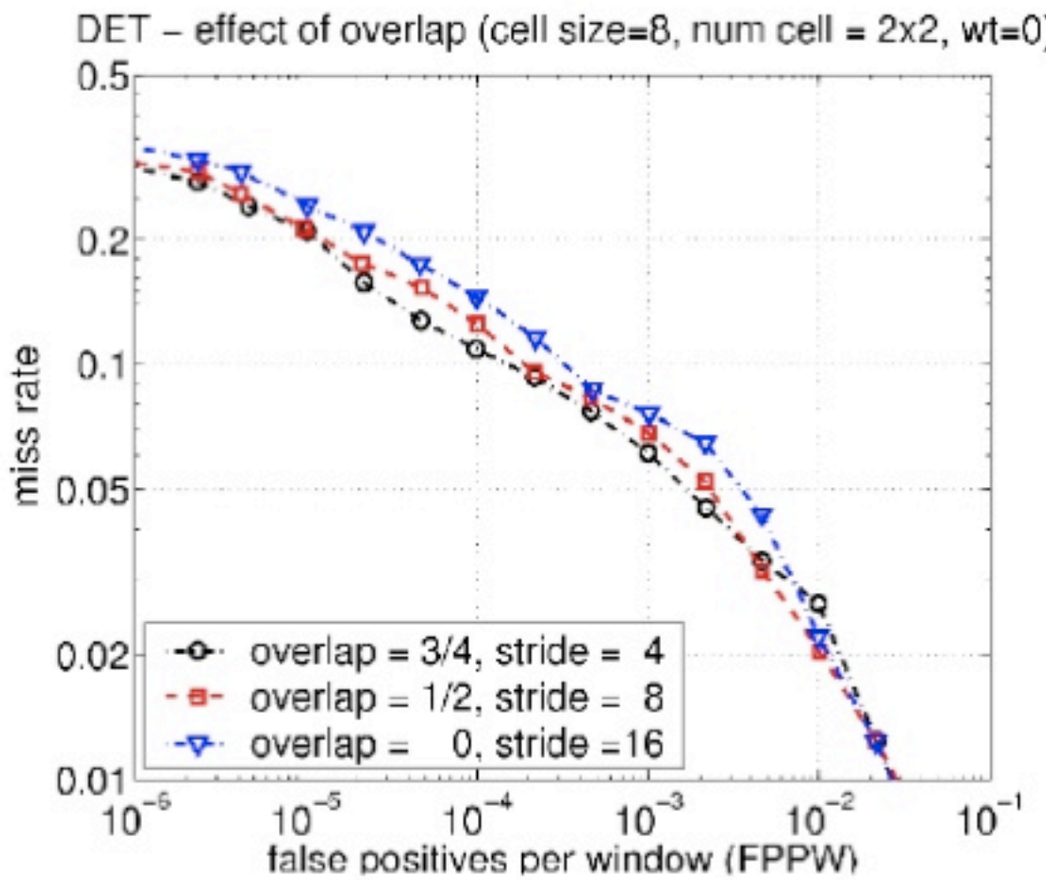
TECHNISCHE UNIVERSITÄT DARMSTADT

Normalisation method



Strong local normalisation is essential

Block overlap



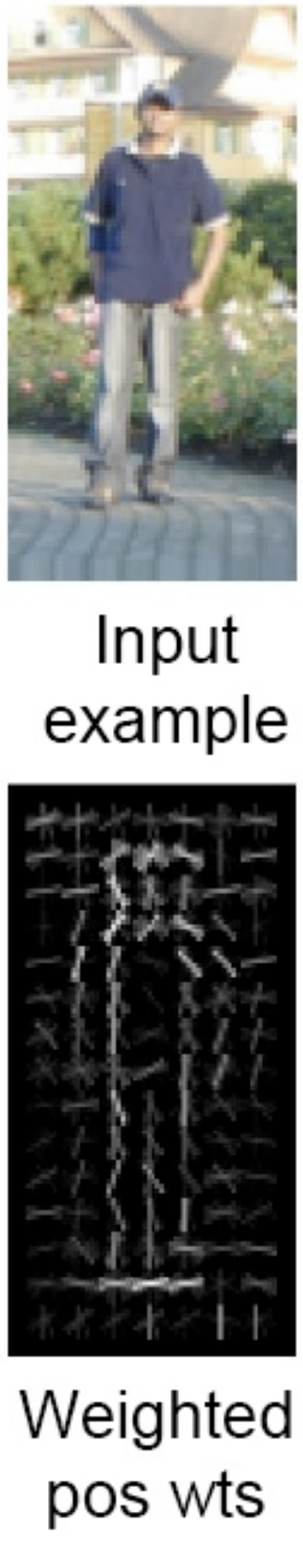
Overlapping blocks improve performance, but descriptor size increases

Descriptor Cues

TECHNISCHE UNIVERSITÄT DARMSTADT

- Most important cues:
  - Head, shoulder, leg silhouettes
  - vertical gradients inside a person are counted as negative
  - overlapping blocks just outside the contour are most important

- "Local context" use:
  - Note that best performance is obtained by including quite substantial context/background around the person



Remaining Failure Cases (for INRIA-people dataset)

TECHNISCHE UNIVERSITÄT DARMSTADT

[Wojek & Schiele, DAGM'08]

- Missing Detections:

(a) Unusual articulation (b) Difficult contrast (c) Occlusion (d) Person carrying goods

- 149 missing detections:
  - 44 difficult contrast & backgrounds
  - 43 occlusion & carried bags
  - 37 unusual articulations
  - 18 over- / underexposure
  - 7 wrong scale (too small/large)

Remaining Failure Cases (for INRIA-people dataset)

TECHNISCHE UNIVERSITÄT DARMSTADT

[Wojek & Schiele, DAGM'08]

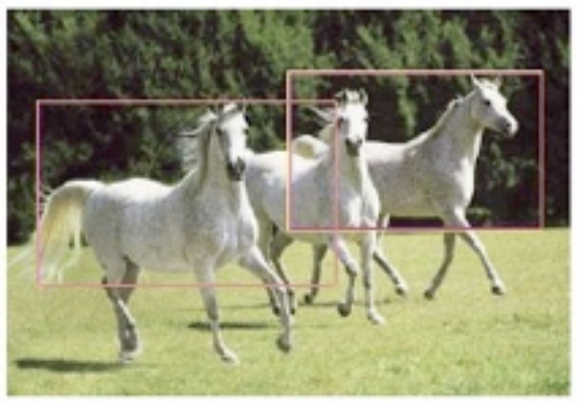
- False Positives

(e) Detection on parts (f) Too large scale (g) Detection on vertical structures (h) Cluttered background (i) Missing annotation

- 149 false positives:
  - 54 vertical structures / street signs
  - 31 cluttered background
  - 28 too small scale (body parts)
  - 24 too large scale detections
  - 12 people that are not annotated :-)

Application to other classes

TECHNISCHE UNIVERSITÄT DARMSTADT



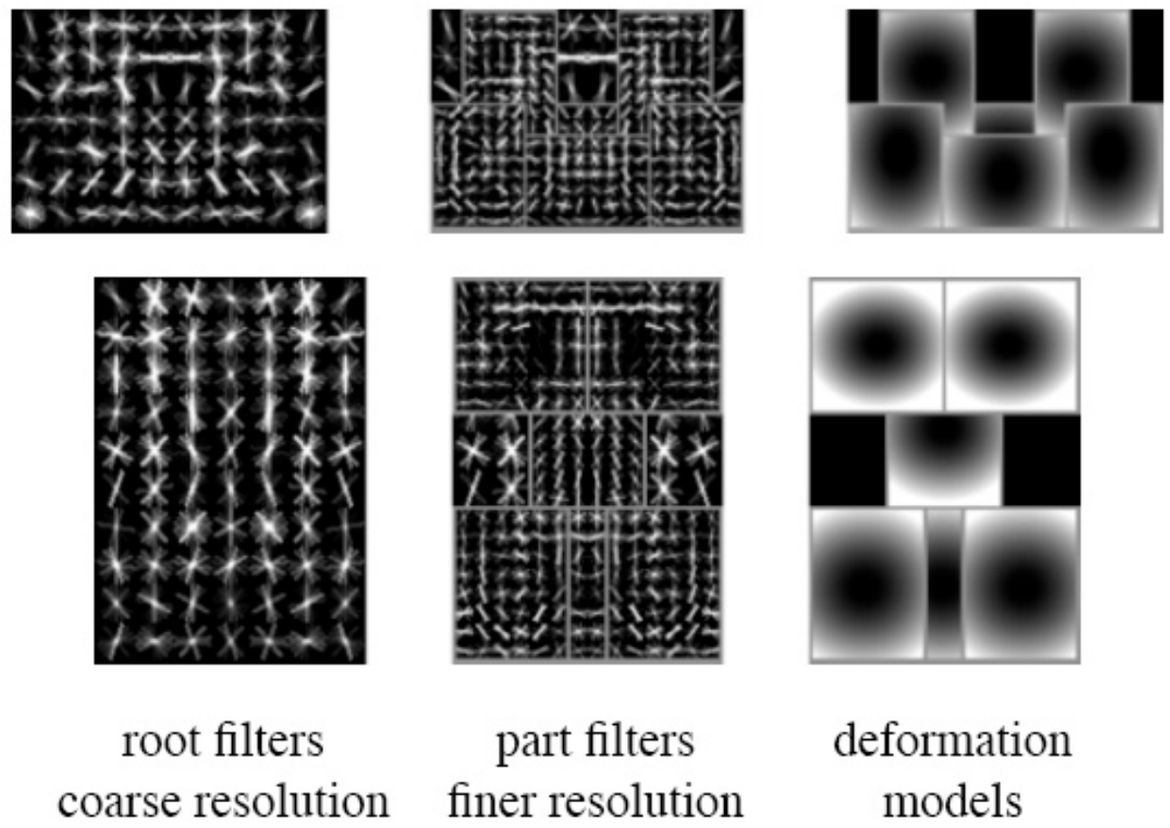
Deformable Part Model

TECHNISCHE UNIVERSITÄT DARMSTADT

- 2-scale model
  - Whole object
  - Parts

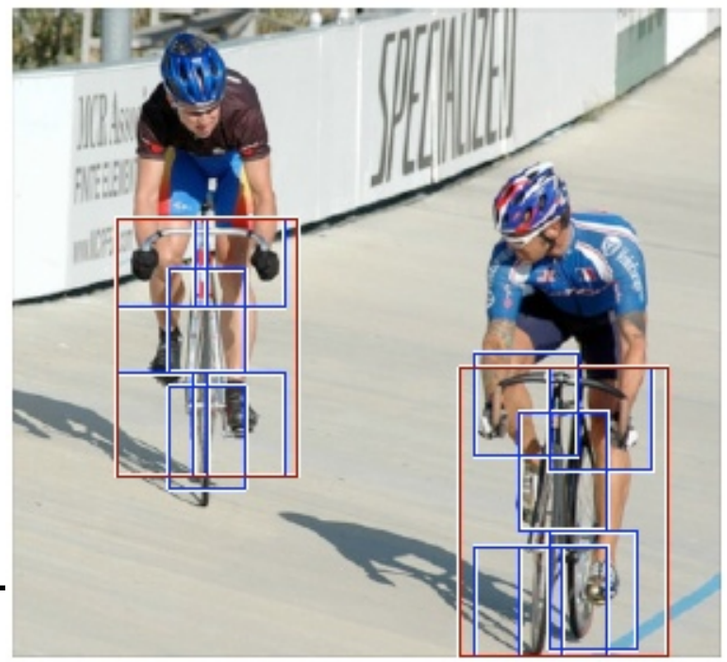
- HOG representation + SVM training to obtain robust part detectors

- Efficient algorithm for detection



root filters coarse resolution | part filters finer resolution | deformation models

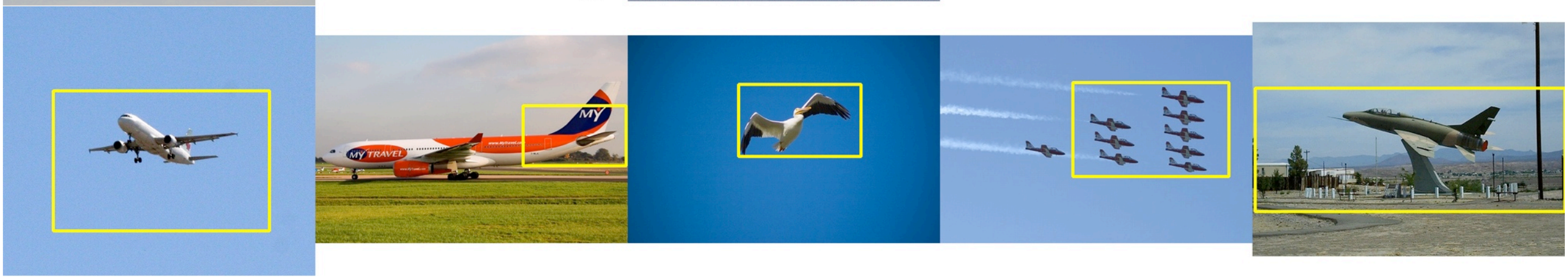
[Felzenszwalb et al.]



HOG + BoW + CRF [Schnitzspan et al, 09]

true positives

false positives



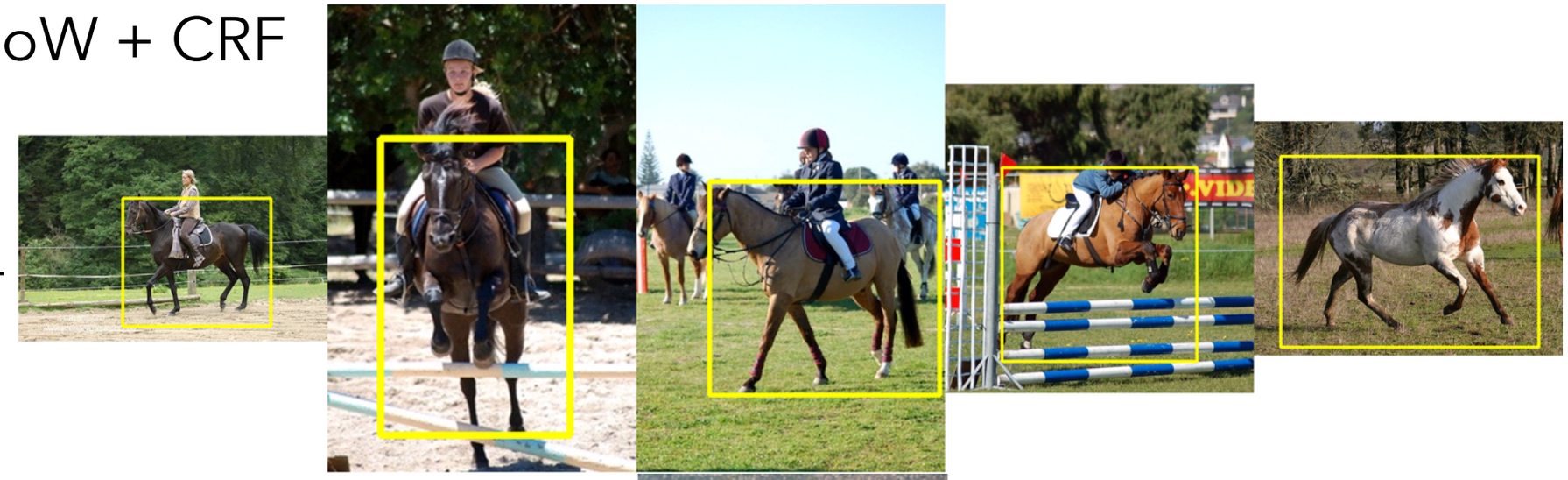
Deformable part model [Felzenszwalb et al., 08]

true positives

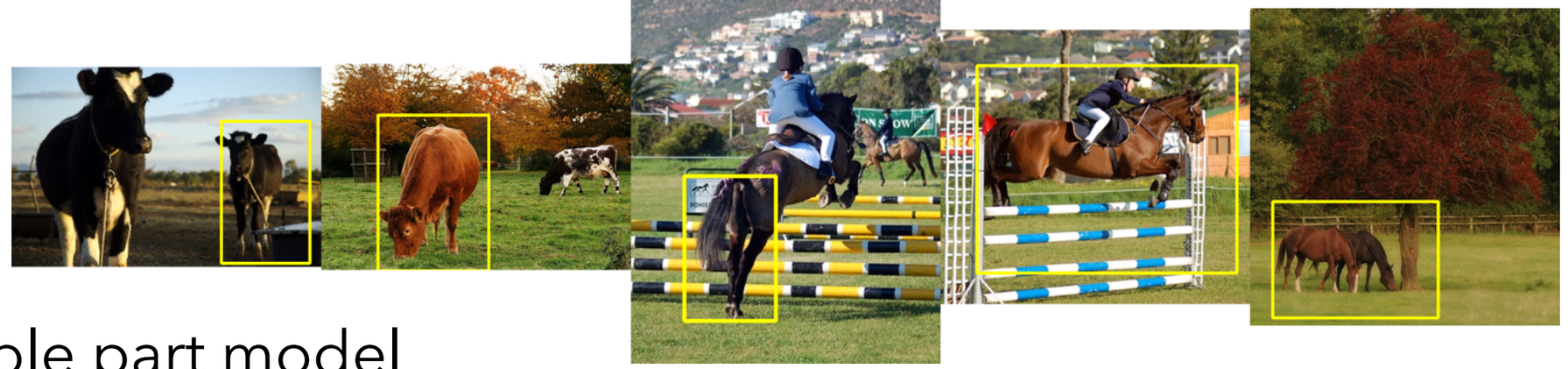
false positives

HOG + BoW + CRF

true positives

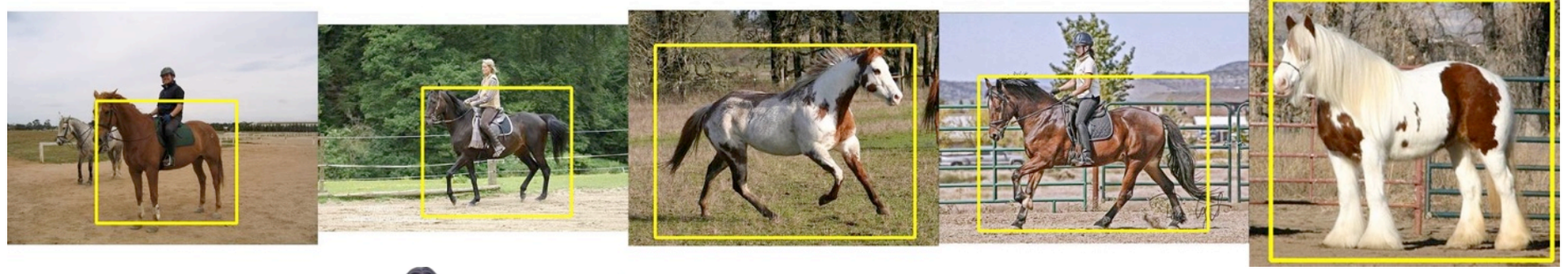


false positives



Deformable part model

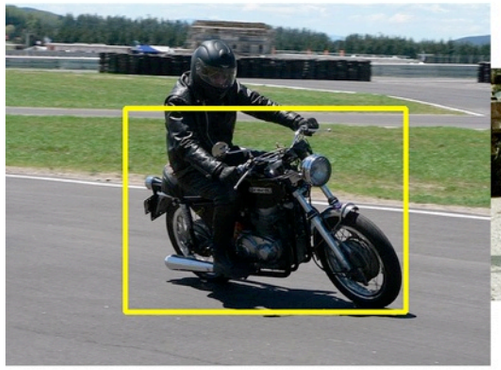
true positives

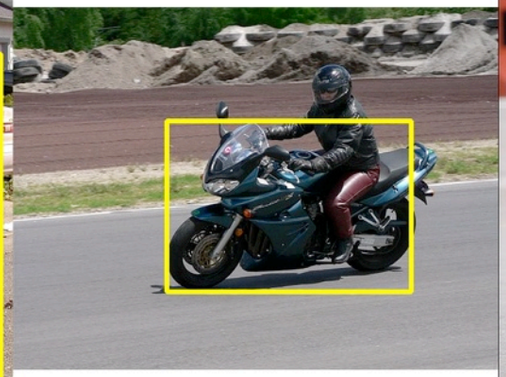
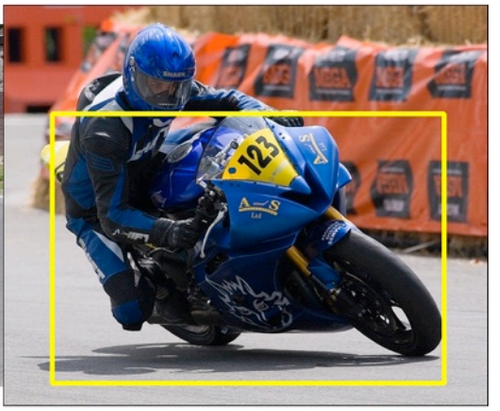


false positives

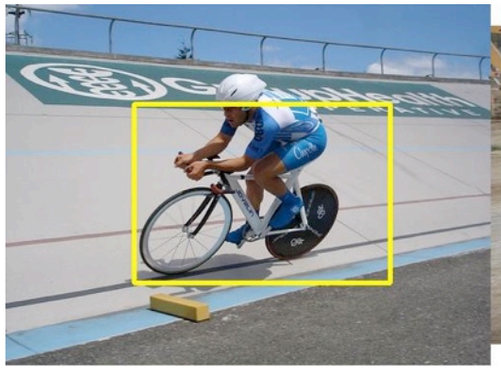
HOG + BoW + CRF [Schnitzspan et al, 09]

true positives



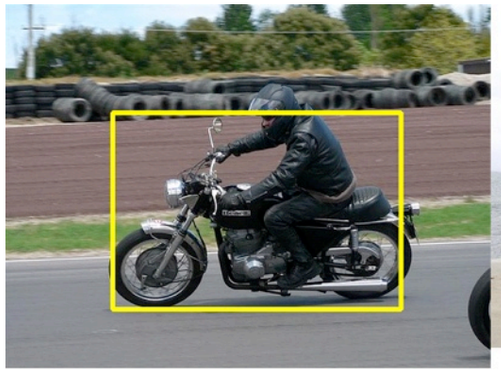
false positives

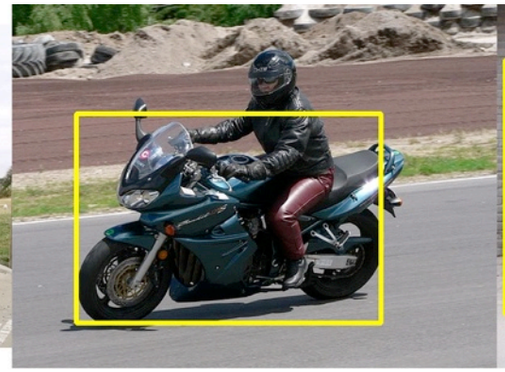


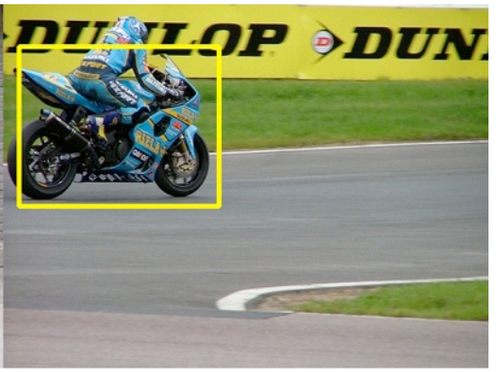


Deformable part model [Felzenszwalb et al., 08]

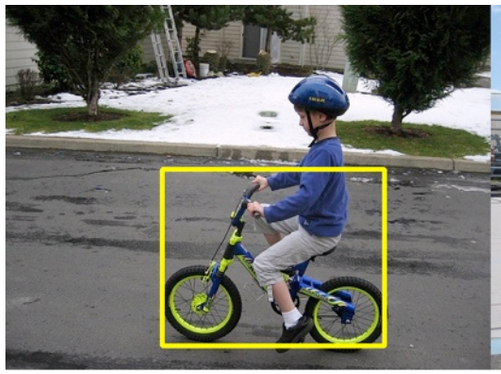
true positives

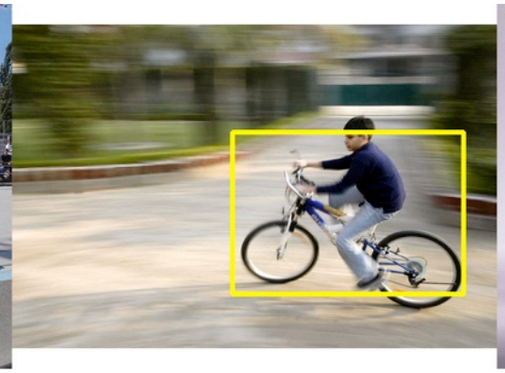
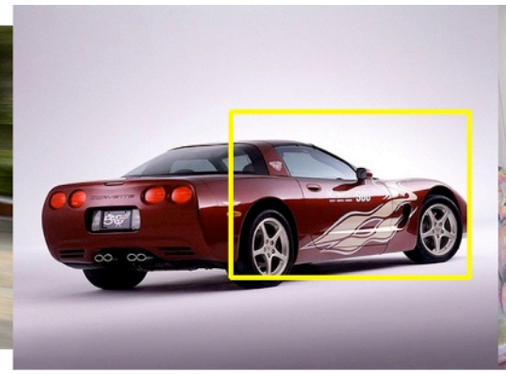






false positives




Computer Vision I

Single-View Geometry - 19.06.2013

TECHNISCHE UNIVERSITÄT DARMSTADT

with slides from:

Konrad Schindler
Svetlana Lazebnik

visual inference

Outline

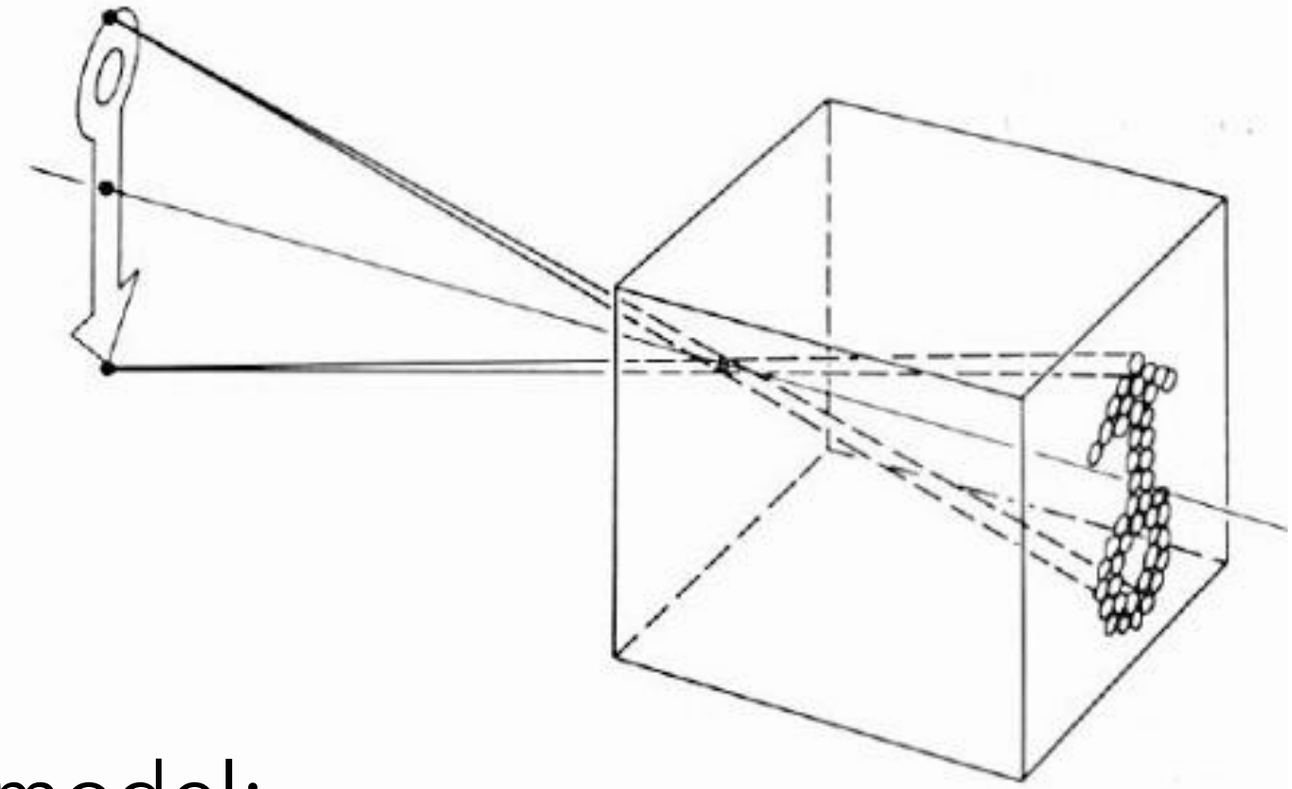
TECHNISCHE UNIVERSITÄT DARMSTADT

- Part 1: Cameras revisited

- Reminder
  - Pinhole camera
  - Projection model

- Camera calibration
  - From known 3D points

Reminder: Pinhole camera model



- Pinhole model:
  - Captures **pencil of rays** – all rays through a single point
  - Projection rays are **straight** lines
  - The point is called **center of projection (focal point)**
  - The image is formed on the **image plane**
  - Two equivalent projections: **positive, negative**

Reminder: Perspective Projection Matrix

- Projection is a matrix multiplication in homogeneous coordinates:

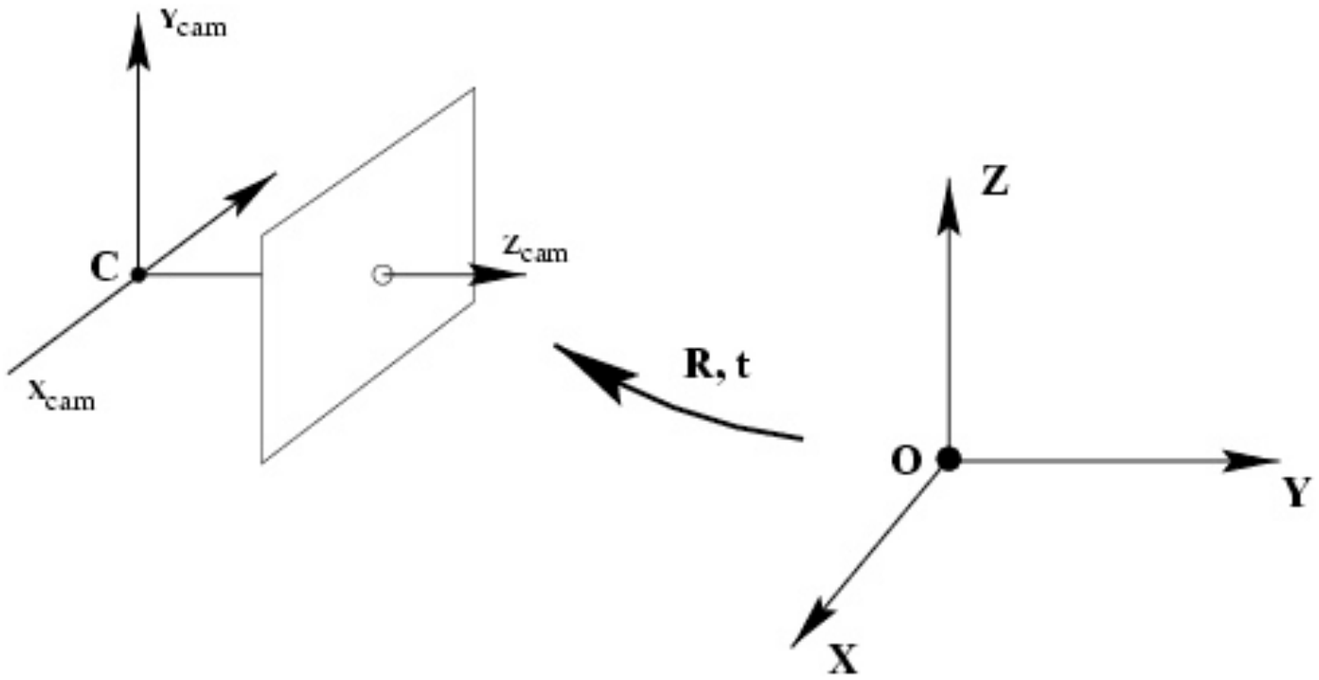
$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1/f' & 0 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix} = \begin{pmatrix} x \\ y \\ z/f' \end{pmatrix} \quad \Rightarrow \quad (f'\frac{x}{z}, f'\frac{y}{z})$$

divide by the third coordinate

- By changing the algebraic representation all relevant transformations become linear

$$\begin{pmatrix} \text{2D point (3x1)} \end{pmatrix} = \begin{pmatrix} \text{Camera to pixel coord. trans. matrix (3x3)} \end{pmatrix} \begin{pmatrix} \text{Perspective projection matrix (3x4)} \end{pmatrix} \begin{pmatrix} \text{World to camera coord. trans. matrix (4x4)} \end{pmatrix} \begin{pmatrix} \text{3D point (4x1)} \end{pmatrix}$$

Reminder: Camera rotation and translation



- the camera coordinate frame is related to the world coordinate frame by a rotation and a translation

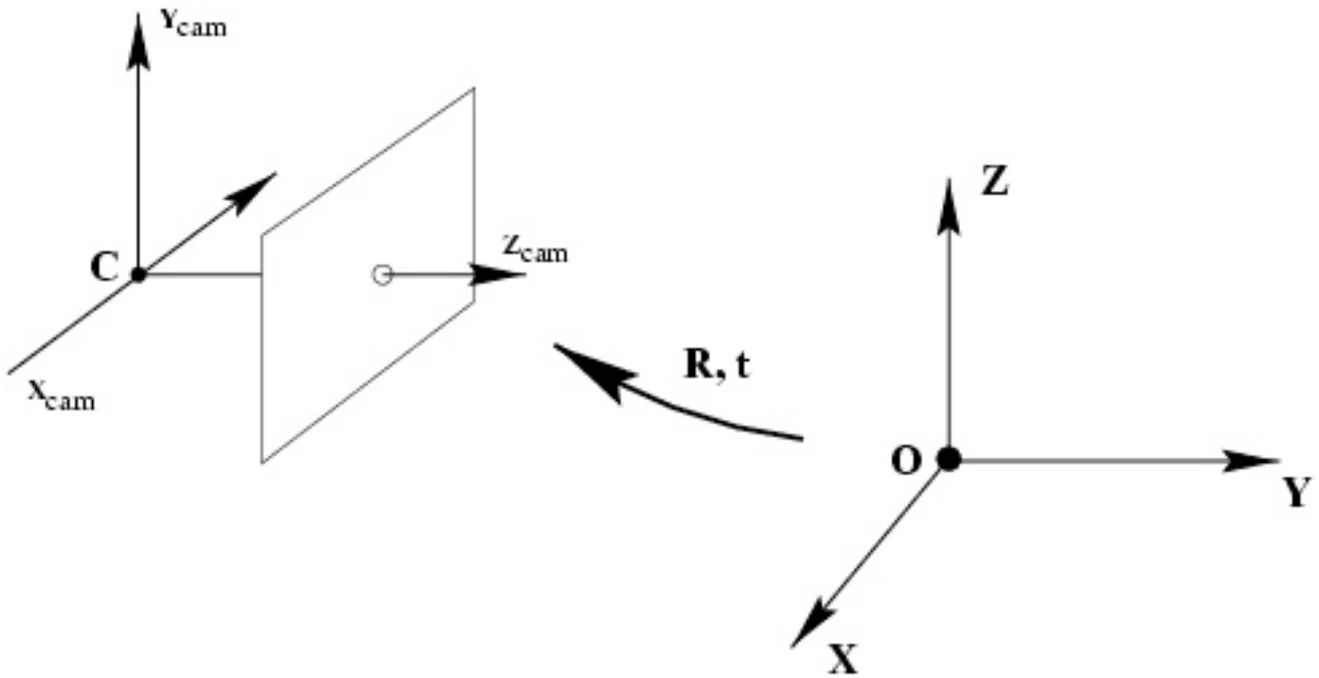
$$\tilde{X}_{cam} = R(\tilde{X} - \tilde{C})$$

coords. of point in camera frame

coords. of a point in world frame (non-homogeneous)

coords. of camera center in world frame

Reminder: Camera rotation and translation



In non-homogeneous coordinates:

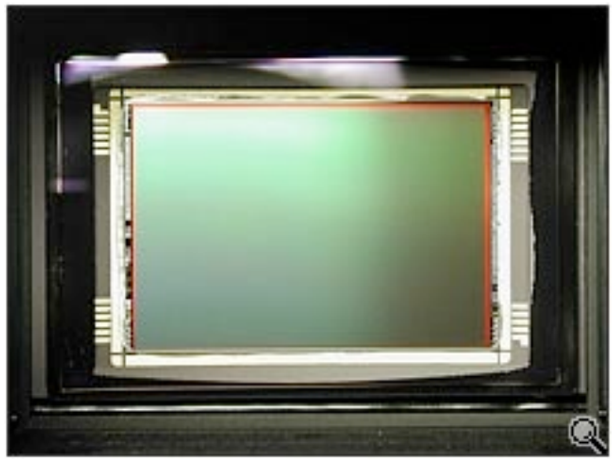
$$\tilde{X}_{cam} = R(\tilde{X} - \tilde{C})$$

$$X_{cam} = \begin{bmatrix} R & -R\tilde{C} \\ 0 & 1 \end{bmatrix} \begin{pmatrix} \tilde{X} \\ 1 \end{pmatrix} = \begin{bmatrix} R & -R\tilde{C} \\ 0 & 1 \end{bmatrix} X$$

$$x = K[I|0]X_{cam} = K[R|-R\tilde{C}]X \qquad P = K[R|t] \qquad t = -R\tilde{C}$$

Note: C is the null space of the camera projection matrix (PC=0)

Reminder: Pixel coordinates



Pixel size: $\frac{1}{m_x} \times \frac{1}{m_y}$

m_x pixels per unit (m,mm,inch,...) in horizontal direction,

m_y pixels per unit in vertical direction

$$K = \begin{bmatrix} m_x & & \\ & m_y & \\ & & 1 \end{bmatrix} \begin{bmatrix} f & & p_x \\ & f & p_y \\ & & 1 \end{bmatrix} = \begin{bmatrix} \alpha_x & & \beta_x \\ & \alpha_y & \beta_y \\ & & 1 \end{bmatrix}$$

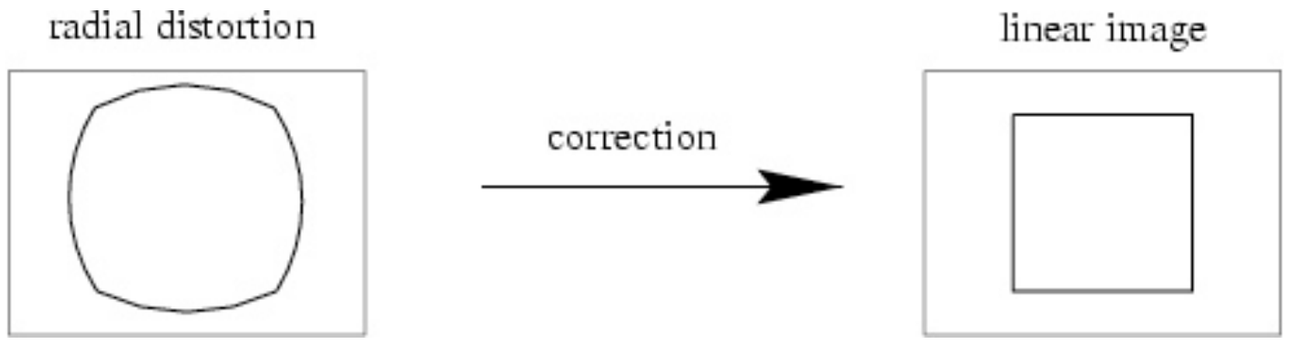
pixels/unit units pixels

Reminder: Camera parameters

TECHNISCHE UNIVERSITÄT DARMSTADT

- Intrinsic parameters
  - Principal point coordinates
  - Focal length
  - Pixel magnification factors
  - Skew (non-rectangular pixels)
  - Radial distortion

$$K = \begin{bmatrix} m_x & & \\ & m_y & \\ & & 1 \end{bmatrix} \begin{bmatrix} f & & p_x \\ & f & p_y \\ & & 1 \end{bmatrix} = \begin{bmatrix} \alpha_x & & \beta_x \\ & \alpha_y & \beta_y \\ & & 1 \end{bmatrix}$$



Reminder: Camera parameters

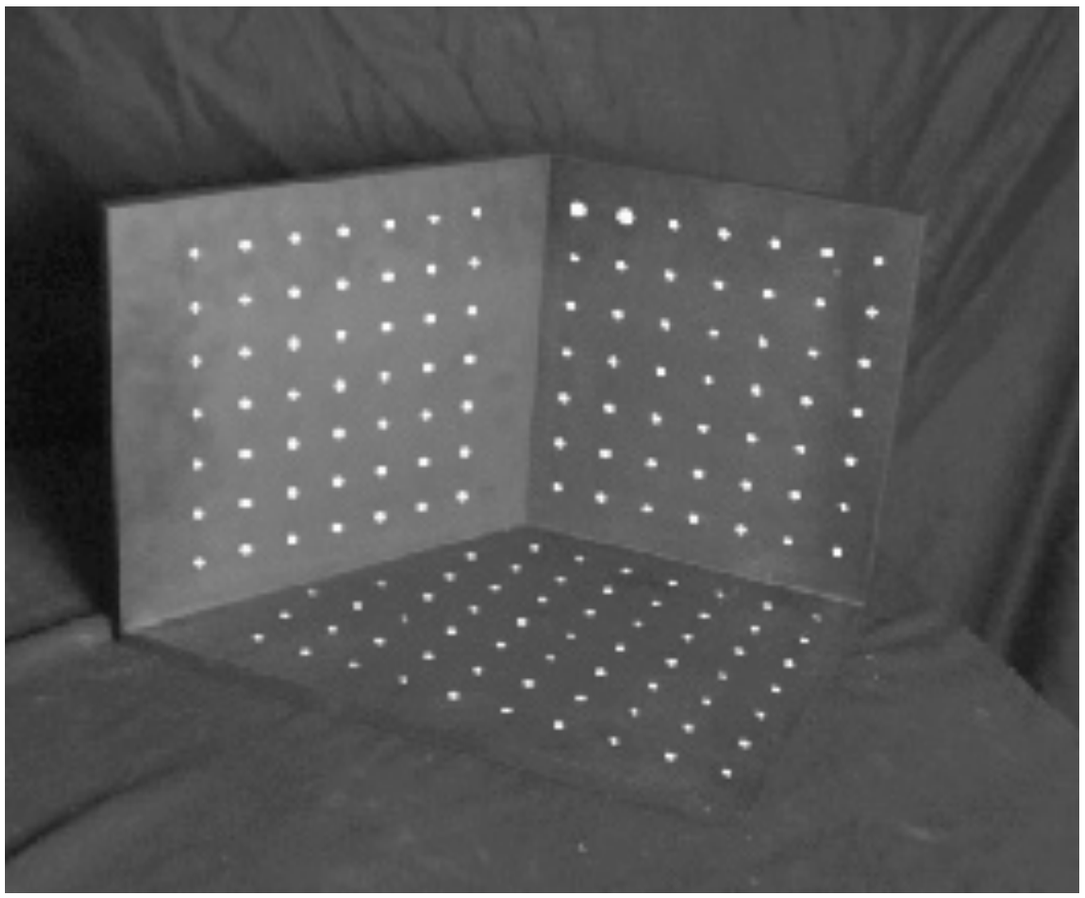
TECHNISCHE UNIVERSITÄT DARMSTADT

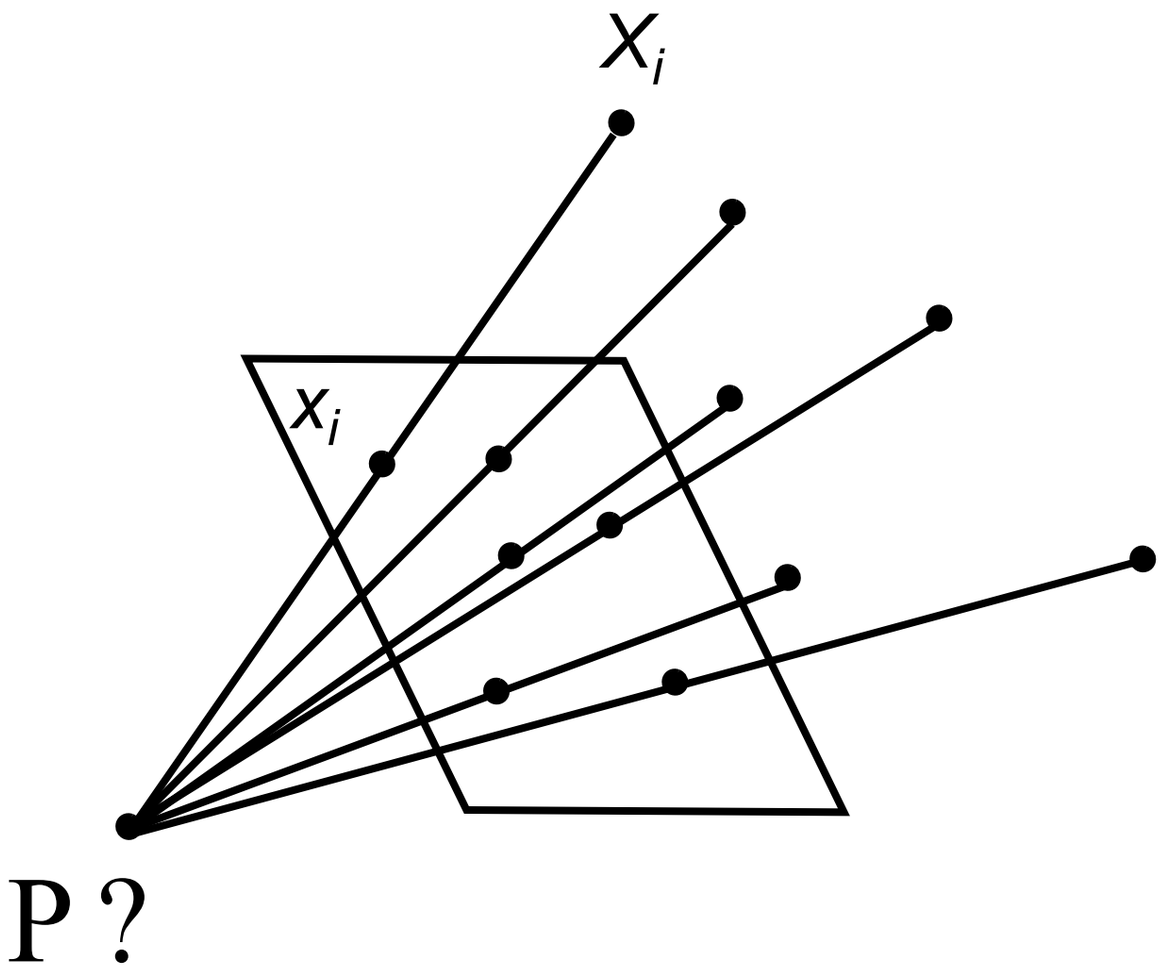
- Intrinsic parameters
  - Principal point coordinates
  - Focal length
  - Pixel magnification factors
  - Skew (non-rectangular pixels)
  - Radial distortion
- Extrinsic parameters
  - Rotation and translation relative to world coordinate system

Camera calibration

TECHNISCHE UNIVERSITÄT DARMSTADT

- Given *n* points with known 3D coordinates X_i and known image projections x_i, estimate the camera parameters

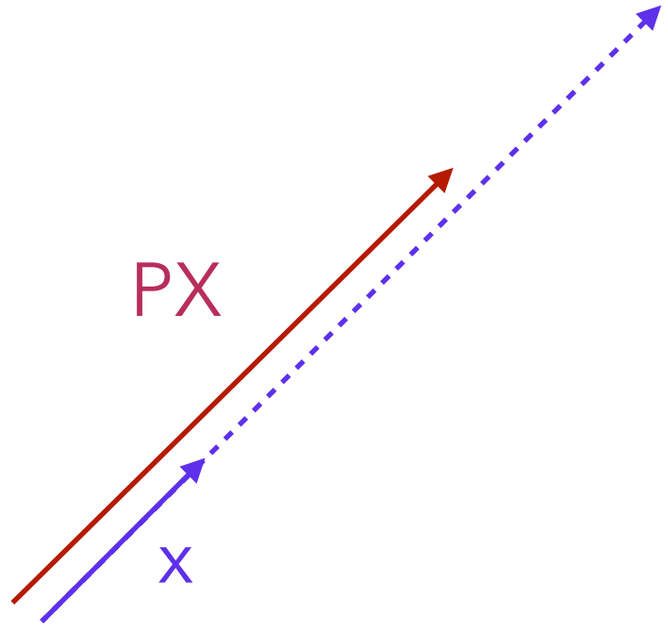




Camera calibration

TECHNISCHE UNIVERSITÄT DARMSTADT

$$\lambda \mathrm{x}_i = \mathrm{P}\mathrm{X}_i \qquad \lambda \begin{bmatrix} x_i \\ y_i \\ 1 \end{bmatrix} = \begin{bmatrix} \mathrm{P}_1^T \\ \mathrm{P}_2^T \\ \mathrm{P}_3^T \end{bmatrix} \mathrm{X}_i$$



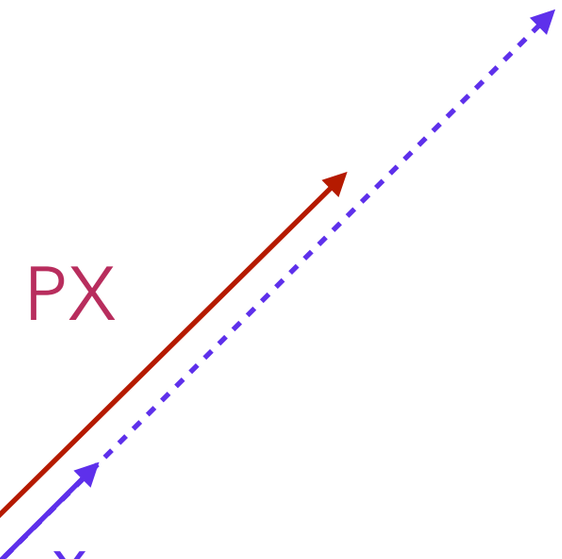
$$\mathrm{x}_i \times \mathrm{P}\mathrm{X}_i = 0$$

$$\begin{pmatrix} y_i P_3^T X_i - P_2^T X_i \\ P_1^T X_i - x_i P_3^T X_i \\ x_i P_2^T X_i - y_i P_1^T X_i \end{pmatrix} = 0$$

Camera calibration

TECHNISCHE UNIVERSITÄT DARMSTADT

$$\lambda \mathrm{x}_i = \mathrm{P}\mathrm{X}_i \qquad \lambda \begin{bmatrix} x_i \\ y_i \\ 1 \end{bmatrix} = \begin{bmatrix} \mathrm{P}_1^T \\ \mathrm{P}_2^T \\ \mathrm{P}_3^T \end{bmatrix} \mathrm{X}_i$$



$$\mathrm{x}_i \times \mathrm{P}\mathrm{X}_i = 0$$

$$\begin{pmatrix} y_i P_3^T X_i - P_2^T X_i \\ P_1^T X_i - x_i P_3^T X_i \\ x_i P_2^T X_i - y_i P_1^T X_i \end{pmatrix} = 0$$

$$\begin{bmatrix} 0 & -X_i^\top & y_i X_i^\top \\ X_i^\top & 0 & -x_i X_i^\top \\ -y_i X_i^\top & x_i X_i^\top & 0 \end{bmatrix} \begin{pmatrix} P_1 \\ P_2 \\ P_3 \end{pmatrix} = 0$$

Two linearly independent equations

Camera calibration

TECHNISCHE UNIVERSITÄT DARMSTADT

$$\begin{bmatrix} 0 & -X_i^\top & y_i X_i^\top \\ X_i^\top & 0 & -x_i X_i^\top \\ \vdots & \vdots & \vdots \\ 0 & -X_n^\top & y_n X_n^\top \\ X_n^\top & 0 & -x_n X_n^\top \end{bmatrix} \begin{pmatrix} P_1 \\ P_2 \\ P_3 \end{pmatrix} = 0 \qquad Ap = 0$$

- P has 11 degrees of freedom (12 parameters, but scale is arbitrary)
- One 2D/3D correspondence gives us two linearly independent equations
- 6 correspondences needed for a minimal solution
- more points: least-squares estimation

Camera calibration

TECHNISCHE UNIVERSITÄT DARMSTADT

- Problem: trivial solution p=0

$$Ap = 0$$
$$\text{s.t. } \|p\| = 1$$

- How to solve homogeneous least squares: find the right nullspace of A

$$UDV^\top = A$$
$$A \rightarrow [U, D, V] \qquad \text{(singular value decomposition)}$$
$$D = diag(d_1 \dots d_{12}) \qquad d_i \geq d_{i+1}$$
$$V = [v_1 \dots v_{12}] \qquad p = v_{12}$$

Introduction

We want to find a $n \times 1$ vector $\mathbf{h}$ satisfying

$$\mathbf{A}\mathbf{h} = \mathbf{0}\,,$$

where $\mathbf{A}$ is $m \times n$ matrix, and $\mathbf{0}$ is $n \times 1$ zero vector. Assume $m \geq n$, and $\mathrm{rank}(\mathbf{A}) = n$. We are obviously not interested in the trivial solution $\mathbf{h} = \mathbf{0}$ hence, we add the constraint

$$\|\mathbf{h}\| = 1\,.$$

Constrained least–squares minimization: Find $\mathbf{h}$ that minimizes $\|\mathbf{A}\mathbf{h}\|$ subject to $\|\mathbf{h}\| = 1$.

[Tomas Svoboda]

Derivation I — Lagrange multipliers

- $\mathbf{h} = \mathrm{argmin}_h \|\mathbf{A}\mathbf{h}\|$ subject to $\|\mathbf{h}\| = 1$. We rewrite the constraint as $1 - \mathbf{h}^\top \mathbf{h} = 0$
- To find an extreme (the sought $\mathbf{h}$) we must solve $\frac{\partial}{\partial \mathbf{h}} \left(\mathbf{h}^\top \mathbf{A}^\top \mathbf{A} \mathbf{h} + \lambda (1 - \mathbf{h}^\top \mathbf{h}) \right) = 0 \, .$
- We derive: $2\mathbf{A}^\top \mathbf{A}\mathbf{h} - 2\lambda \mathbf{h} = 0$.
- After some manipulation we end up with: $(\mathbf{A}^\top \mathbf{A} - \lambda \mathbf{E})\mathbf{h} = 0$ which is the characteristic equation. Hence, we know that $\mathbf{h}$ is an eigenvector of $(\mathbf{A}^\top \mathbf{A})$ and λ is an eigenvalue.
- The least-squares error is $e = \mathbf{h}^\top \mathbf{A}^\top \mathbf{A} \mathbf{h} = \mathbf{h}^\top \lambda \mathbf{h}$.
- The error will be minimal for $\lambda = \min_i \lambda_i$ and the sought solution is then the eigenvector of the matrix $(\mathbf{A}^\top \mathbf{A})$ corresponding to the smallest eigenvalue.

[Tomas Svoboda]

Derivation II — SVD

- Let $\mathtt{A} = \mathtt{U}\mathtt{S}\mathtt{V}^\top$, where $\mathtt{U}$ is $m \times n$ orthonormal, $\mathtt{S}$ is $n \times n$ diagonal with descending order, and $\mathtt{V}^\top$ is $n \times n$ also orthonormal.
- From orthonormality of $\mathtt{U}, \mathtt{V}$ follows that $\|\mathtt{U}\mathtt{S}\mathtt{V}^\top\mathbf{h}\| = \|\mathtt{S}\mathtt{V}^\top\mathbf{h}\|$ and $\|\mathtt{V}^\top\mathbf{h}\| = \|\mathbf{h}\|$.
- Substitute $\mathbf{y} = \mathtt{V}^\top\mathbf{h}$. Now, we minimize $\|\mathtt{S}\mathbf{y}\|$ subject to $\|\mathbf{y}\| = 1$.
- Remember that $\mathtt{S}$ is diagonal and the elements are sorted descendently. Than, it is clear that $\mathbf{y} = [0, 0, \ldots, 1]^\top$.
- From substitution we know that $\mathbf{h} = \mathtt{V}\mathbf{y}$ from which follows that sought $\mathbf{h}$ is the last column of the matrix $\mathtt{V}$.

[Tomas Svoboda]

Camera calibration

TECHNISCHE UNIVERSITÄT DARMSTADT

$$\begin{bmatrix} 0 & -X_i^\top & y_i X_i^\top \\ X_i^\top & 0 & -x_i X_i^\top \\ \vdots & \vdots & \vdots \\ 0 & -X_n^\top & y_n X_n^\top \\ X_n^\top & 0 & -x_n X_n^\top \end{bmatrix} \begin{pmatrix} P_1 \\ P_2 \\ P_3 \end{pmatrix} = 0 \qquad Ap = 0$$

- Note: for coplanar points that satisfy $\Pi^T X=0$, we will get degenerate solutions $(\Pi,0,0)$, $(0,\Pi,0)$, or $(0,0,\Pi)$

Camera calibration

TECHNISCHE UNIVERSITÄT DARMSTADT

- Once we've recovered the values of the camera matrix, we still have to figure out the intrinsic and extrinsic parameters
- decompose the left (3x3)-submatrix into an upper triangular and an orthonormal matrix

$$P = [M|m] = [KR| - KR\tilde{C}]$$
$$M \rightarrow [K, R] \qquad \text{(RQ-decomposition)}$$
$$PC = 0 \rightarrow C \qquad \text{(SVD)}$$

(if required, extract 3 rotation axis and angle from R - its columns are the axes of the rotated coordinate system)