Human Activity and Vision Summer School

visual representation of people

background subtraction

jean-marc odobez

01.10.2012

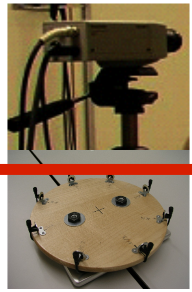
IDIAP RESEARCH INSTITUTE

EPFL
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

the overall goal: to infer relevant information from audio-visual human scenes



audio-visual scenes



representation (what is a person?)

detection (are there any people?)

localization (where are they?)

tracking (where do they go?)

identification (who are they?)

activity recognition & discovery (what do they do? what do they look at?, do they interact? who do they interact with? what do they do together?, …)

outline

- introduction
  - why is observing people from visual sensors difficult?

- standard methods for visual representation of people

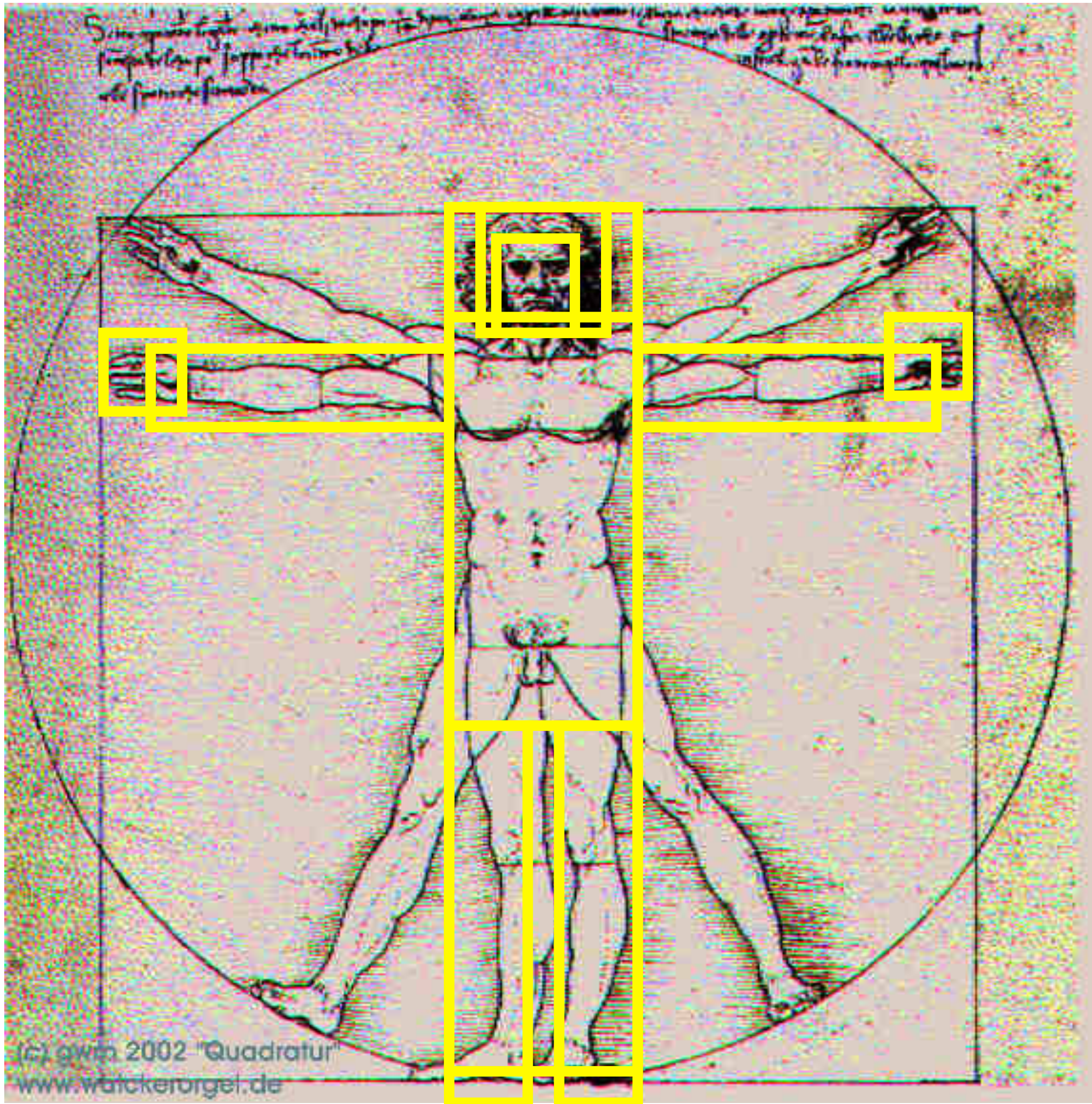
  how to build observation models $p(Y|X)$
  - contour-based
  - patch-based

  - blob-based

    => background subtraction

Note: detector based approach addressed in other talks

what is a person?



head

face

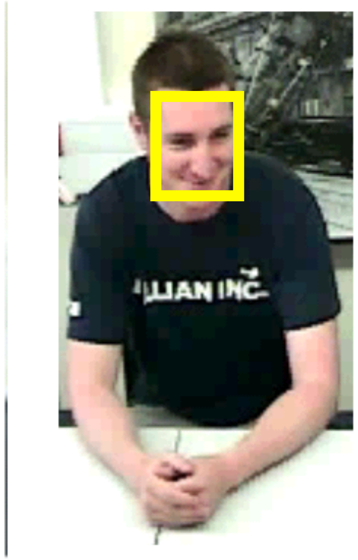
hands

body

full body

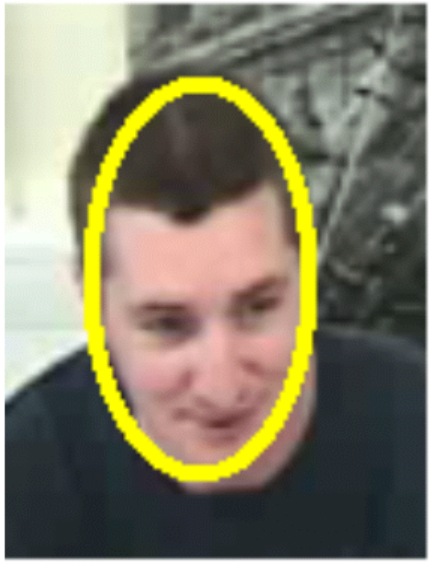
© I. da vinci

what is a person? (2)



a patch

(appearance, color, texture)

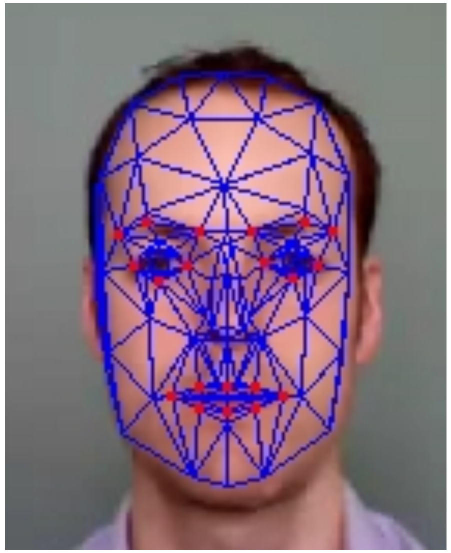


a contour

(set of points)

a blob

(binary mask)



a mesh

(set of points)

why is measuring people difficult?

- observations are **noisy**
  - image processors/ detectors are not perfect

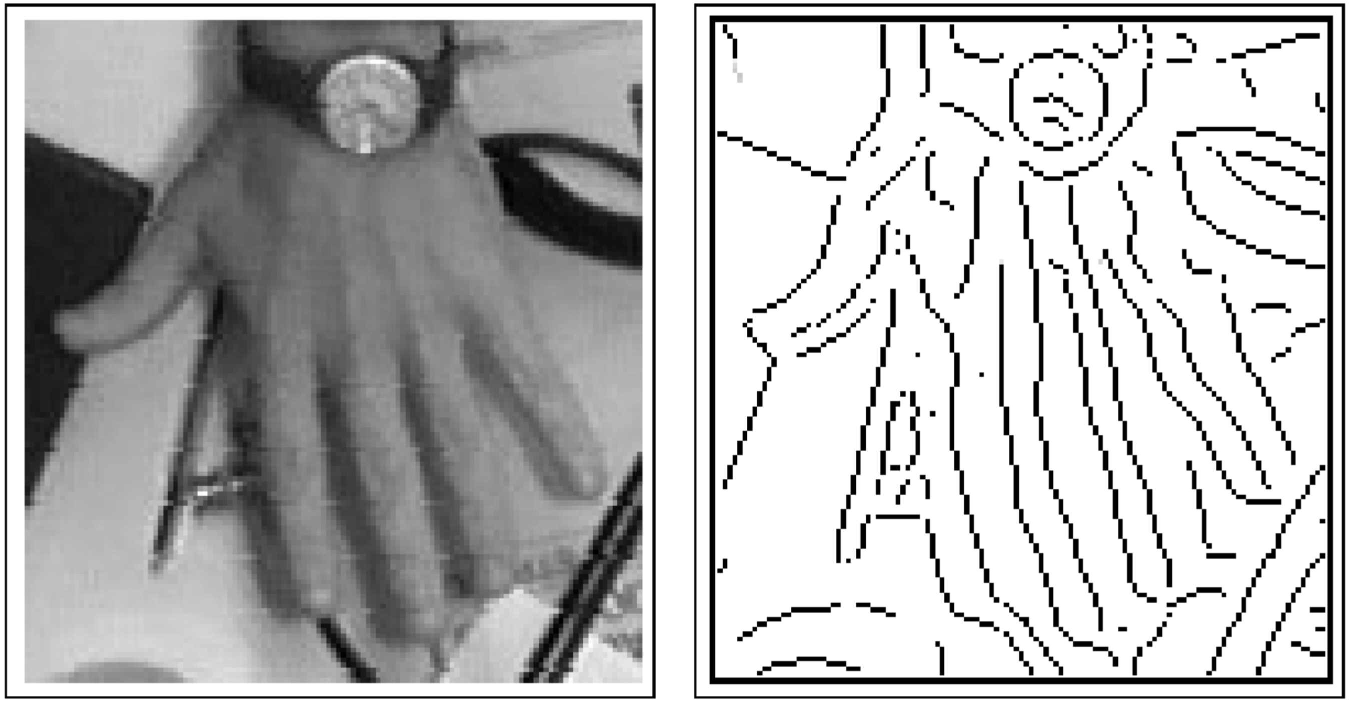
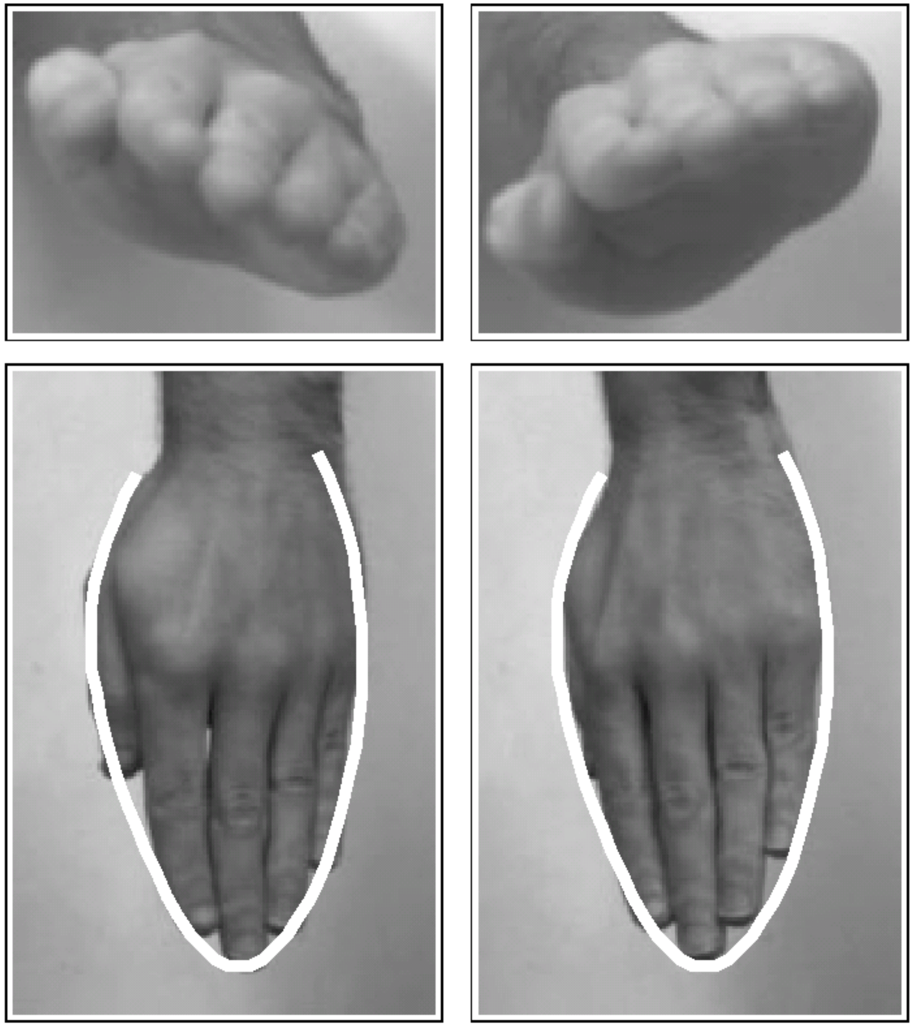


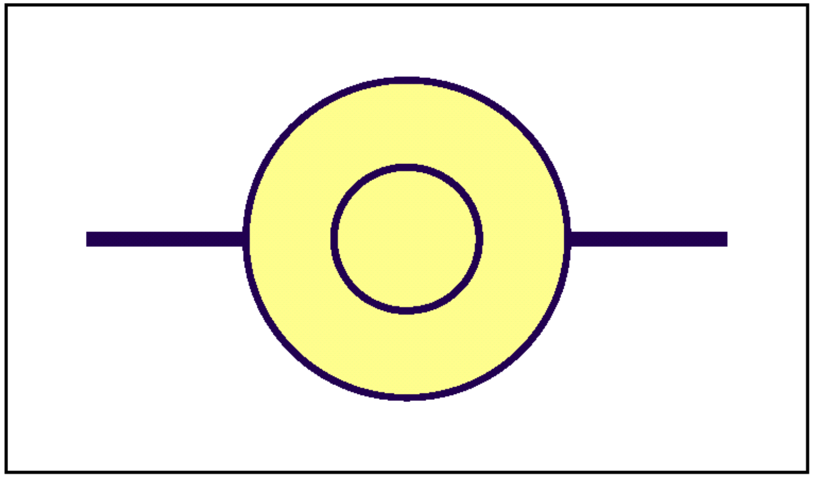
Figure 2.2: Detecting edges. *Edges (right) are generated from the image (left) using horizontally and vertically oriented masks and a decision process (Canny, 1986) that attempts to repair gaps. Nonetheless, there are breaks at critical locations such as corners or junctions, and spurious fragments that disrupt the topology of the hand.*

why is measuring people difficult? (2)

- observations are **ambiguous**
  - multiple configurations can explain the observations



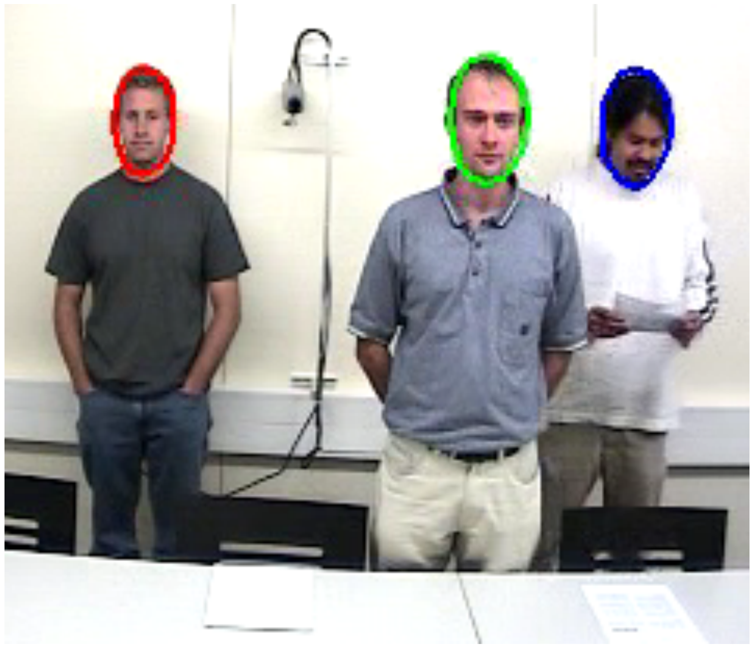
ambiguous observations…



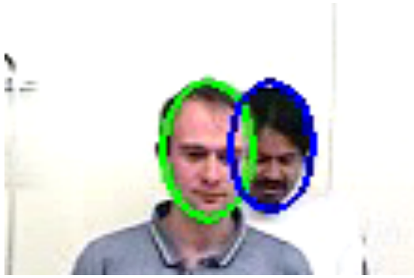
what is this?

why is measuring people difficult? (3)

- observations are **incomplete**
  - (self) occlusion



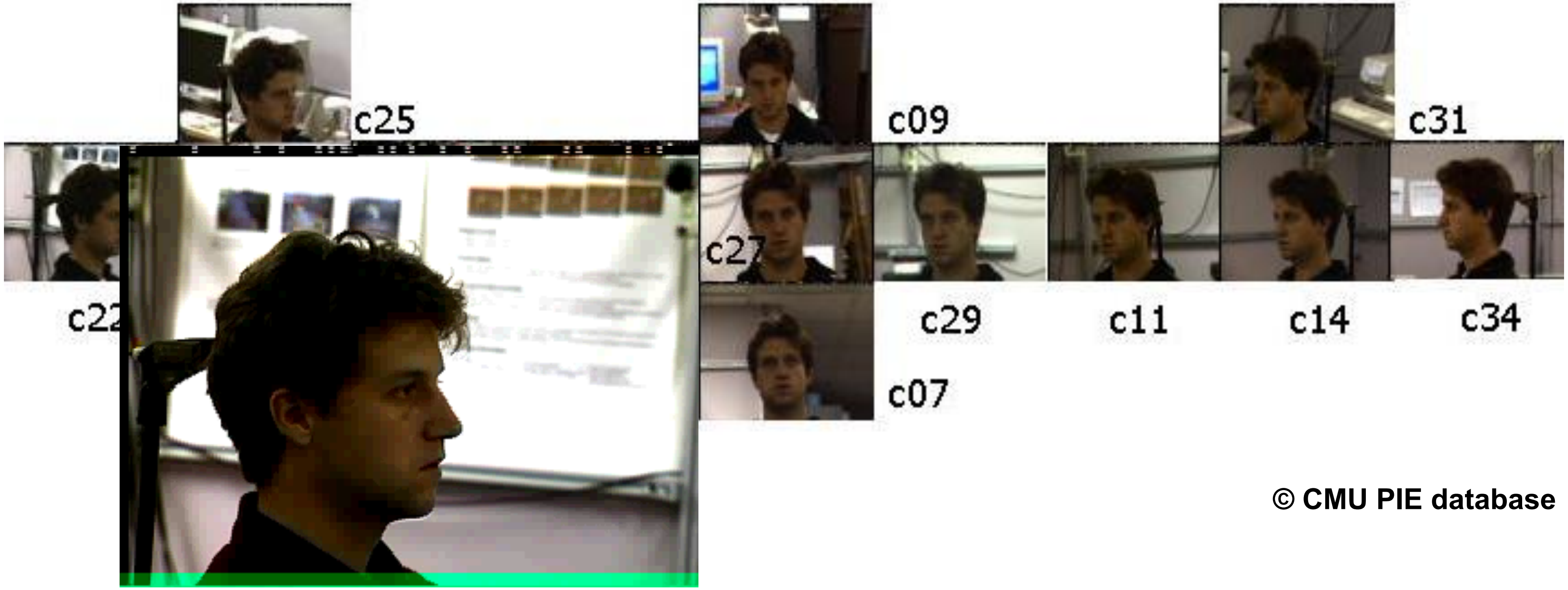
all faces are visible



not quite

why is measuring people difficult? (4)

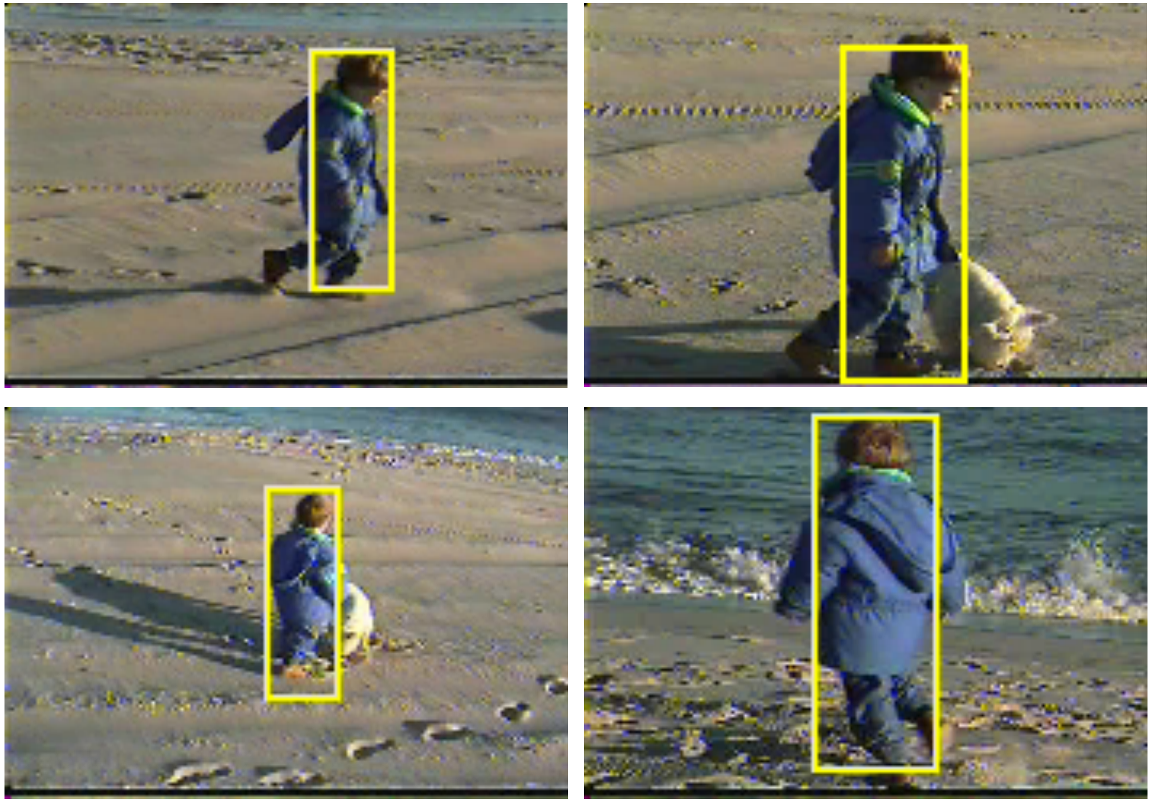
- observations have large **intra-class variation**
  - the same person can look very different under different conditions (illumination, pose, clothing)
  - people look different from each other



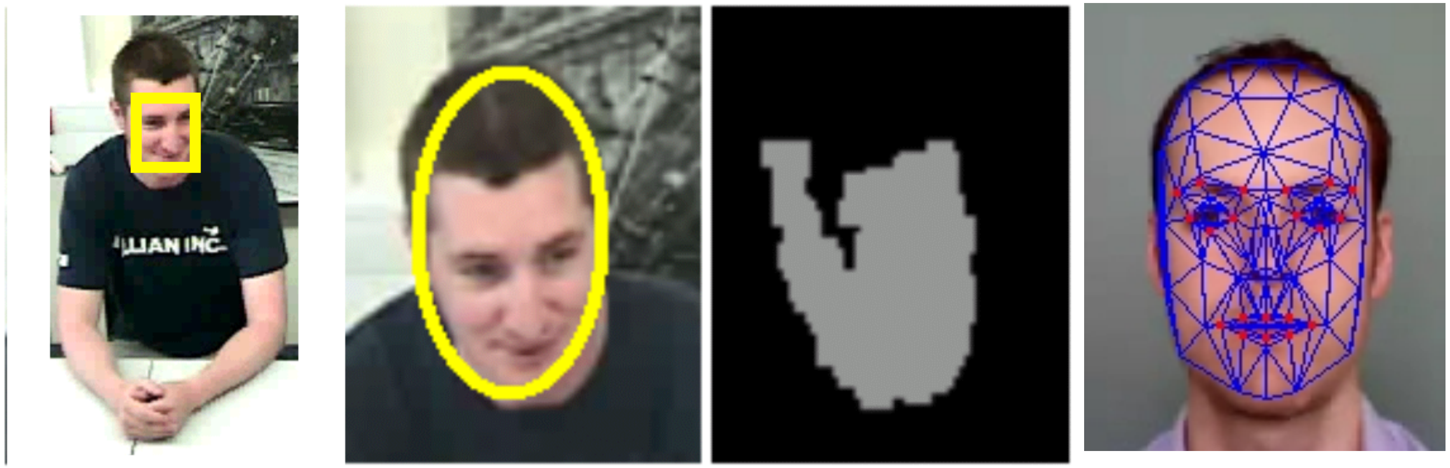
© CMU PIE database

why is measuring people difficult? (4)

- observations have large variation due to **motion**
  - combos of translation, scale, non-rigid transformations



person model: ‘template’ + set of attributes

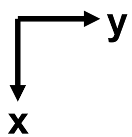


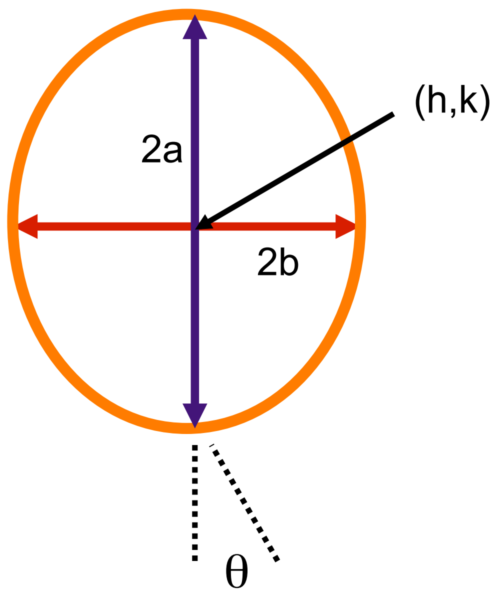
representation type	possible attributes
patch	vector of pixel values
	transformation of pixel values (e.g. PCA)
	distribution of pixel values (histogram)
contour	set of reference points
	contour parameters
blob	geometric moments
mesh	set of mesh vertices

case 1:
contour-based visual representation

contour-based representation

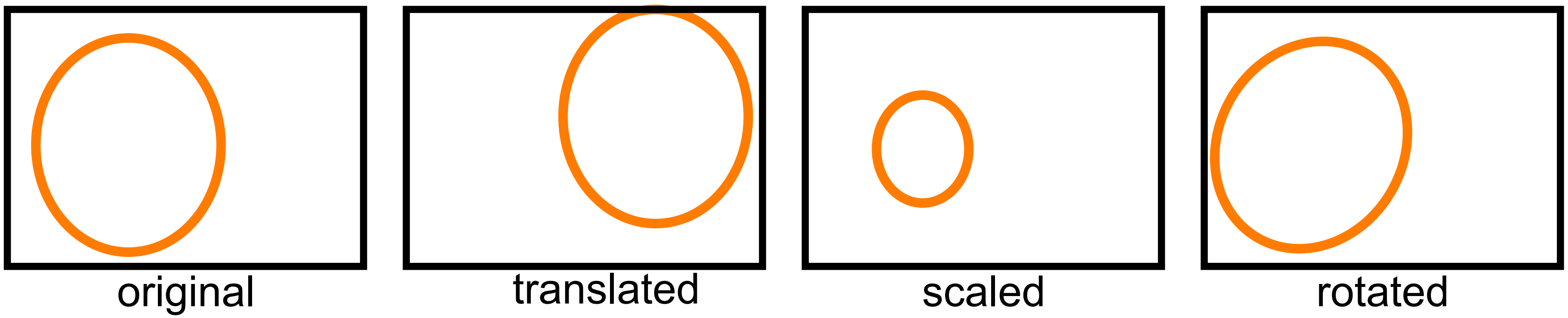
- assume a simple **parametric shape** S
  - bounding box: characterized by its center, width, height
  - ellipse: characterized by its center, main axes, orientation





shape-space: template + transformation

- **shape space**
  - class of geometric transformations **X** (rigid motion) applied to template shape
  - euclidean transformations (4 parameters)
    - translation, rotation, isotropic scaling



S **X**(S)

shape-space (2)

- this concept can be generalized for arbitrary contours represented by splines (Blake and Isard)

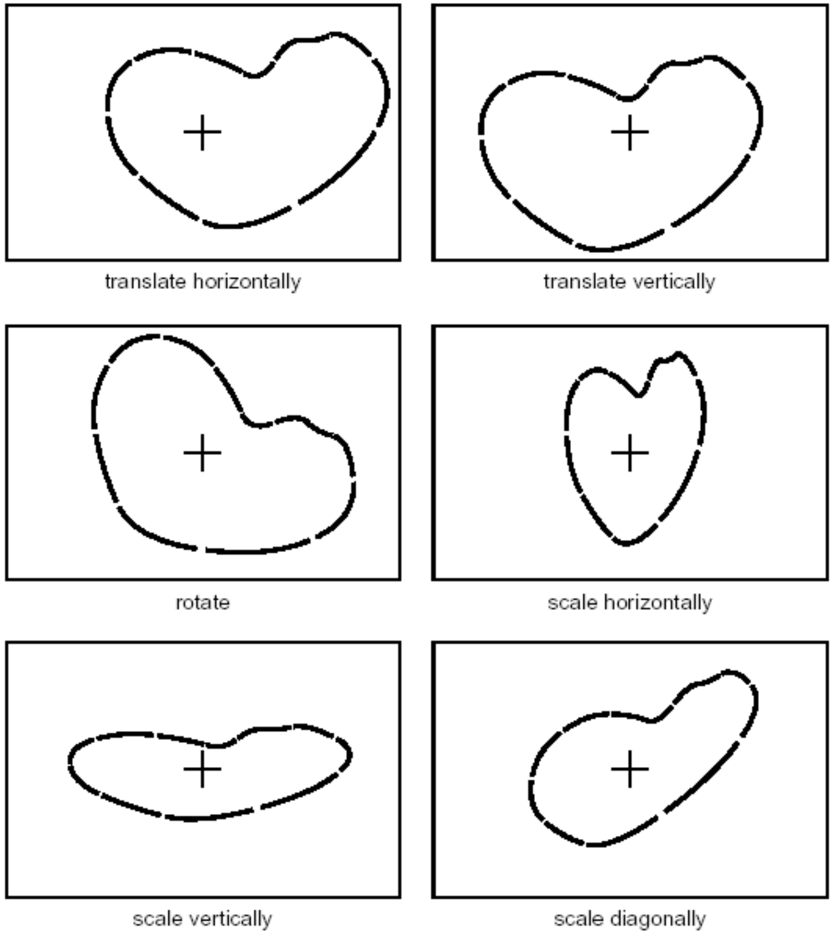


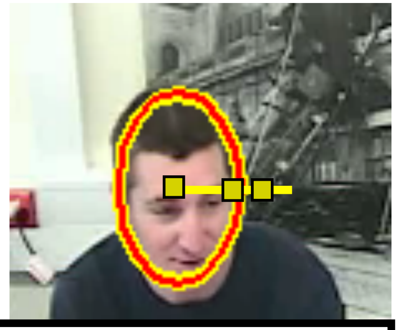
Figure 4.3: **Planar affine basis.** *The planar affine transformation group has 6 degrees of freedom. A basis for them is depicted here, as applied to the pretzel outline from figure 4.2 on page 71. The first three elements of the basis correspond to the first three for the Euclidean similarities. The last three elements span a subspace that includes the fourth element — scaling — for the Euclidean similarities and two further degrees of freedom for directional scaling. Directional scaling occurs when a planar object, initially co-planar with the image, is allowed to rotate about an axis that lies parallel to the image plane.*

shape-based observation model (1)

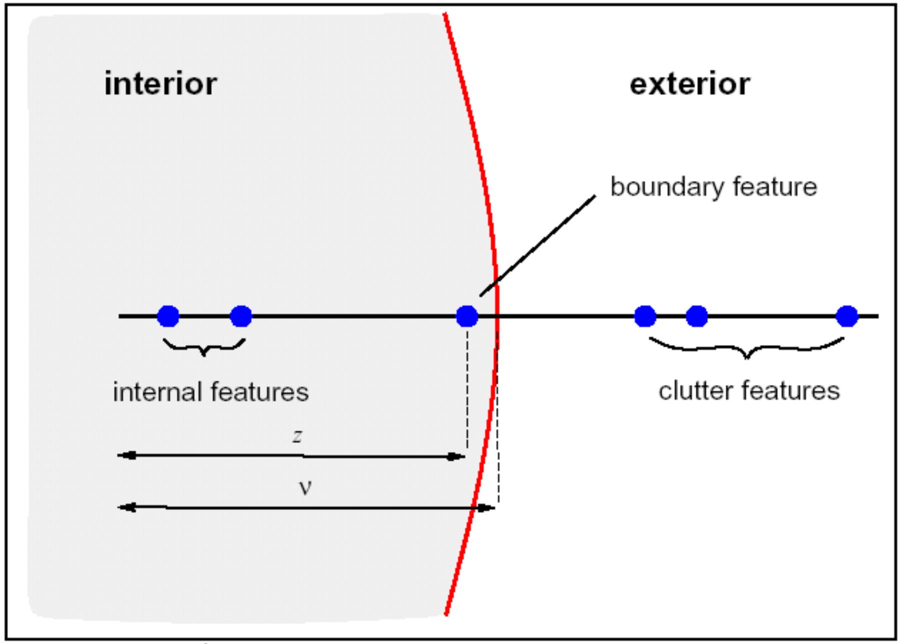
- model: bent wire immersed in clutter
- observations: detected edges along L normal lines $\mathrm{y}^{l} = \{v_{m}^{l}\}$

shape-based observation model (2)

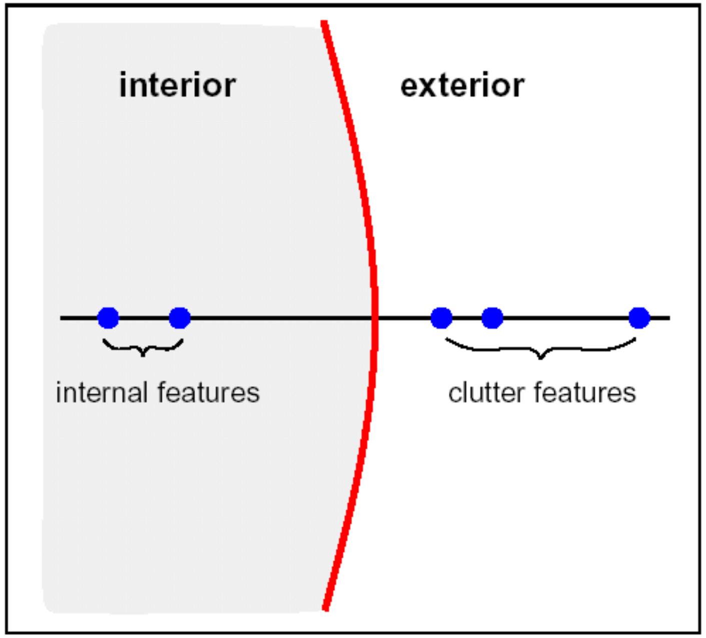
- observations: detected edges along normal



this case should give high response	this case should not



person boundary present and detected

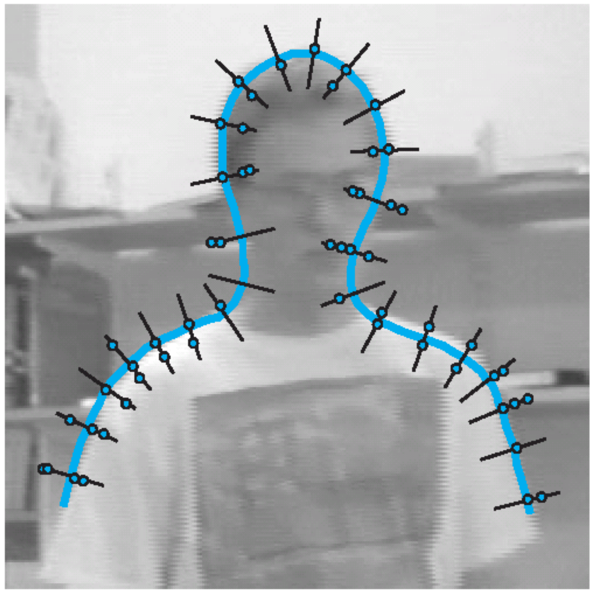


person boundary present and not detected

shape-based observation model (3)

- assumption: conditional independence on 1-D measurements
- an observation model p(*Y*|*X*) can be defined

image observations state (configuration in shape-space)



© a. blake & m. isard

$$\mathrm{y}^l = \{v_m^l\}$$

1-D positions of detected edges

$$p(Y \mid X) \propto \prod_{l=1}^{L} p(\mathrm{y}^l \mid X)$$

$$\propto \prod_{l=1}^{L} \max\left(K, \exp\left(-\frac{\|\hat{v}_m^l - v_0^l\|^2}{2\sigma^2}\right)\right)$$

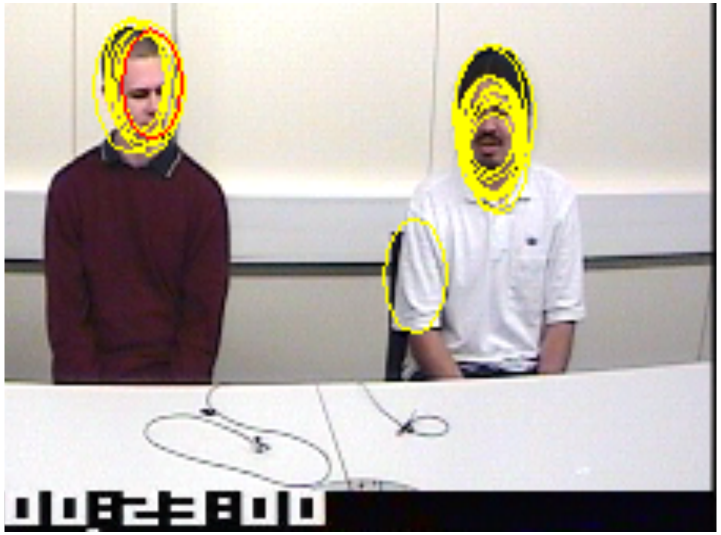
factor accounting for cases with no measurements

closest 1-D measurement to the contour

1-D position of the contour

limitations of the shape observation model

- often, high response in presence of clutter



30 samples from area around heads

head motion: translation + scaling

1000 samples from a uniform prior

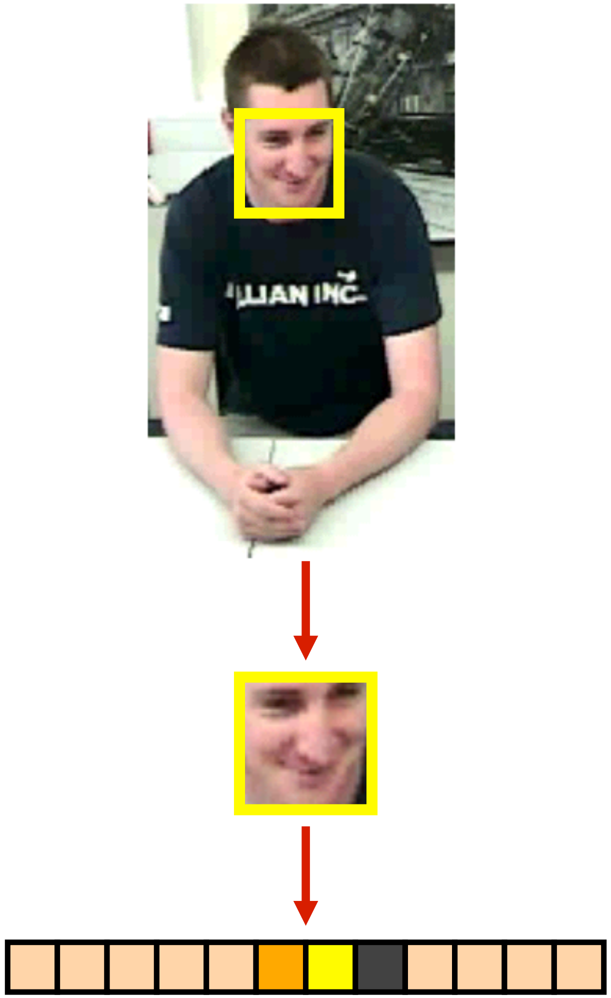
head motion: planar affine

- remedy: combine with other cues

Case 2:
patch-based visual representation

raw patches

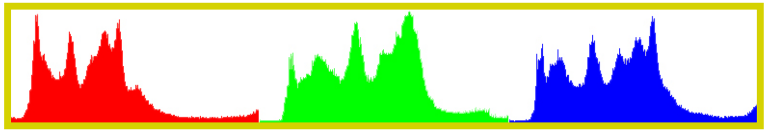
- **observations:** pixels inside template are concatenated in a vector
- templates to be compared need to be made of same size
- preprocessing needed (e.g. normalization) to build in some invariance to illumination
- comparing patches is straightforward
- not very robust if person's appearance varies (e.g. due to non-rigid motion)



color histograms

- **observations:** the person's color histogram of their template
- **non-parametric** estimate of color distribution
  - simple to compute
  - robust to many factors
  - discards all spatial structure (bag-of-colors)
  - joint vs.marginal histogram

© eastman kodak



- modeling elements
  - choice of color space
  - binning size
    - large enough (generalization)
    - Small enough (discrimination)

color histograms (2)

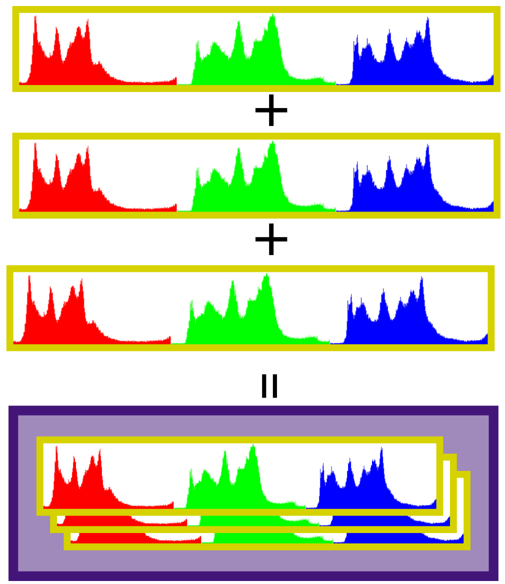
- **variation**: add some spatial structure: histogram-by-parts

- improves discrimination of the model
- parameters increase linearly with the number of parts

More thorough on descriptors for re-identification task

S. Bak : Friday talk 4pm

© eastman kodak



comparing color histograms

- use distributional measures
- **Bhattacharyya distance**

$$\rho_{BT} = \rho(Y_1, Y_2) = \sum_b (Y_1(b) Y_2(b))^{1/2}$$

histogram 1 histogram 2

© eastman kodak

$$d\left(X_T, X\right) = \sqrt{1 - \rho_{BT}(Y(X_T), Y(X))}$$

- observation model

$$p(Y \mid X) \propto e^{-\lambda\, d\left(X_T, X\right)}$$

how peaked the observation model is

localizing people with color histograms

Case 3:

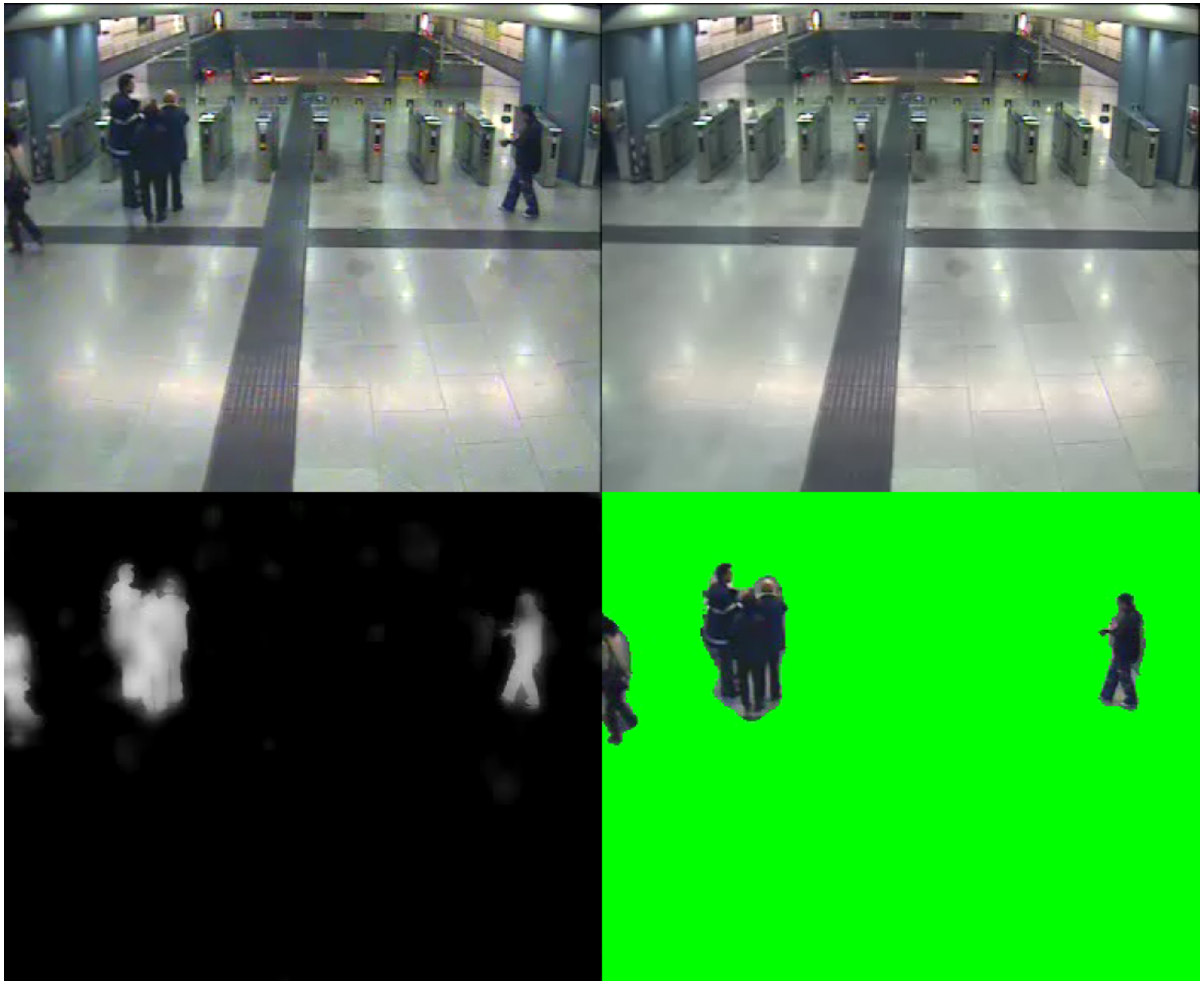
blob-based visual representation

=> background subtraction

Background subtraction

Detect foreground object by comparing the current image with an image of the background

- Top-left. Original
- Top-right: Current background representation
- Bottom-Left: distance between current image and background image
- Bottom-Right: foreground pixel



Background subtraction

- Main question : how to (automatically) build the background ?
  needs to adapt to changes
  - illumination (gradual, sudden, eg indoor light)
  - motion jittering (camera oscillation; or scene content –water, rain…)
  - scene structure (parked car; moved content like chair in an office…)

- Related question
  how to compare the current image to the background model ?

- Rest of the talk outline
  - parametric methods (GMM)
  - some improvement
  - recent method evaluation
  - non-parametric state-of-the-art method

GMM foreground blob extraction
(background subtraction, Stauffer and Grimson)

key ideas:

1. model the color of **each pixel** over time with a separate GMM
2. determine what mixture components are background and foreground
3. associate new observations with a mixture => pixel classification

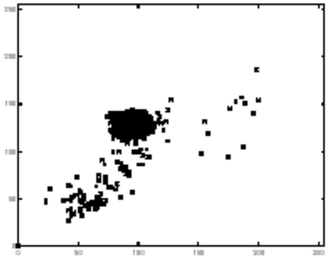
$$p(\mathbf{x}) = \sum p(\mathbf{z})p(\mathbf{x}|\mathbf{z}) = \sum^{K} \pi_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)$$

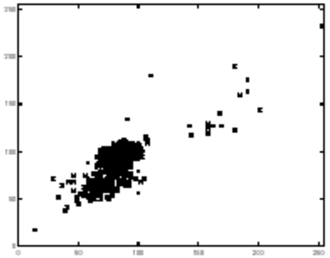
$$p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{z})p(\mathbf{x}|\mathbf{z}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)$$

$$p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{z})p(\mathbf{x}|\mathbf{z}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)$$

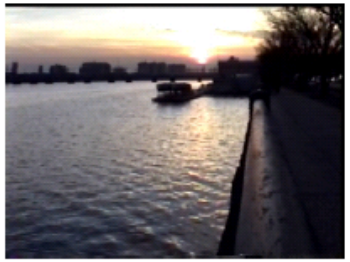
why is this necessary? background is non-Gaussian

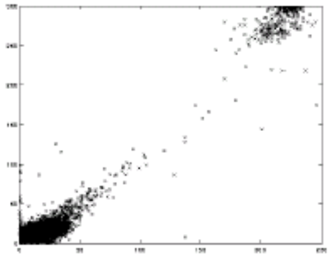
G and B values for one pixel over time



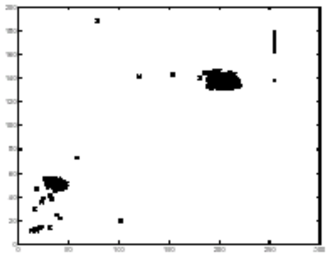


two plots taken 2 minutes apart





bimodal distribution due to specularities in water



bimodal distribution due to monitor flickering

learning the GMM model for each pixel

accumulated evidence for each pixel (color pixels)

$$\{X_1, ..., X_t\} = \{I(x_0, y_0, i) : 1 \leq i \leq t\}$$

$$P(X_t) = \sum_{i=1}^{K} \omega_{i,t} * \eta(X_t, \mu_{i,t}, \Sigma_{i,t})$$

assumptions:

+ small number of components *K* (3-5)

+ diagonal covariance matrix

GMM on-line learning

+ EM learning at each time step is not computationally feasible

+ use a new observation to choose and adapt only one Gaussian component

1. search for the closest existing component (within 2.5 std)

2. if no component is good, start new component (mean= pixel, large variance, low weight), and eliminate smallest-weight component

3. if a component is good, adapt its parameters (all other parameters fixed)

Update weights of all component

$$\omega_{k,t} = (1-\alpha)\omega_{k,t-1} + \alpha(M_{k,t})$$

learning rate

1 for chosen component
0 otherwise

Update Gaussian parameters of selected component

$$\mu_t = (1-\rho)\mu_{t-1} + \rho X_t$$

$$\sigma_t^2 = (1-\rho)\sigma_{t-1}^2 + \rho(X_t - \mu_t)^T(X_t - \mu_t)$$

where

$$\rho = \alpha\eta(X_t|\mu_k, \sigma_k)$$

background model estimation

key idea: what GMM components represent the background?

background: components with most supporting evidence and least variance

=> rank the components using weight/variance ratio

⇒ choose the B ranked components whose accumulated weight are above a threshold as the background's

$$B = argmin_b \left(\sum_{k=1}^{b} \omega_k > T \right)$$

T: minimum portion of data accounted for the background

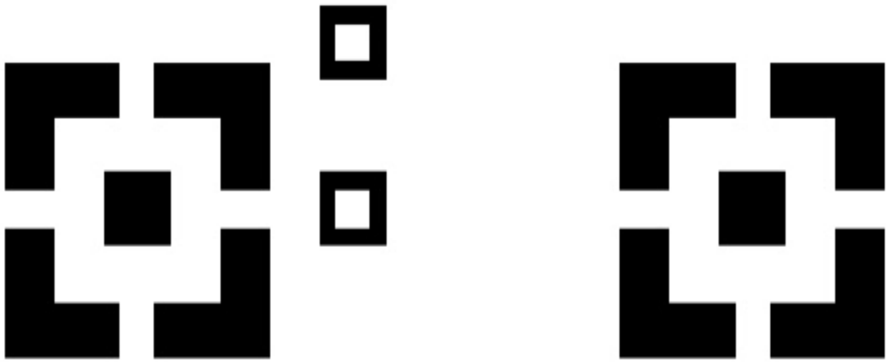
Summary

- background: B first GMM components
- foreground: the other ranked GMM components

Classification: pixel = background if associated with one of the B component

= foreground otherwise

Sample result



Issue (1) cast shadow from moving object

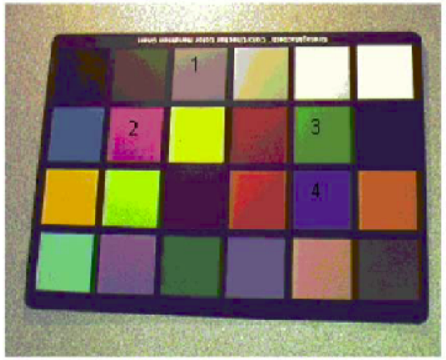
[Sanin/Sanderson/Lowell, Shadow detection survey, 2012]

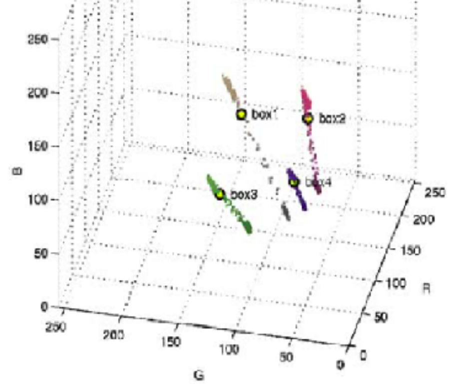
- Principles/methods to deal with shadow
  - intensity/physical/chromaticity:

    scene pixels keep their color,

    become darkers

Distribution of four pixels measurements for various illumination conditions





  - geometry

    account for objects/illumination configuration

    useful only in specific scene cases

  - texture

    invariant to shadow – most effective, but usually slower

  - temporal features

    shadow moves in the same way as objects => used as (tracking) post-filter

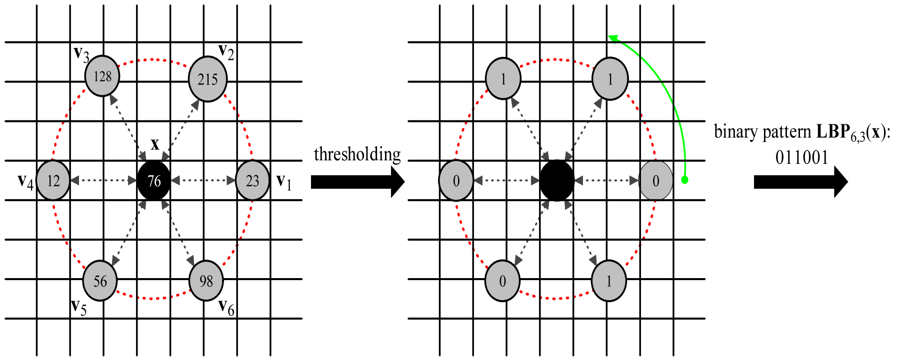
Example: Robust Multi-Layer Background Subtraction

[Yao & Odobez CVPR-Visual Surveillance workshop, 2007] code available: www.idiap.ch/~odobezTexture & Color

- Texture+color features - Distance map D => distance to nearest mode

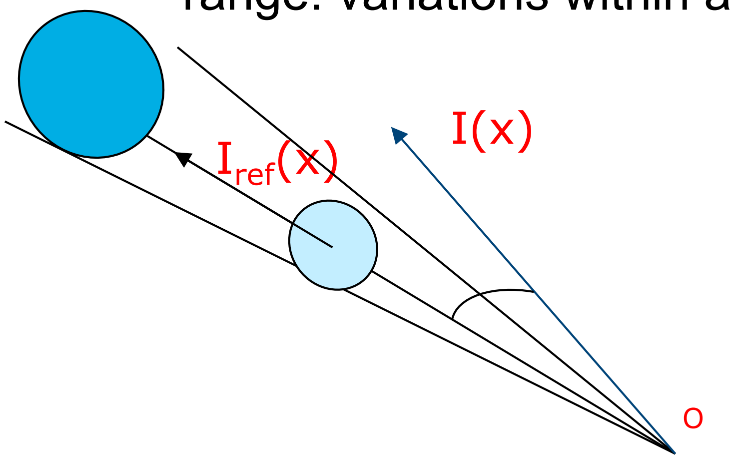
$$D = \lambda\, Dist_{texture} + (1- \lambda)Dist_{color}$$

- Texture - local binary pattern (LBP)
  - differential feature
  - robust to shadow/illumination changes
  - but: not very informative in uniform regions

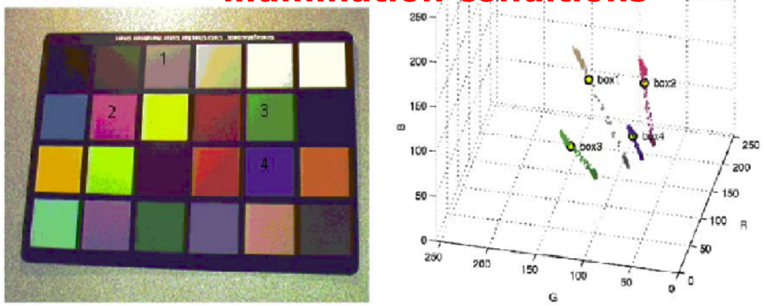


- Color - RGB

  Perceptual shadow invariant distance measure based on
  - angle between I(x) and $I_{ref}(x)$ w.r.t. the RGB origin O
  - range: variations within an interval can be due to cast shadow



Distribution of four pixels measurements for various illumination conditions



Issue (2) global decision ?

- Most methods : pixel based background model

  but: correlation between models and decision at nearby pixels

- Post-processing of decision maps
  - morphological operators, MRF at pixel level
  - image aware filtering : avoid smoothing across gradients
    - bilateral filter (cf next slide)
    - MRF on super-pixels (regions)

- Eigenbackground:

  learn full correlation of background pixel colors by applying Principal Component Analysis (PCA) on training images
  - pros: useful to learn correlated intensity variations (indoor switching on/off lights), predict pixel color
  - cons: can be slow, difficulty for local adaptation, robust

Foreground Regional Detection

[Yao & Odobez, 2007]

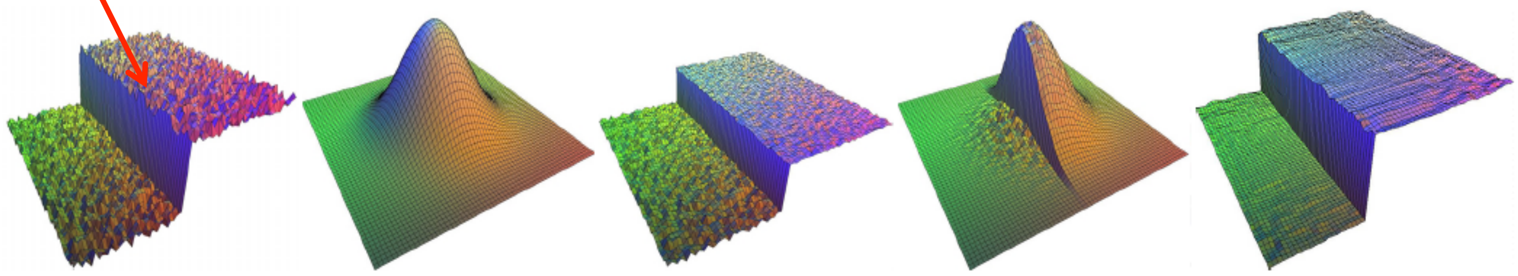
- Distance computation to nearest mode

  $$D = \lambda\, Dist_{texture} + (1-\lambda)Dist_{color}$$

- Cross bilateral filter
  - Combines spatial and intensity smoothing

    $$D_{bf}(x) = 1/C(x) \sum_v G_{\sigma\_s}(||v-x||)\, G_{\sigma\_r}(|\, I(v)-I(x)\, |)\, D(v)$$

x



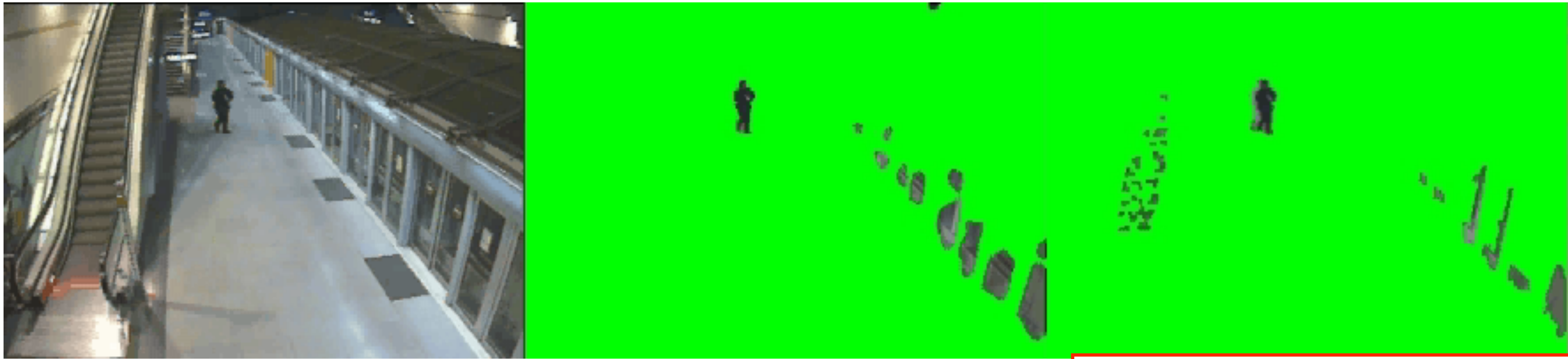
input — spatial kernel f — influence g in the intensity domain for the central pixel — weight $f \times g$ for the central pixel — output

  - Implicit edge-preserving smoothing
- Foreground detection: Thresholding on D_{bf}

Robust Multi-Layer Background Subtraction

[Yao & Odobez CVPR-Visual Surveillance workshop, 2007] code available: www.idiap.ch/~odobez/

Method robust in different environmental conditions



Original video

New method

- correct detection of doors
- less shadow problems
- better object segmentation

Classical approach (Stauffer & Grimson, OpenCV version)

- shadow issues
- more miss detection
- more false alarms (escalators)

better handling of moving background object, cast shadows, local camouflage

Evaluation : the change detection challenge

- Evaluation: difficult task
  - ground truth painful to obtain
  - multiple conditions, applications etc
- Few evaluation datasets
  - difficult to compare methods
  - difficult to know which methods work in which conditions/scenario
- http://www.changedetection.net/
  - Workshop at the Computer Vision and Pattern Recognition conference, 2012
  - Provide datasets/platforms, links to resources (code) and papers

=> Summary of some of their findings (see their website for more)

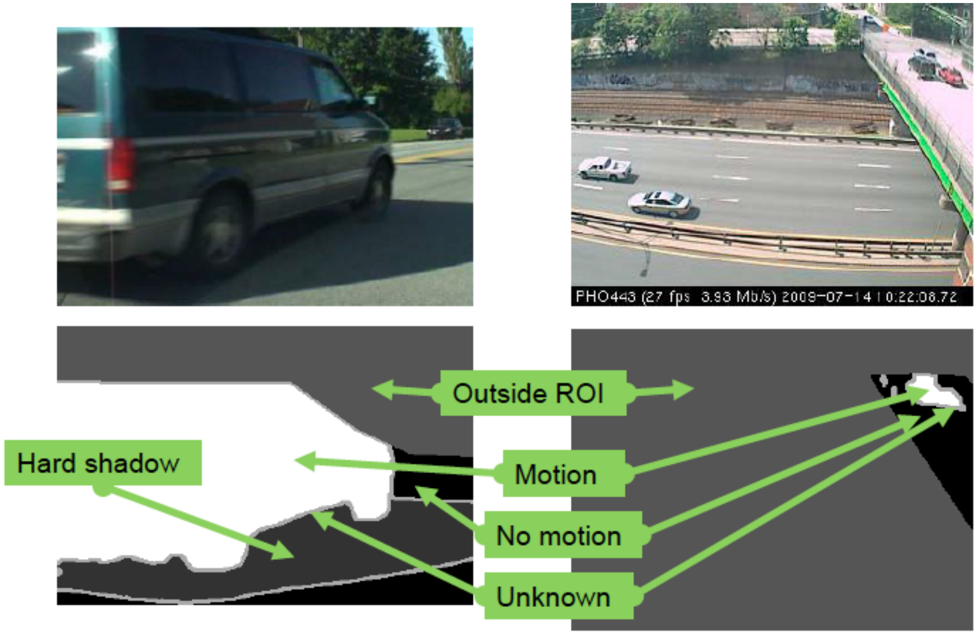
Datasets


Dynamic background
(6 outdoor videos)
Camera Jitter
(4 videos – 3 outdoor, 1 indoor)
Intermittent Object Motion
(6 videos, 5 outdoor, 1 indoor)

Datasets


Shadow
(6 videos, 3 indoor, 3 outdoor)
Baseline
(4 videos, 2 indoor, 2 outdoor)
Infrared
(5 videos, 3 indoor, 2 outdoor)

Evaluation & Results



© changedetection challenge

- Annotation
  - Ground truth

    object, animal, man-made objects
  - Not of interest

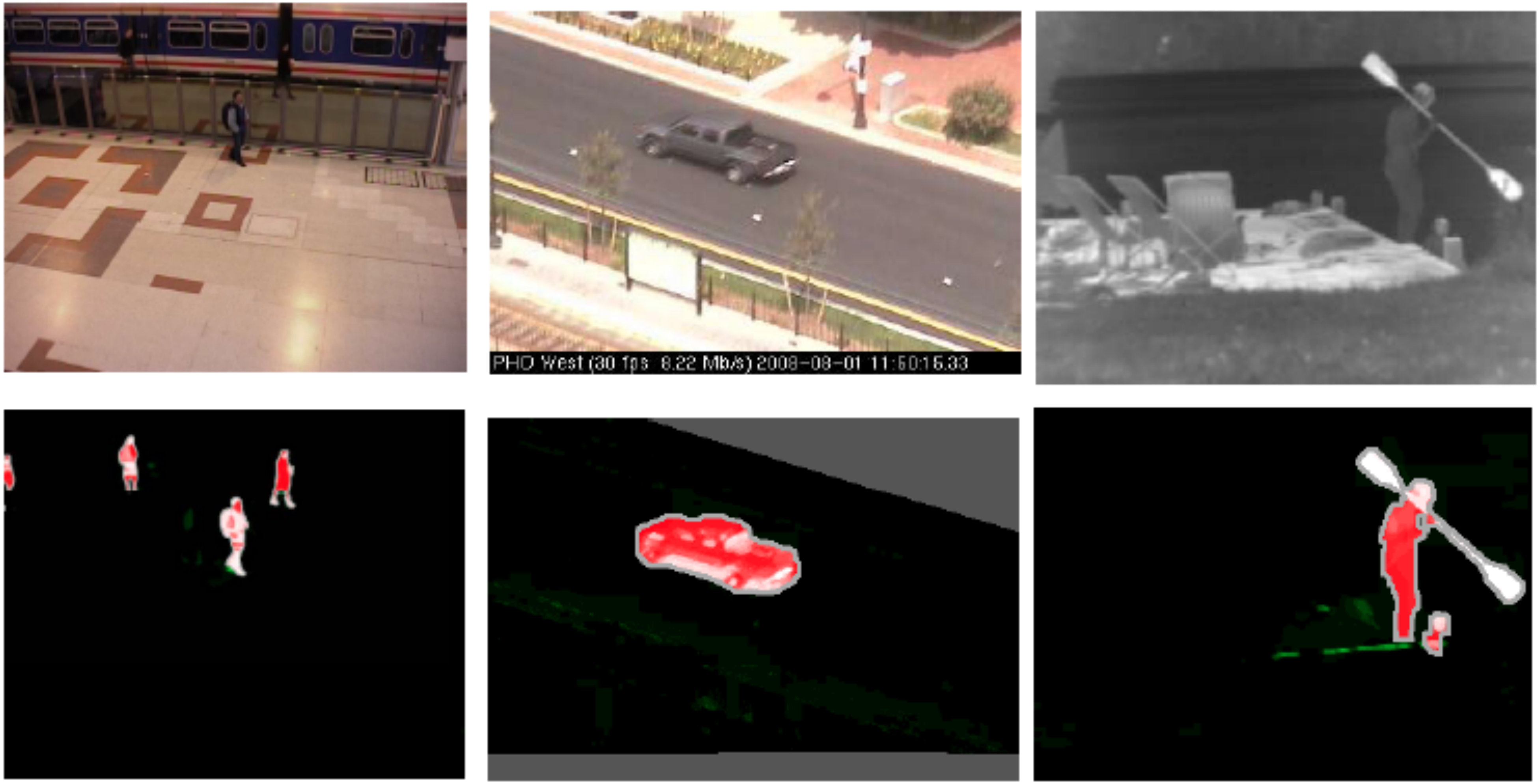
    moving water, rain, shadow, etc

- Different metrics
  - Precision, recall, percentage of wrong classification…
  - Average (video, categories) + ranking

- Main results/conlusions
  - Foreground detection in baseline is more or less solved
  - Background motion (small repetitive motion) not a problem for most methods

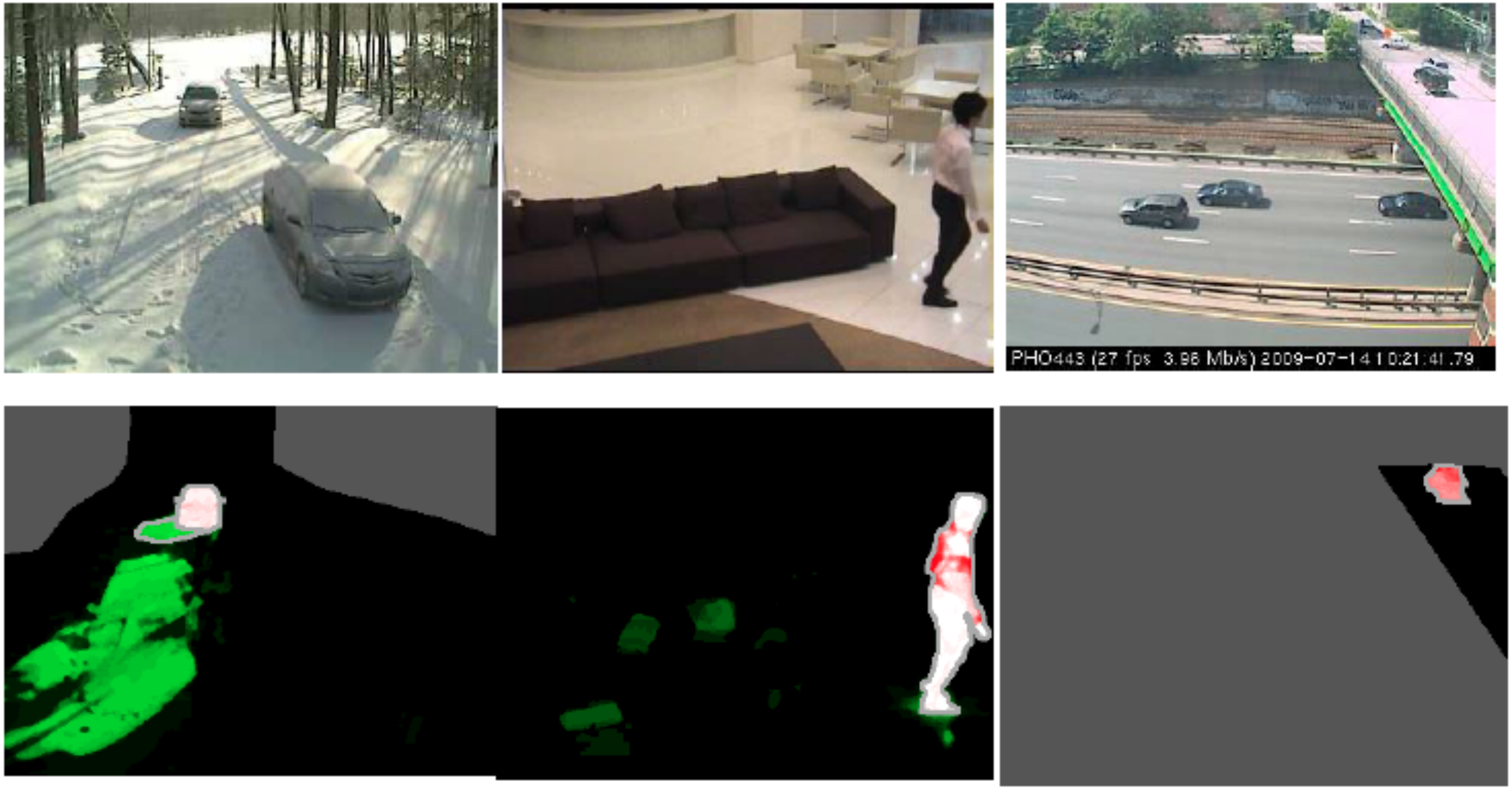
- Failure modes

Results (2) – failure modes (camouflage)



Results (2) – failure modes (hard shadows)

Results (2) – failure modes (intermittent motion)



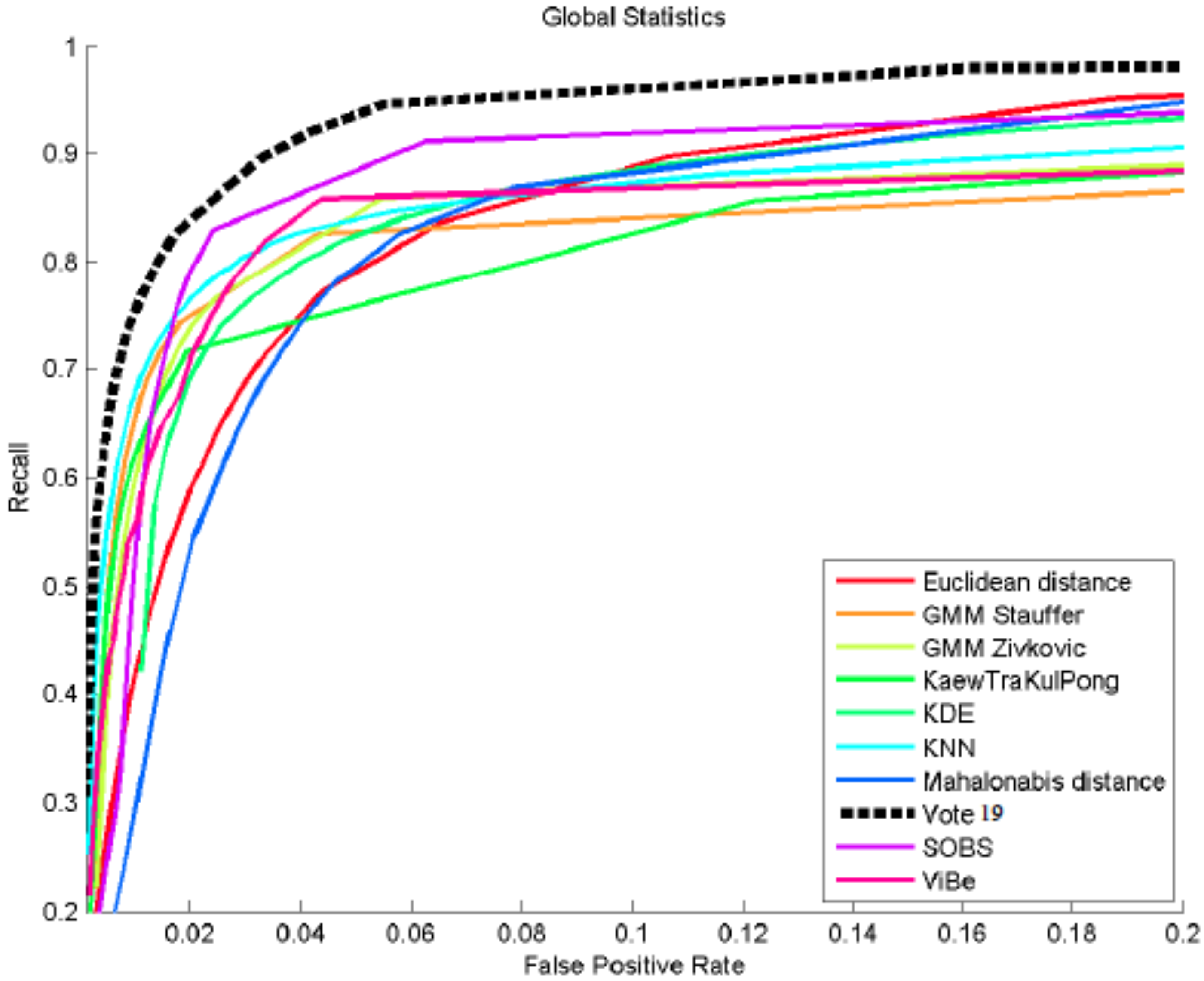
© changedetection challenge

- This issue reaches semantic/high level limit

  When does an object becomes part of a background ?

  example: a stopped car ? may stay for 1 minutes, 1 hour, 1 day…

Combination (majority votes) is best …



Winning methods ?

Recent methods

KDE methods

GMM methods

Simplistic methods

Method	Average ranking accross categories
PBAS [16]	3.00
ViBe+ [11]	4.83
PSP-MRF [10]	5.50
SC-SOBS [18]	6.67
Chebyshev prob. with Static Object detection [17]	7.00
KNN [21]	8.50
SOBS [1]	8.83
KDE - Integrated Spatio-temporal Features [9]	9.33
KDE - ElGammal [4]	10.00
ViBe [3]	10.17
GMM | KaewTraKulPong [2]	10.50
KDE - Spatio-temporal change detection [12]	11.17
Bayesian Background [19]	11.83
GMM | Stauffer & Grimson [20]	12.33
GMM | Zivkovic [5]	14.50
GMM | RECTGAUSS-Tex [13]	14.83
Local-Self similarity [8]	15.17
Mahalanobis distance [6]	16.17
Euclidean distance [7]	17.67

KDE : Kernel density estimators

(eg. Elgammal, Harwood, Davis, 2000)

$$p(x) \propto \sum_{i=1}^{N} w_i \mathcal{N}(x; x_i, \Sigma)$$

- Move towards non parametric methods
  - Background Probability Density p(x) built from N most recent points (from the same pixels or including neighboors)
  - If p(x) > Threshold => the point belongs to the background;
  - only points labelled as background are used for update

- Issues
  - N usually needs to be quite large => large memory requirement
  - Computing the density is expensive
    - Look Up Tables (LUT) can be used somehow
    - Recursive approach (Sequential Kernel Density Approx.)

      => back to GMM with more Gaussians

New methods ?
(VIBE+, Droogenbroeck et al, 2010/2012; SaCon, Wang & Sutter, 2007,…)

- Very simple methods, non parametricity pushed to the extreme
- Example VIBE
  - Background of a pixel represented by N samples (x_1, x_2,…x_{20}) (eg 20)
  - New pixel observation x
    - If Card $\{x_i / d(x, x_i) < R\} >= N_{min}$ (eg 2) : we have a background pixel
  - BG model update: update RANDOMLY only from x classified as background
    - At random time (on average one time every T (eg 16) frames)
    - Model point x_i to be replaced (by new one) is selected randomly

      => there is no temporal order in the current samples
  - BG model update from neighboors: at RANDOM time and RANDOM position

    (i) account for texture motion (foliage…) (ii) avoids the BG model to never update
- Very Fast
- PBAS [Hofmann, Tiefenbacher, Rigoll, 2012] (change detection Winning method)
  - Adapt R and T (update rate) per pixel

Conclusion

- **visual representation** of people
  - large, classic field, we covered a few representations (contours, patches/histograms, blob…)
  - other methods available
  - tradeoff: discrimination vs. flexibility
  - some models are better to characterize individual people
  - others to characterize people classes
  - nowdays: trend is to build detectors for whole body/parts

    Cf talks today and tomorrow
- background subtraction
  - Real-time powerful algorithms exist (with free software)
  - Stauffer & Grimson is not the state-of-the-art
  - Main issues: shadows, and intermittent motion

references

- A. Blake and M. Isard, *Active Contours,* Springer, 1998.
- S. Gong, S. McKenna and A. Psarrou, *Dynamic Vision. From Images to Face Recognition*, Imperial College Press, 2000.
- J. MacCormick, *Probabilistic Modeling and Stochastic Algorithms for Visual Localization and Tracking*, PhD Thesis, University of Oxford, 2000.
- C. Stauffer and E. Grimson, *Adaptive Background Mixture Models for Real-time Tracking*, CVPR 1999.
- J. Yao and J.M. Odobez, *Multi-Layer Background Subtraction based on Color and Texture*, CVPR workshop on Visual Surveillance, 2007.
- O. Barnich and M. van Droogenbroeck *ViBe: A Universal Background Subtraction Algorithm for Video Sequences*, IEEE Trans. Image Processing, June 2011
- N. Goyette, P.-M. Jodoin, F. Porikli, J. Konrad, and P. Ishwar, *Changedetection.net: A new change detection benchmark dataset*, in Proc. IEEE Workshop on Change Detection (CDW’12) at CVPR’12, Providence, RI, 16-21 Jun., 2012
- M. Hofmann, P.Tiefenbacher, G. Rigoll "*Background Segmentation with Feedback: The Pixel-Based Adaptive Segmenter*", in proc of IEEE Workshop on Change Detection, 2012