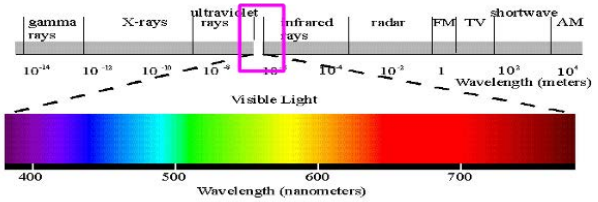
Introduction to Computer Vision
Handcrafted methods

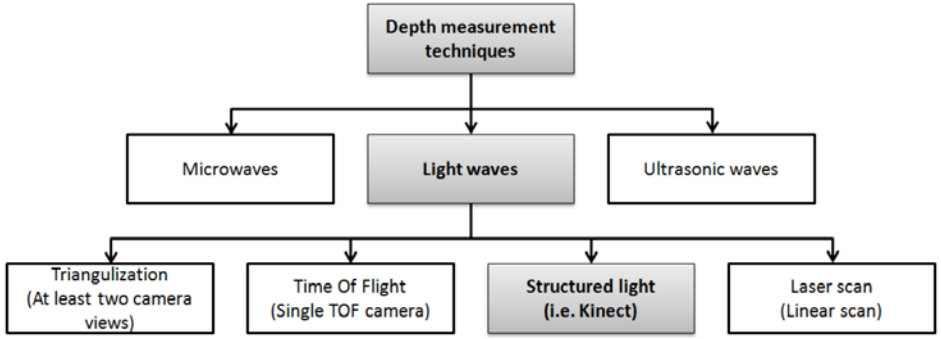
Dr. Sergio Escalera
University of Barcelona

UNIVERSITAT DE BARCELONA

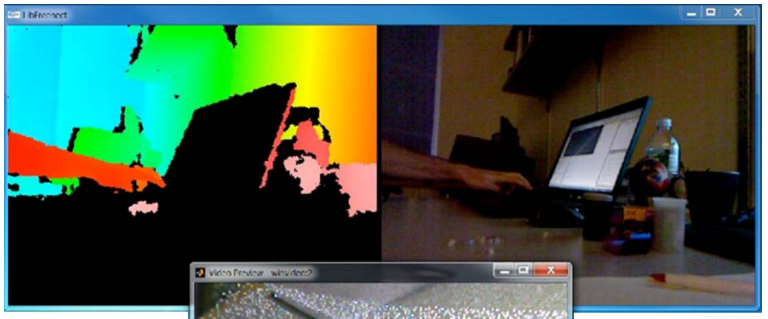
Any kind of visual data (even other kind of data)

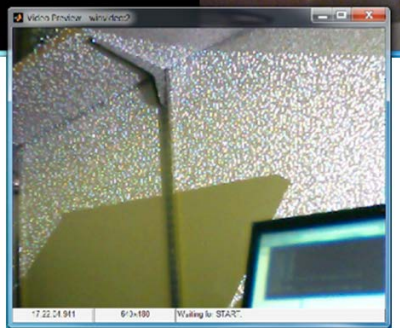
Images, images and … images !!!

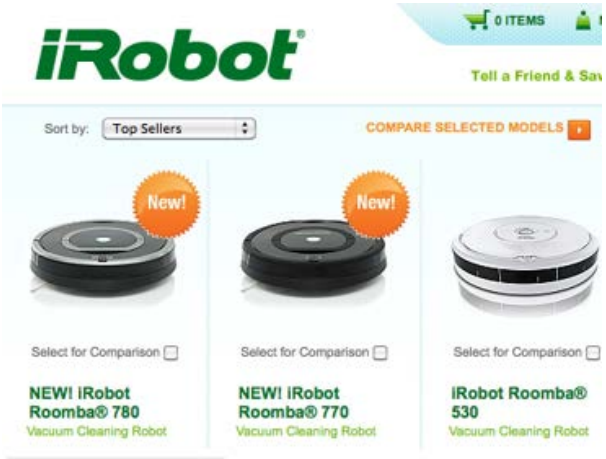




Weareable sensors







Thermal

detect *the same* interest points regardless of *image changes*

- Feature points are used also for:
  - Object recognition
  - Image alignment (homography)
  - 3D reconstruction
  - Motion tracking
  - Indexing and database retrieval
  - Robot navigation
  - … other

Features from different modalities

A) Identify points of interest

-Corners

-Stable visual points

-Saliency points

-Pixels or regions at different sizes from detectors (outputs from classifiers/detectors/segmentation strategies learnt from B)

B) Describe points of interest to match features

-Template-based:

Intensity-based

-Statistical descriptors:

SIFT, HOG, etc.

-Structural descriptors

Contours, silluethes, fragments, etc.

Features from different modalities

A) Identify points of interest

International Journal of Computer Vision 37(2), 151–172, 2000

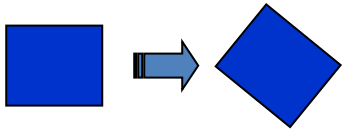
Evaluation of Interest Point Detectors

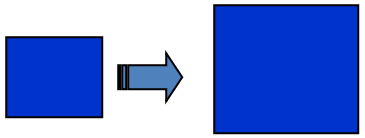
CORDELIA SCHMID, ROGER MOHR AND CHRISTIAN BAUCKHAGE
INRIA Rhône-Alpes, 655 av. de l'Europe, 38330 Montbonnot, France
Cordelia.Schmid@inrialpes.fr

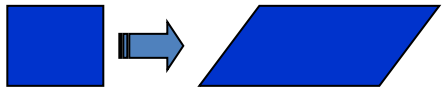
Abstract. Many different low-level feature detectors exist and it is widely agreed that the evaluation of detectors is important. In this paper we introduce two evaluation criteria for interest points: repeatability rate and information content. Repeatability rate evaluates the geometric stability under different transformations. Information content measures the distinctiveness of features. Different interest point detectors are compared using these two criteria. We determine which detector gives the best results and show that it satisfies the criteria well.

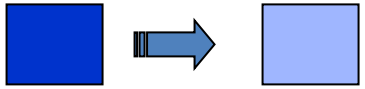
A) Identify points of interest

- Geometry
  - Rotation
  - Similarity (rotation + uniform scale)
  - Affine (scale dependent on direction)
    valid for: orthographic camera, locally planar object
- Photometry
  - Affine intensity change ($I \rightarrow a\ I + b$)



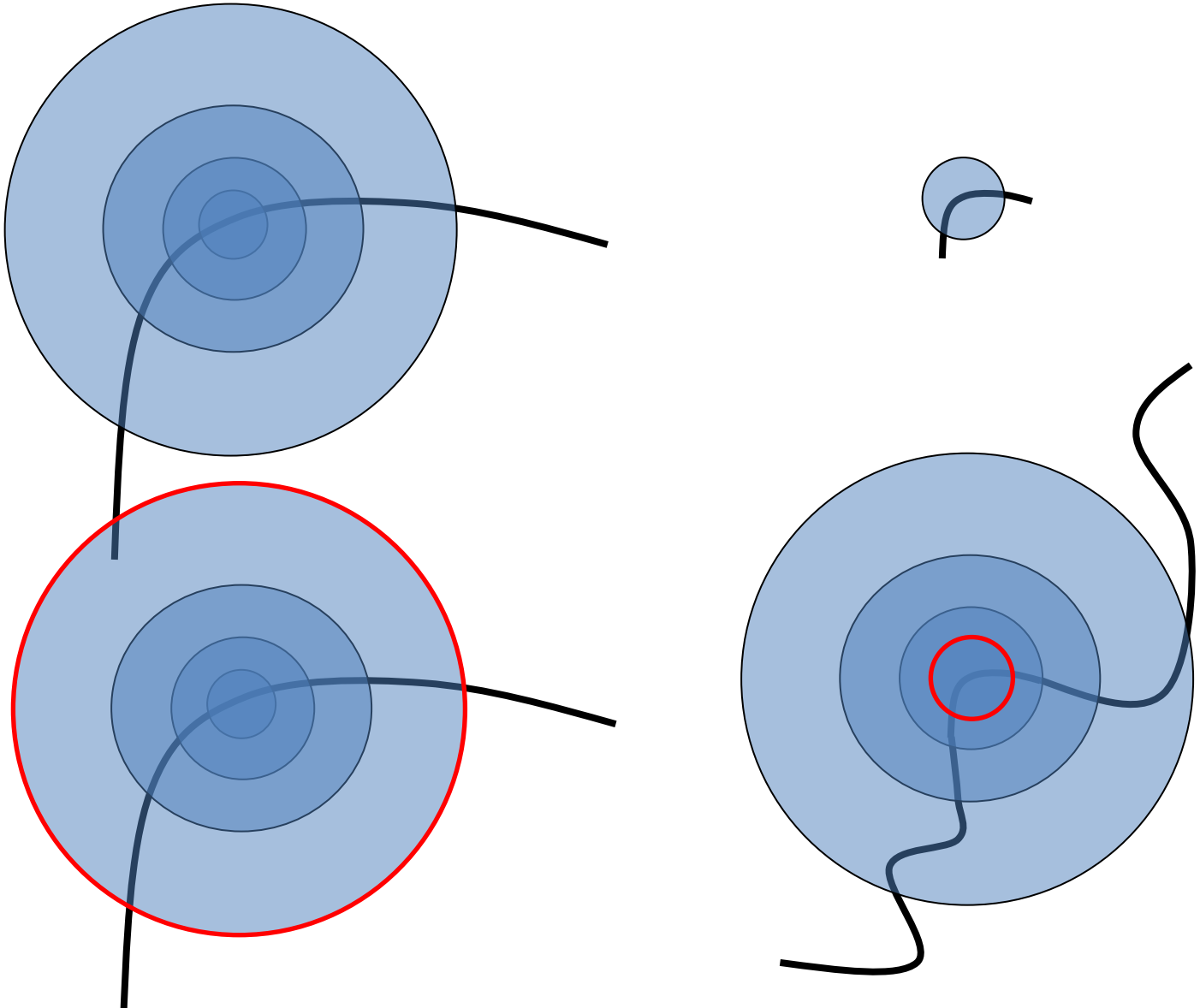






Features from different modalities

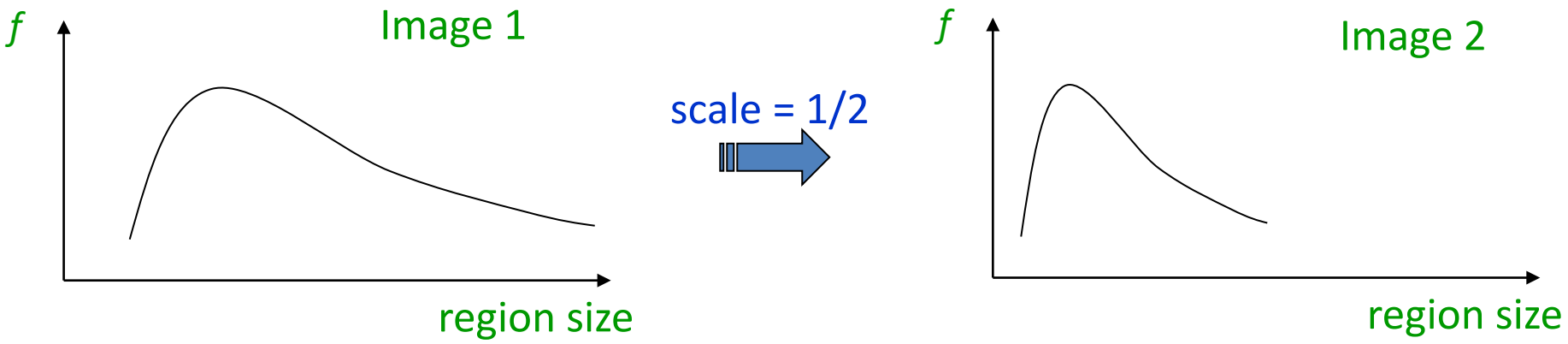
A) Identify points of interest



A) Identify points of interest

- Solution:
  - Design a **function** on the region (circle), which is "**scale invariant**" (the same for corresponding regions, even if they are at different scales)

    Example: average intensity. For corresponding regions (even of different sizes) it will be the same.
  - For a point in one image, we can consider it as a function of region size (circle radius)



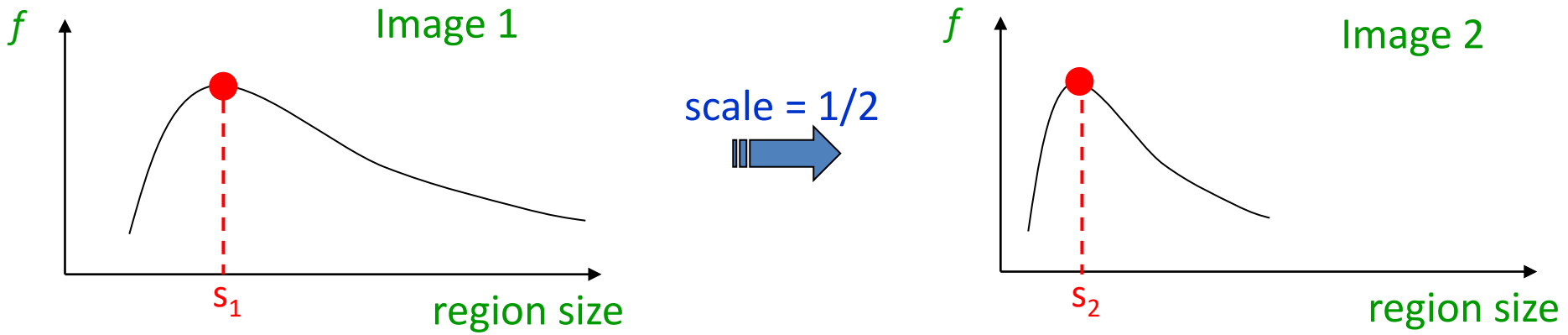
A) Identify points of interest

- Common approach:

Take a local maximum of this function

Observation: region size, for which the maximum is achieved, should be *invariant* to image scale.

Important: this scale invariant region size is found in each image independently!



A) Identify points of interest

- Functions for determining scale $f = \text{Kernel} * \text{Image}$

Kernels:

$$L = \sigma^2 \left(G_{xx}(x, y, \sigma) + G_{yy}(x, y, \sigma) \right)$$

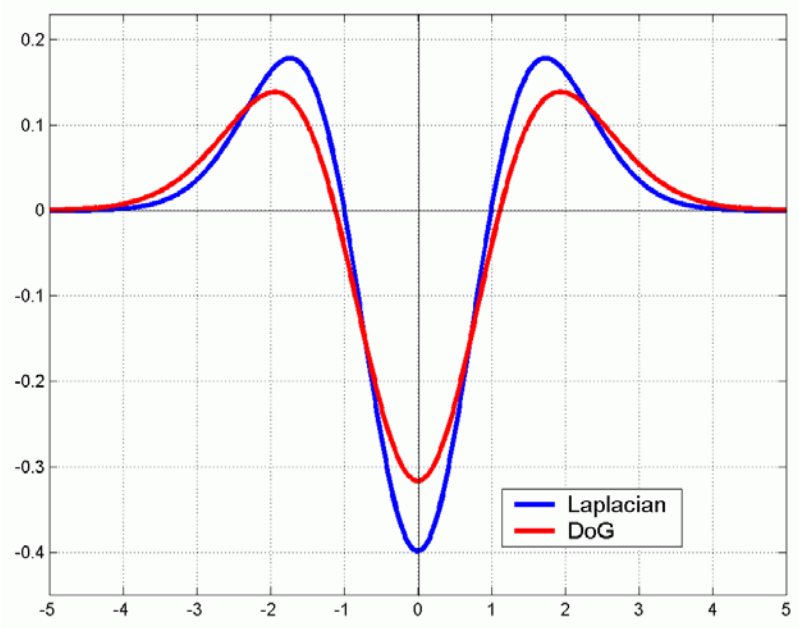
(Laplacian)

$$DoG = G(x, y, k\sigma) - G(x, y, \sigma)$$

(Difference of Gaussians)

where Gaussian

$$G(x, y, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2 + y^2}{2\sigma^2}}$$



Note: both kernels are invariant to *scale* and *rotation*

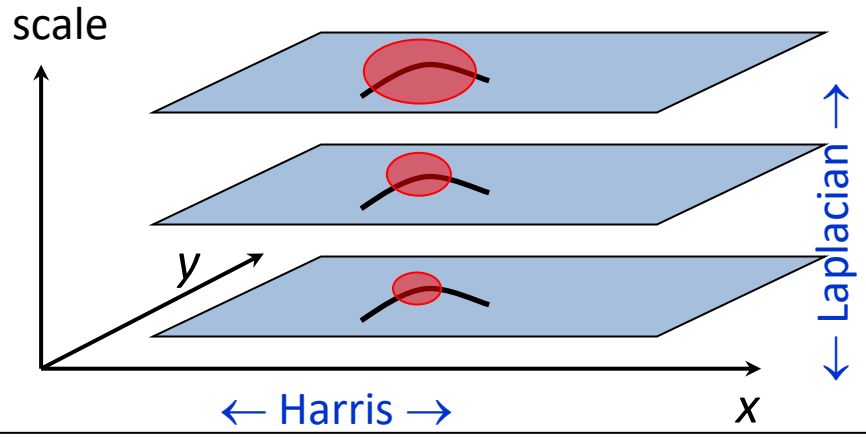
A) Identify points of interest
Scale Invariant Detectors

- Harris-Laplacian[1]

  Find local maximum of:

  – Harris corner detector in space (image coordinates)

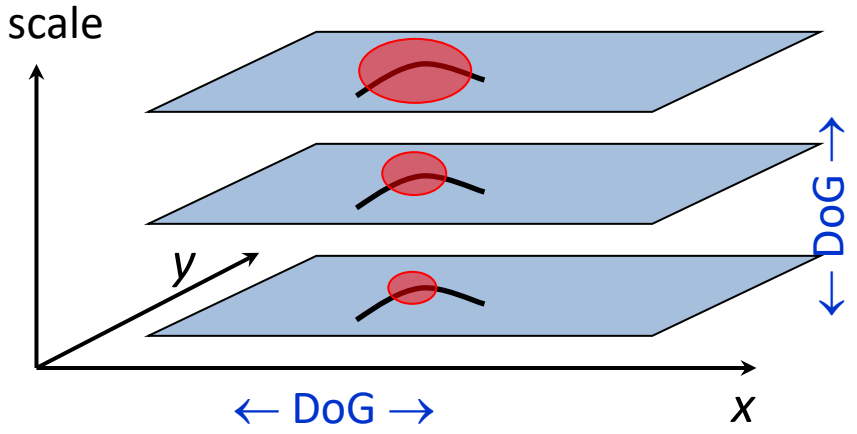
  – Laplacian in scale



- SIFT (Lowe)[2]

  Find local maximum of:

  – Difference of Gaussians in space and scale



[1] K.Mikolajczyk, C.Schmid. "Indexing Based on Scale Invariant Interest Points". ICCV 2001

[2] D.Lowe. "Distinctive Image Features from Scale-Invariant Keypoints". IJCV 2004

A) Identify points of interest

Affine patches

Detection of salient image regions

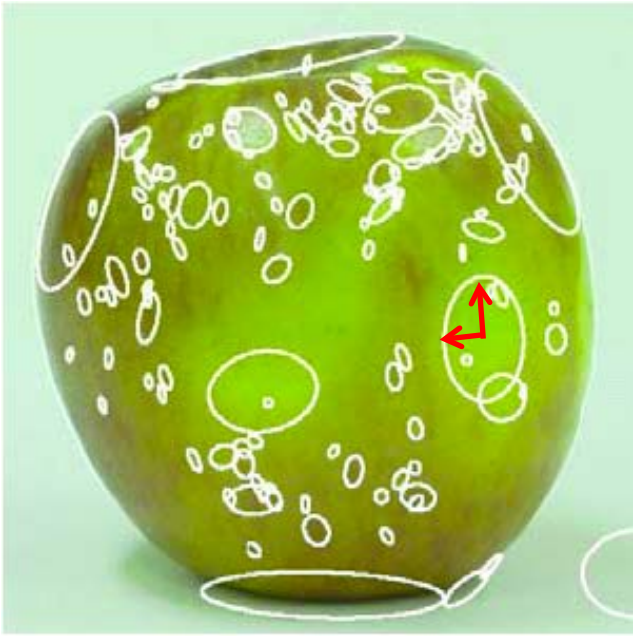
(Garding and Lindeberg, 96; Mikolajczyk and Schmid, 02)

a) an elliptical image region is deformed to maximize the isotropy of the corresponding brightness pattern.

b) its characteristic scale is determined as a local extreme of the normalized Laplacian in scale space.

c) the Harris (1988) operator is used to refine the position of the ellipse's center.

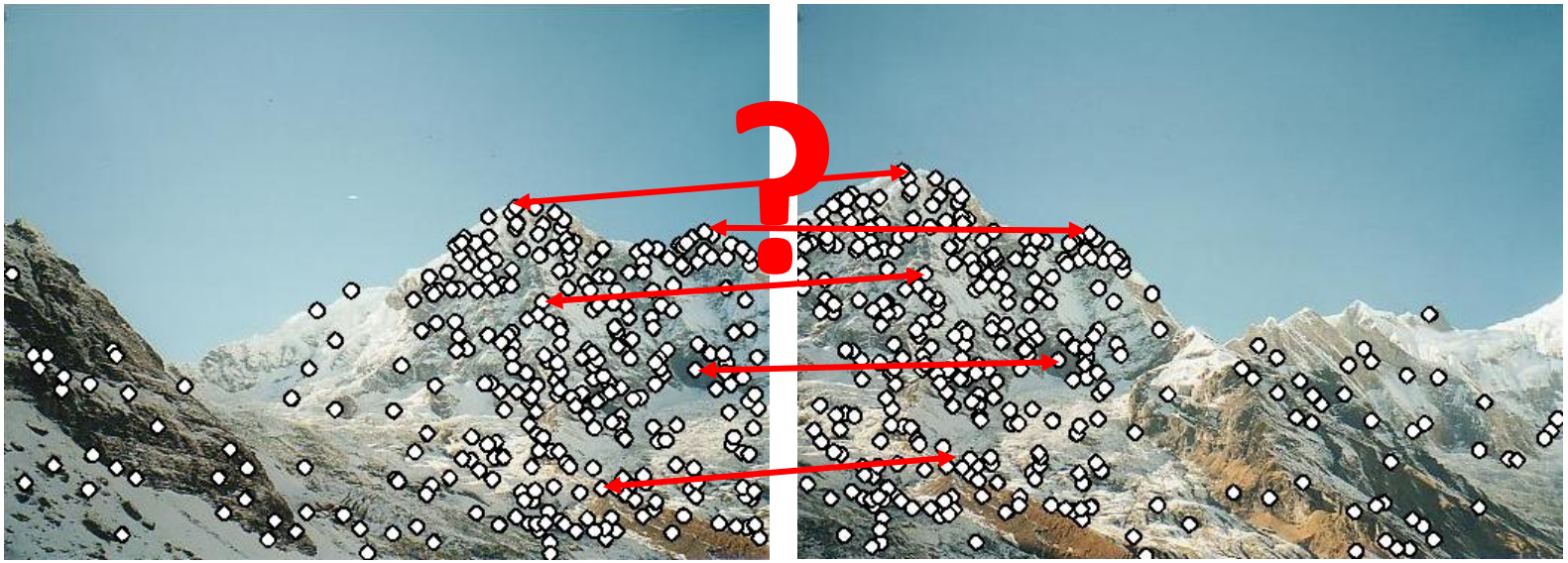
The elliptical region obtained at convergence can be shown to be covariant under affine transformations **(based on two main orthogonal eigenvectors)**.



B) Describe points of interest to match features

- We know how to detect points
- Next question:

How to match them?

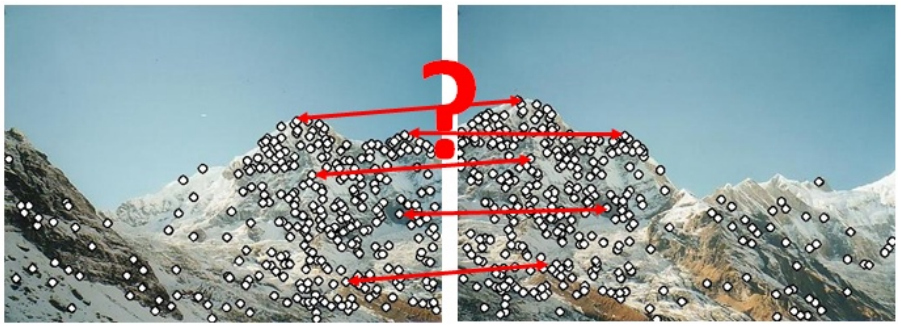


Point descriptor should be:

1. Invariant → **Main orientation normalization**
2. Distinctive

B) Describe points of interest to match features

- **How to match them?**
  RANSAC



WIKIPEDIA

•The input to the RANSAC algorithm is a set of observed data values, a parameterized model which can explain or be fitted to the observations, and some confidence parameters.

•RANSAC achieves its goal by iteratively selecting a random subset of the original data. These data are *hypothetical inliers* and this hypothesis is then tested as follows:

•A model is fitted to the hypothetical inliers, i.e. all free parameters of the model are reconstructed from the inliers.

•All other data are then tested against the fitted model and, if a point fits well to the estimated model, also considered as a hypothetical inlier.

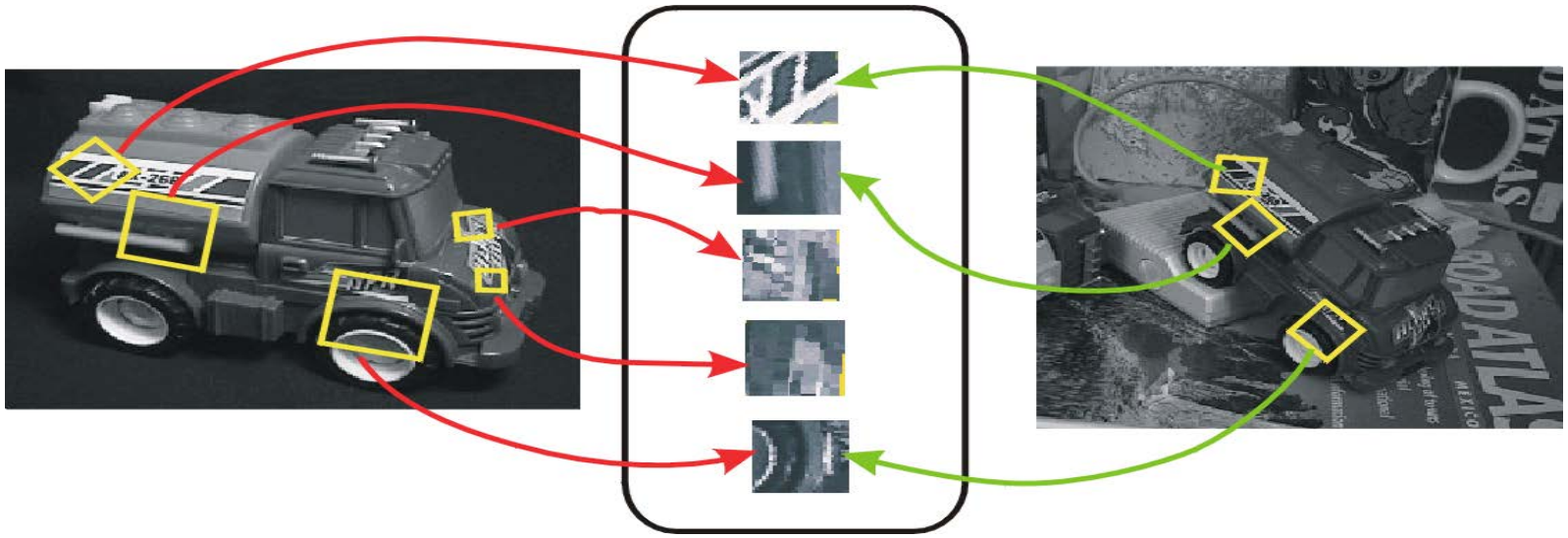
•The estimated model is reasonably good if sufficiently many points have been classified as hypothetical inliers.

•The model is reestimated from all hypothetical inliers, because it has only been estimated from the initial set of hypothetical inliers.

•Finally, the model is evaluated by estimating the error of the inliers relative to the model.

•This procedure is repeated a fixed number of times, each time producing either a model which is rejected because too few points are classified as inliers or a refined model together with a corresponding error measure. In the latter case, we keep the refined model if its error is lower than the last saved model.

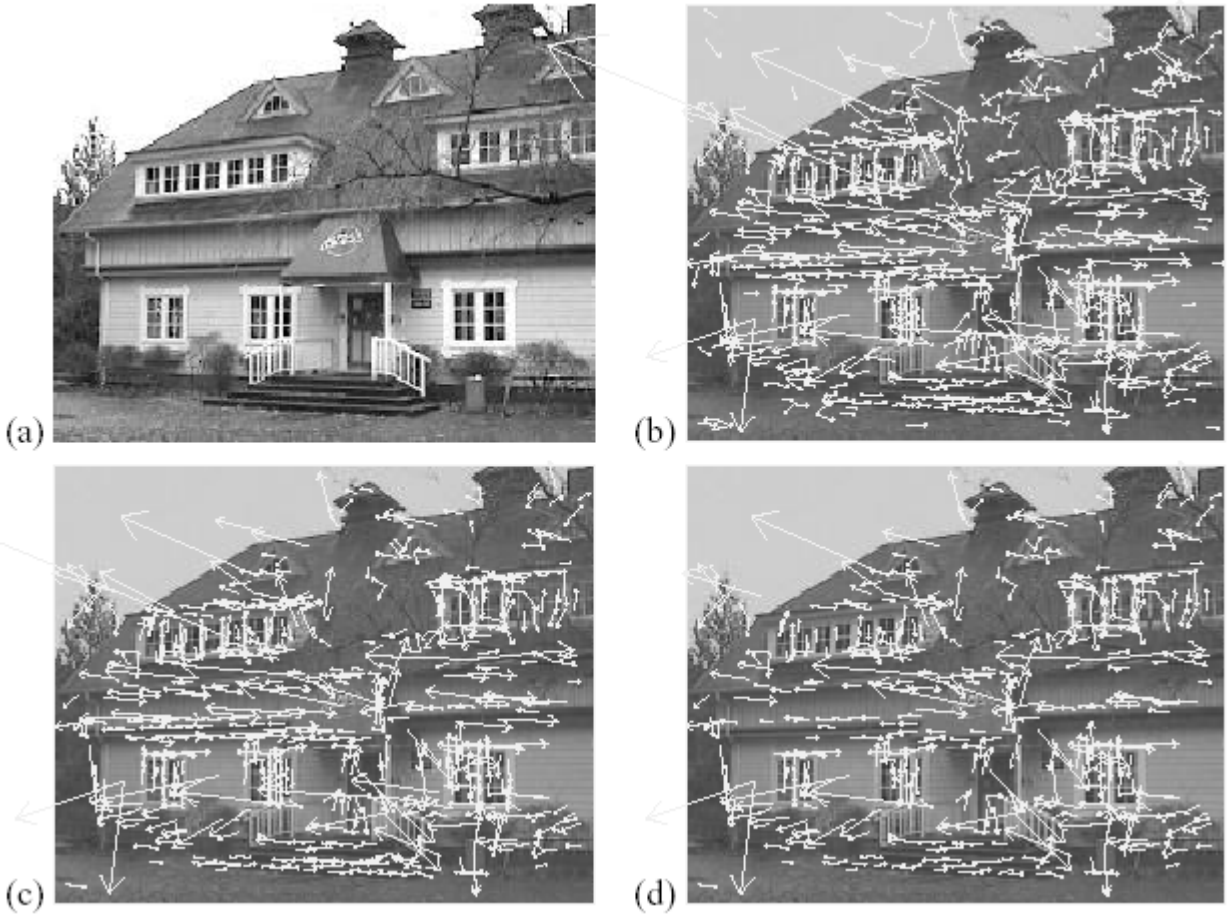
B) Describe points of interest to match features



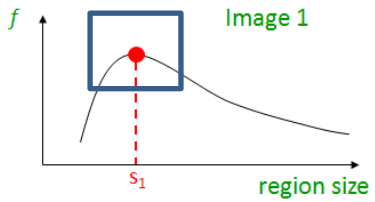
- **Locality:** features are local, so robust to occlusion and clutter (no prior segmentation)
- **Distinctiveness:** individual features can be matched to a large database of objects
- **Quantity:** many features can be generated for even small objects
- **Efficiency:** close to real-time performance
- **Extensibility:** can easily be extended to wide range of differing feature types, with each adding robustness

B) Describe points of interest to match features

SIFT - Threshold on value at DOG peak and on ratio of principle curvatures (Harris approach)



(a) 233x189 image
(b) 832 DOG extrema
(c) 729 left after peak value threshold



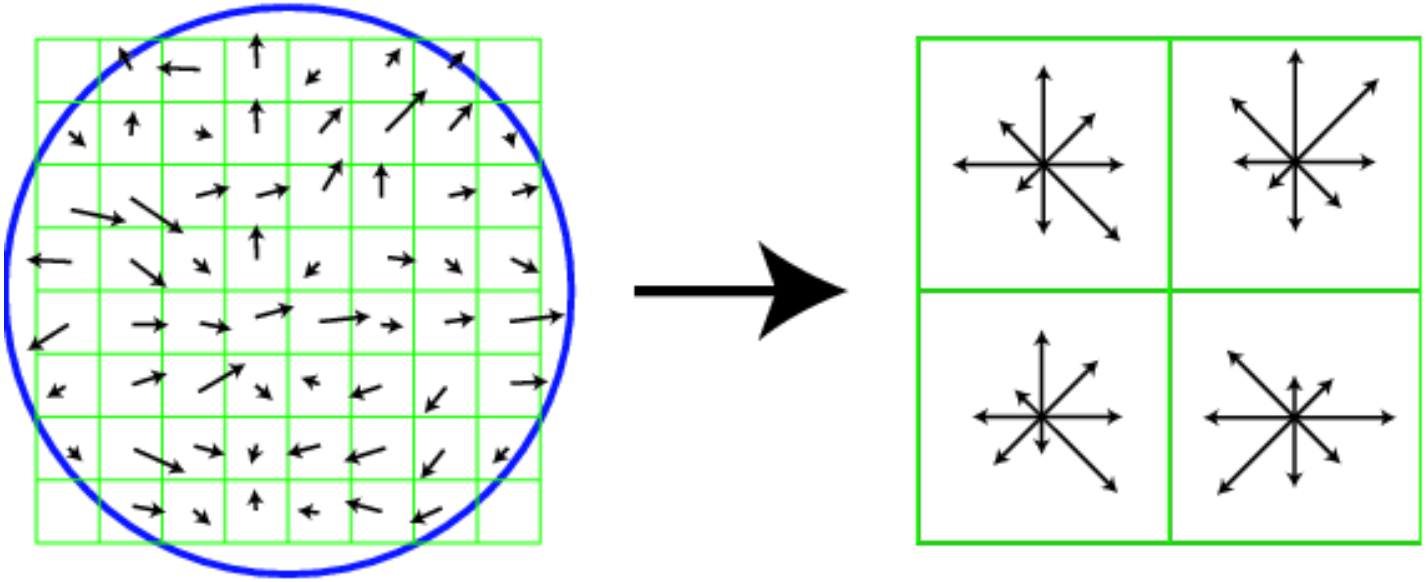
(d) 536 left after testing ratio of principle curvatures

(inside amount of gradient information)

B) Describe points of interest to match features

SIFT keypoint feature vector

- Thresholded image gradients are sampled over 16x16 array of locations in scale space
- Create array of orientation histograms
- 8 orientations x 4x4 histogram array = 128 dimensions



Features from different modalities

B) Describe points of interest to match features

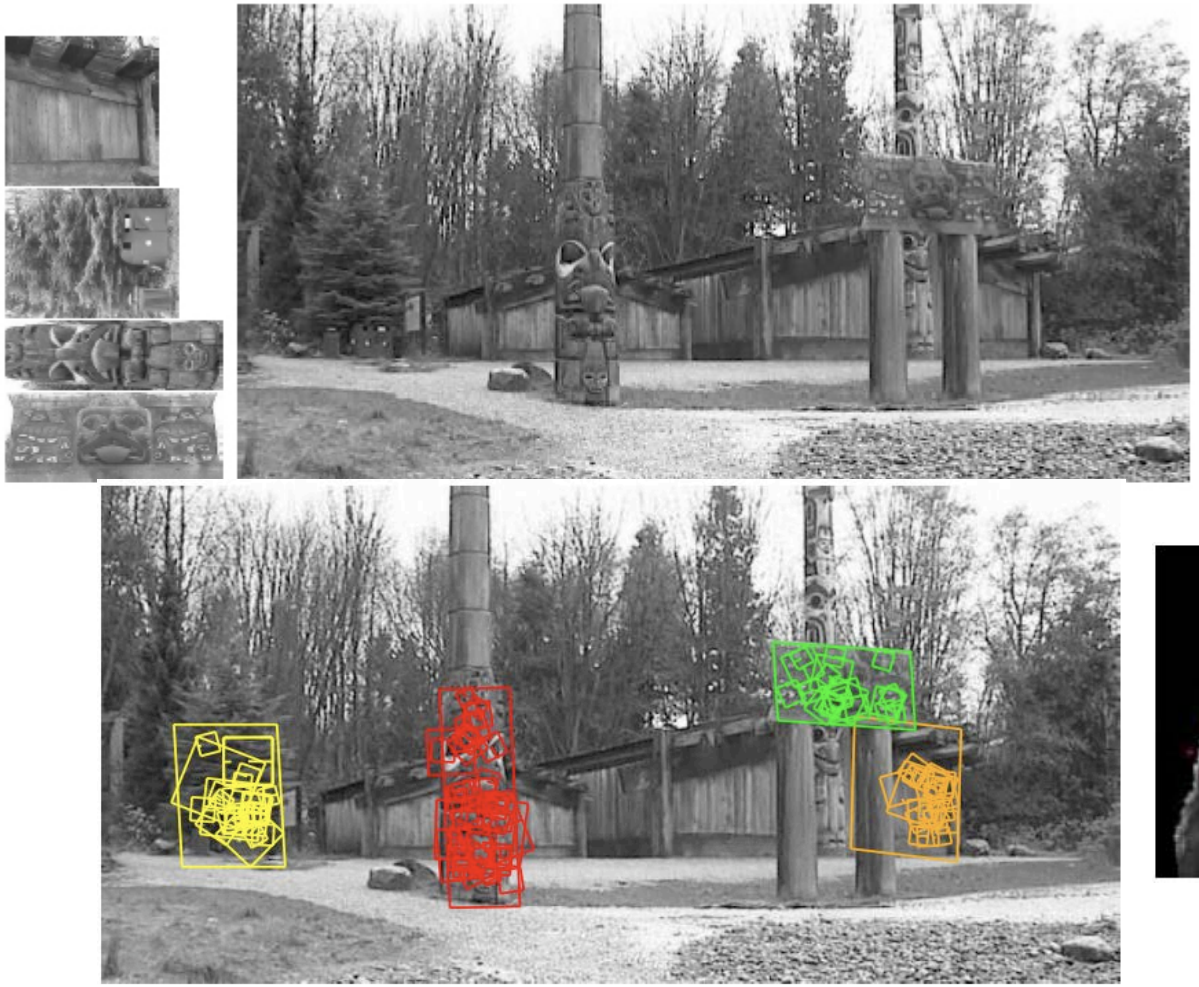
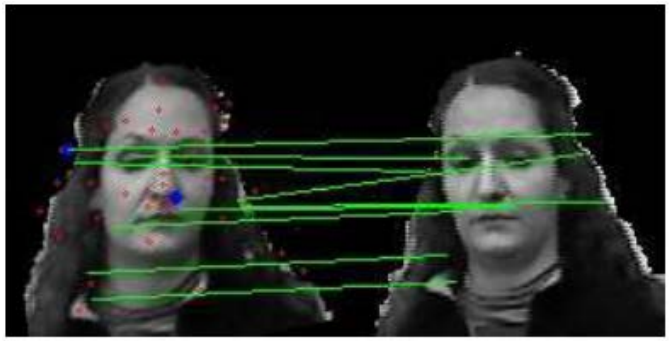


Figure 13: This example shows location recognition within a complex scene. The training images for locations are shown at the upper left and the 640x315 pixel test image taken from a different viewpoint is on the upper right. The recognized regions are shown on the lower image, with keypoints shown as squares and an outer parallelogram showing the boundaries of the training images under the affi ne transform used for recognition.

¿Useful for face analysis?



Features from different modalities

B) Describe points of interest to match features

Rapid Object Detection Using a Boosted Cascade of Simple Features

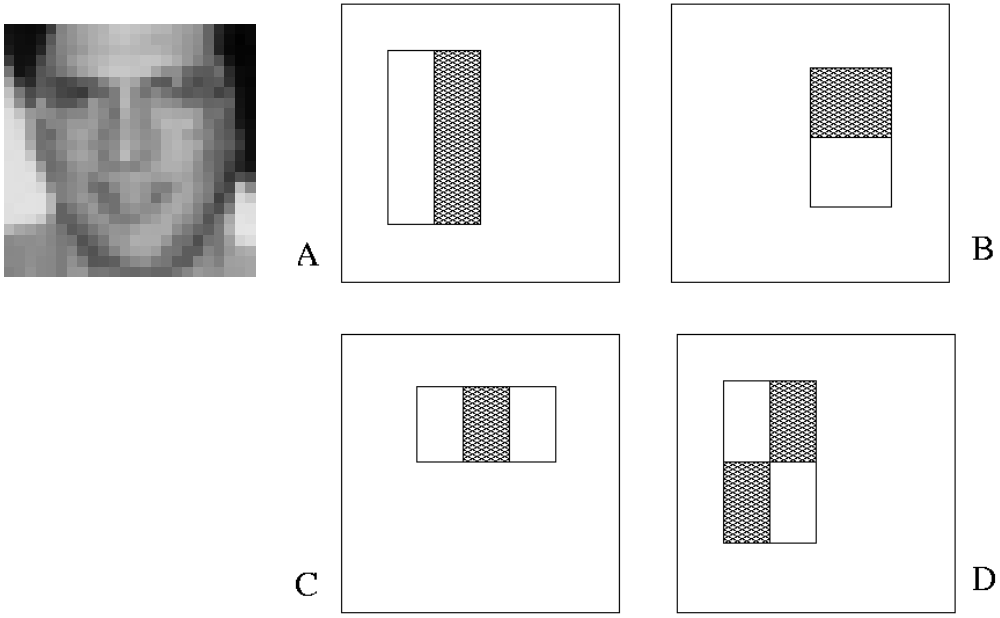
Paul Viola Michael J. Jones
Mitsubishi Electric Research Laboratories (MERL)
Cambridge, MA

"Rectangle filters"

Similar to Haar wavelets

Differences between sums of pixels in adjacent rectangles

$$h_t(x) = \begin{cases} +1 & \text{if } f_t(x) > \theta_t \\ -1 & \text{otherwise} \end{cases}$$



$$160{,}000 \times 100 = 16{,}000{,}000$$

Unique Features

B) Describe points of interest to match features

Integral Image

- Define the Integral Image

$$I'(x,y) = \sum_{\substack{x' \le x \\ y' \le y}} I(x',y')$$

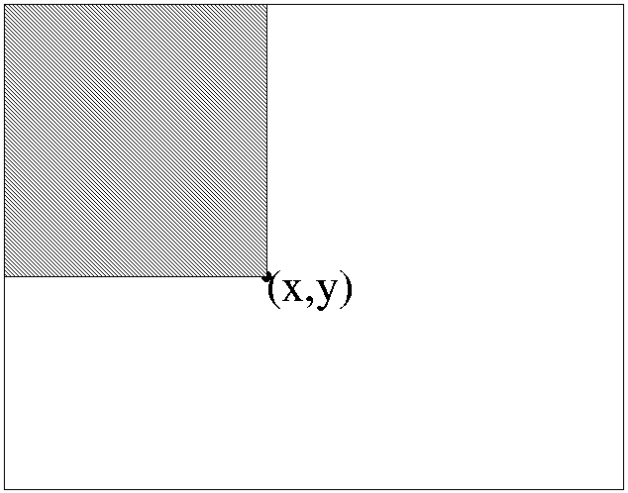
- Any rectangular sum can be computed in constant time:

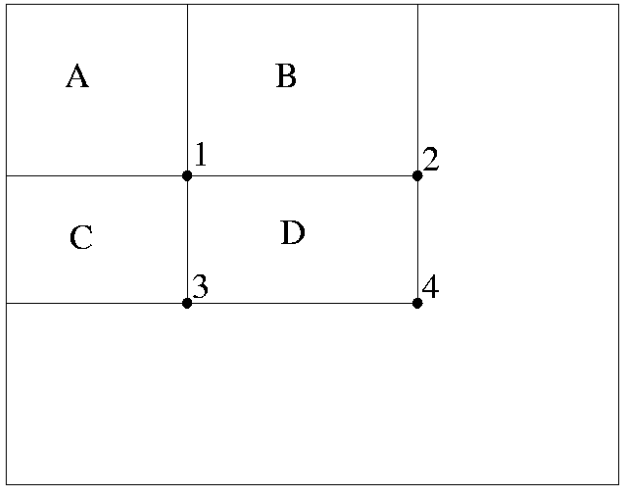
$$D = 1 + 4 - (2 + 3)$$

$$= A + (A + B + C + D) - (A + C + A + B)$$

$$= D$$

- Rectangle features can be computed as differences between rectangles

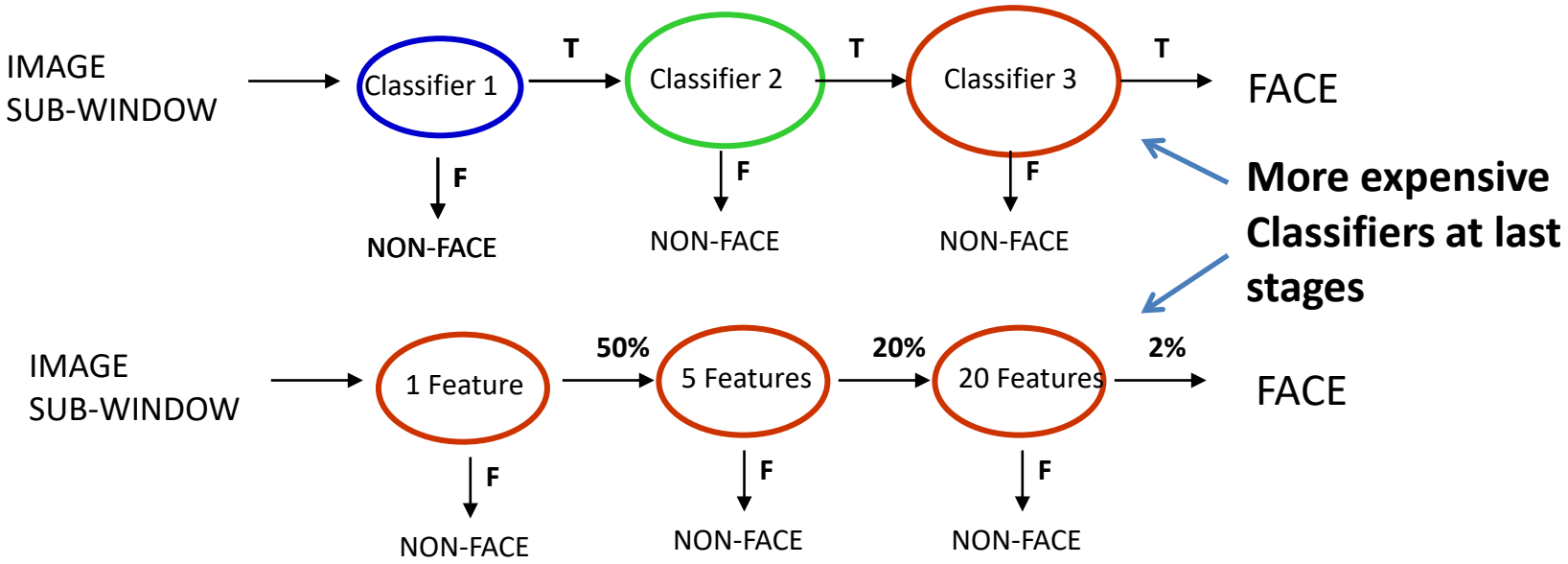




Features from different modalities

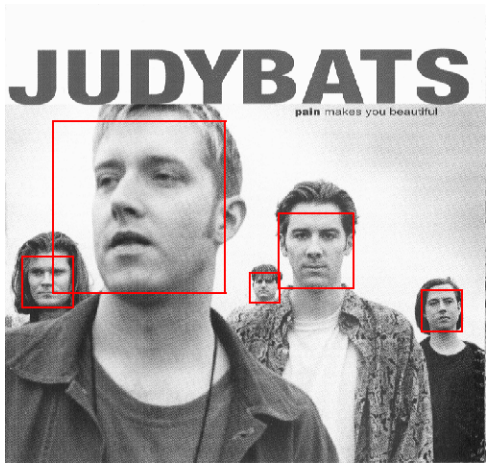
B) Describe points of interest to match features

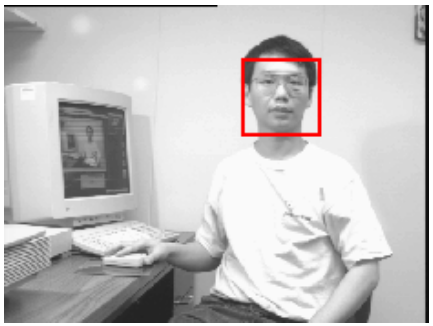
Cascade of Adaboost classifiers

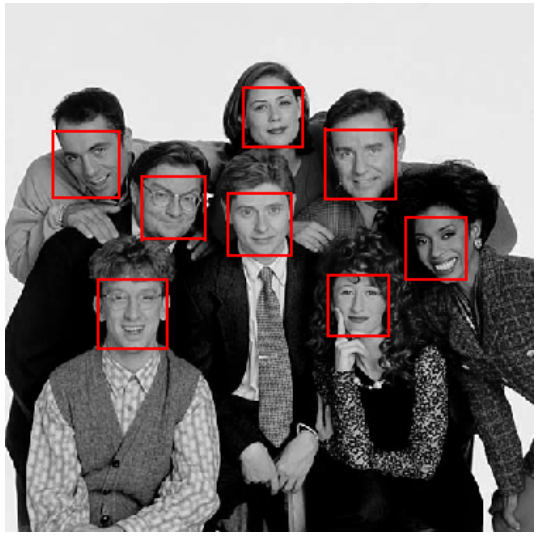


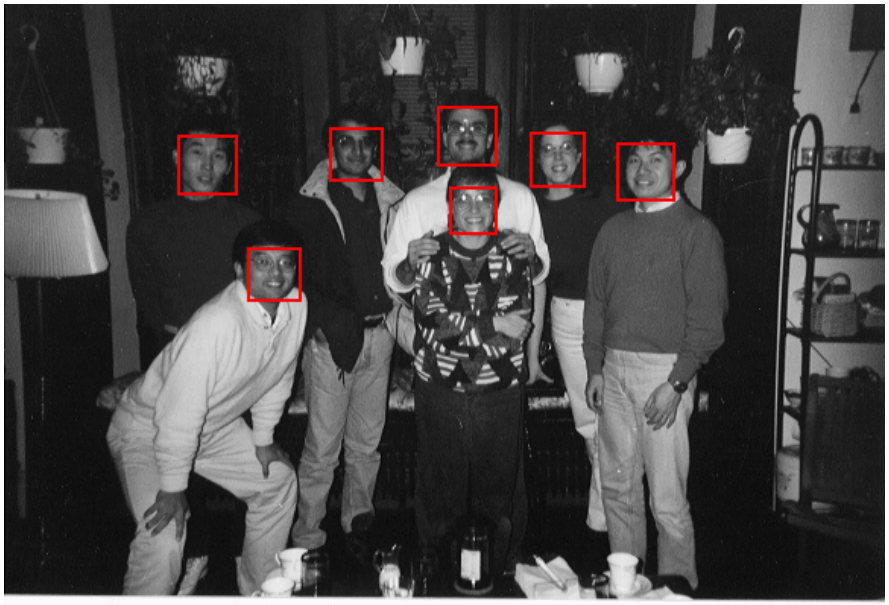
- A 1 feature classifier achieves 100% detection rate and about 50% false positive rate.
- A 5 feature classifier achieves 100% detection rate and 40% false positive rate (20% cumulative)
  - using data from previous stage.
- A 20 feature classifier achieve 100% detection rate with 10% false positive rate (2% cumulative)

B) Describe points of interest to match features









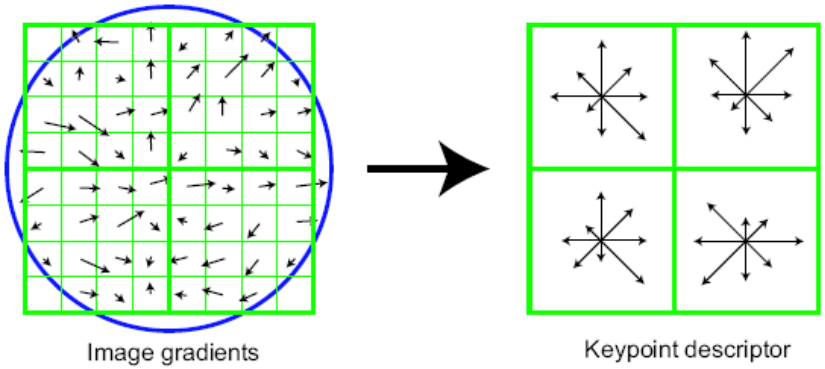
Features from different modalities

B) Describe points of interest to match features

Histograms of oriented gradients

Statistical basis

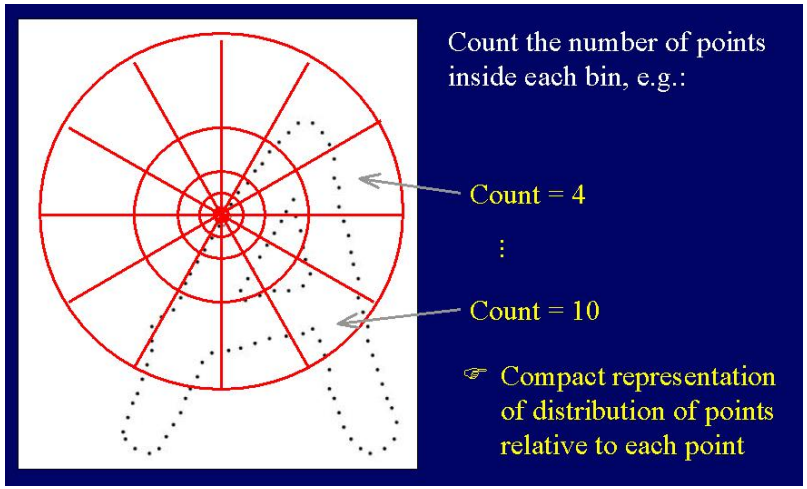
SIFT, D. Lowe, ICCV 1999



Structural basis

Shape context

Belongie, Malik, Puzicha, NIPS 2000



Features from different modalities

B) Describe points of interest to match features

Histograms of Oriented Gradients for Human Detection

Navneet Dalal and Bill Triggs

INRIA Rhône-Alps, 655 avenue de l'Europe, Montbonnot 38334, France
{Navneet.Dalal,Bill.Triggs}@inrialpes.fr, http://lear.inrialpes.fr

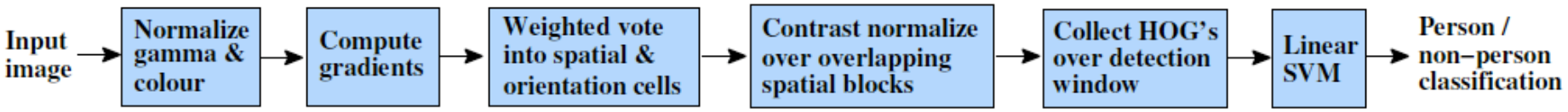
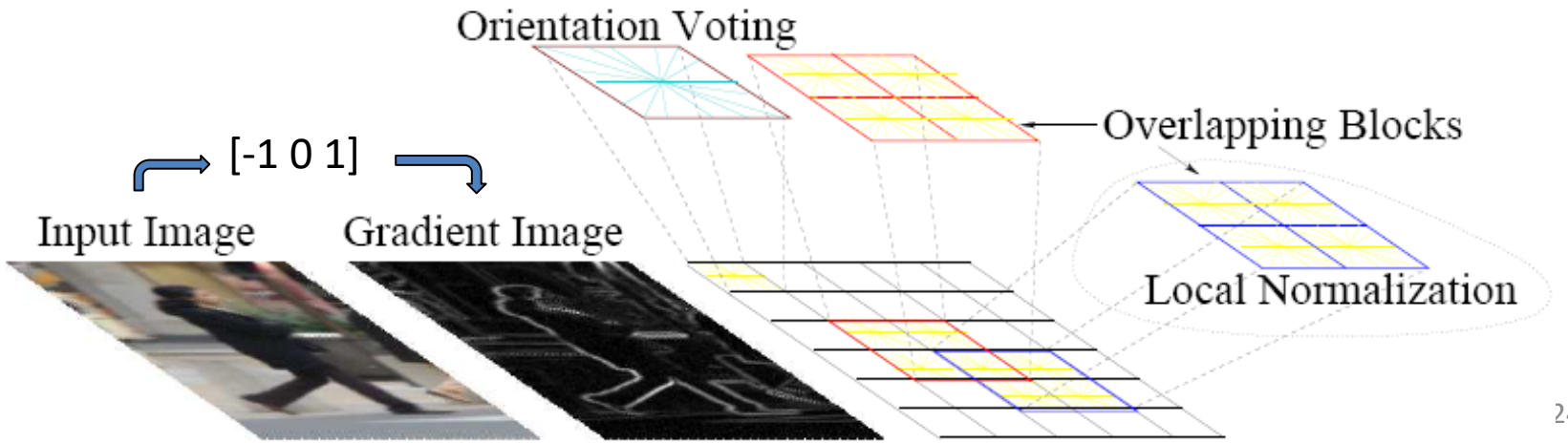


Figure 1. An overview of our feature extraction and object detection chain. The detector window is tiled with a grid of overlapping blocks in which Histogram of Oriented Gradient feature vectors are extracted. The combined vectors are fed to a linear SVM for object/non-object classification. The detection window is scanned across the image at all positions and scales, and conventional non-maximum suppression is run on the output pyramid to detect object instances, but this paper concentrates on the feature extraction process.



Features from different modalities

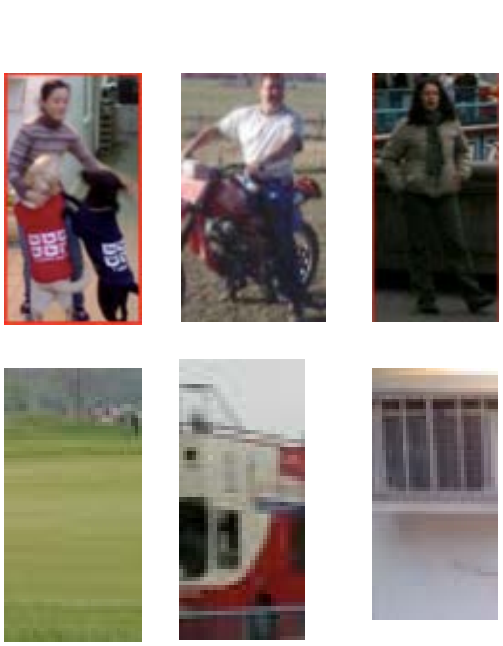
B) Describe points of interest to match features

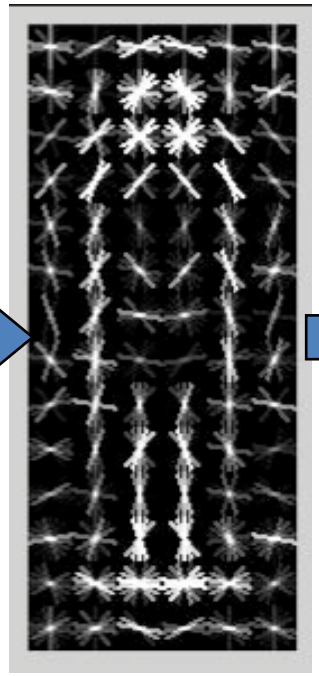
Dalal and Triggs CVPR05 (HOG)

Papageorgiou and Poggio ICIP99 (wavelets)

pos

neg





w **More weights for particular orientations based on SVM block weights**

w = weights for orientation and spatial bins

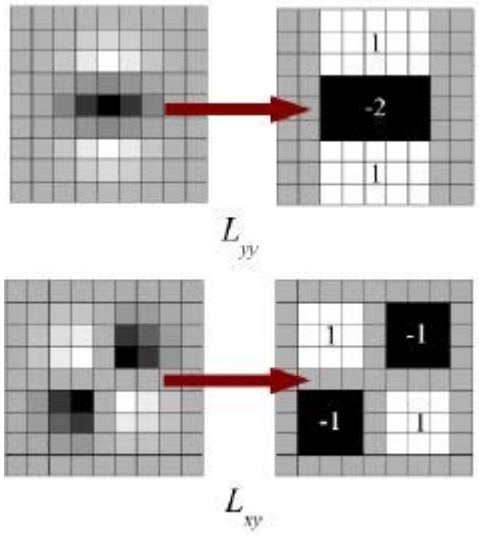
$w \cdot x > 0$

Train with a linear classifier (perceptron, logistic regression, SVMs...)

Features from different modalities

B) Describe points of interest to match features

Surf: Speeded up robust features
H Bay, T Tuytelaars, L Van Gool
Computer vision–ECCV 2006, 404-417

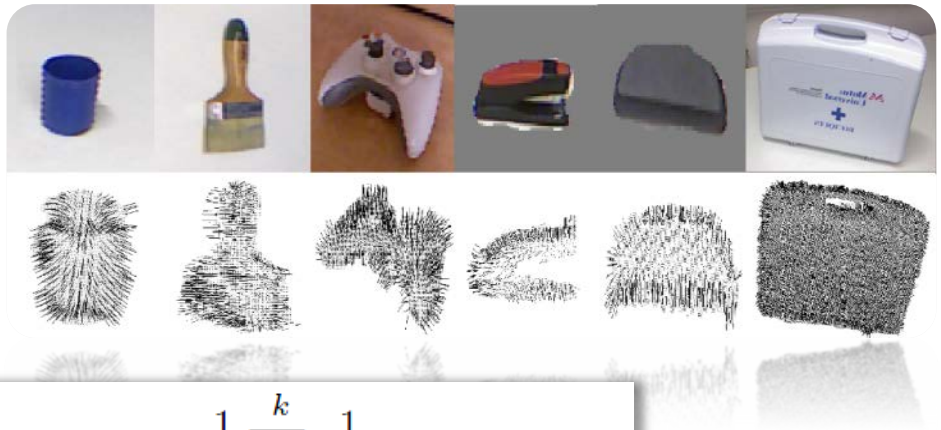


DoF -> haar-like style discrete approximation

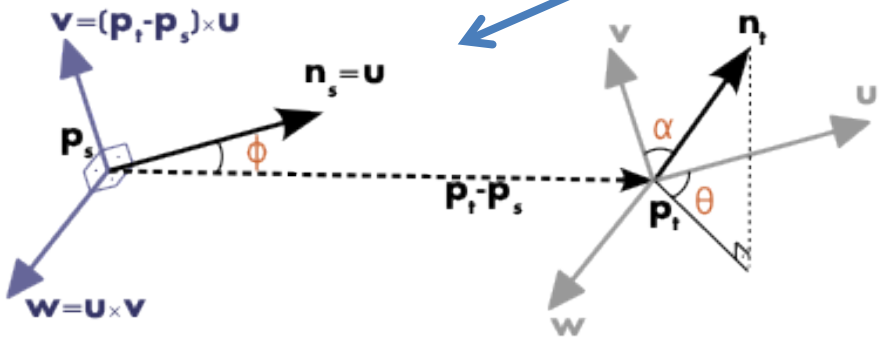
Sift-based descriptor -> -> haar-like style discrete approximation

Depth-based features

- ***Fast Point Feature Histogram*** (***FPFH***) method ***FPFH*** method is integrated in the Point Cloud Library (PCL).



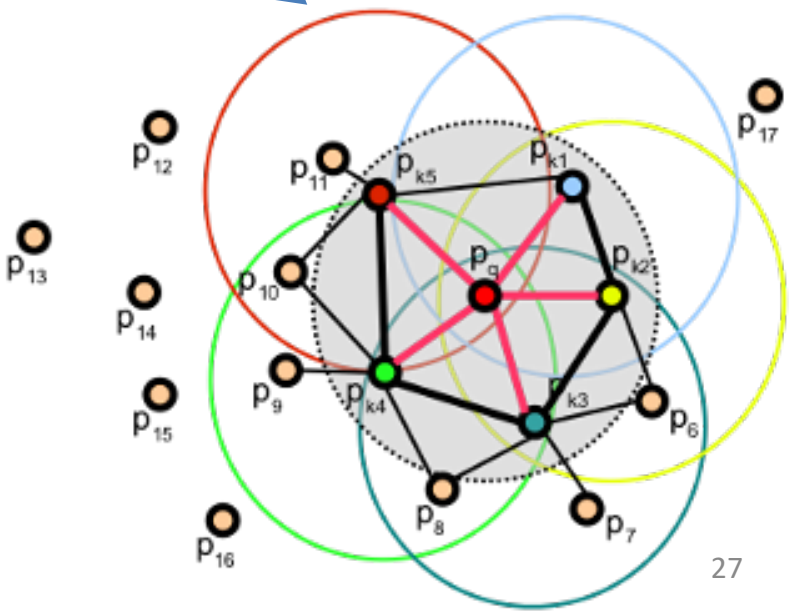
$$FPFH(\mathbf{p}_q) = SPFH(\mathbf{p}_q) + \frac{1}{k}\sum_{i=1}^{k}\frac{1}{\omega_k}\cdot SPFH(\mathbf{p}_k)$$



$$\alpha = \mathbf{v}\cdot \boldsymbol{n}_t$$

$$\phi = \mathbf{u}\cdot\frac{(\boldsymbol{p}_t - \boldsymbol{p}_s)}{d}$$

$$\theta = \arctan(\mathbf{w}\cdot \boldsymbol{n}_t, \mathbf{u}\cdot \boldsymbol{n}_t)$$



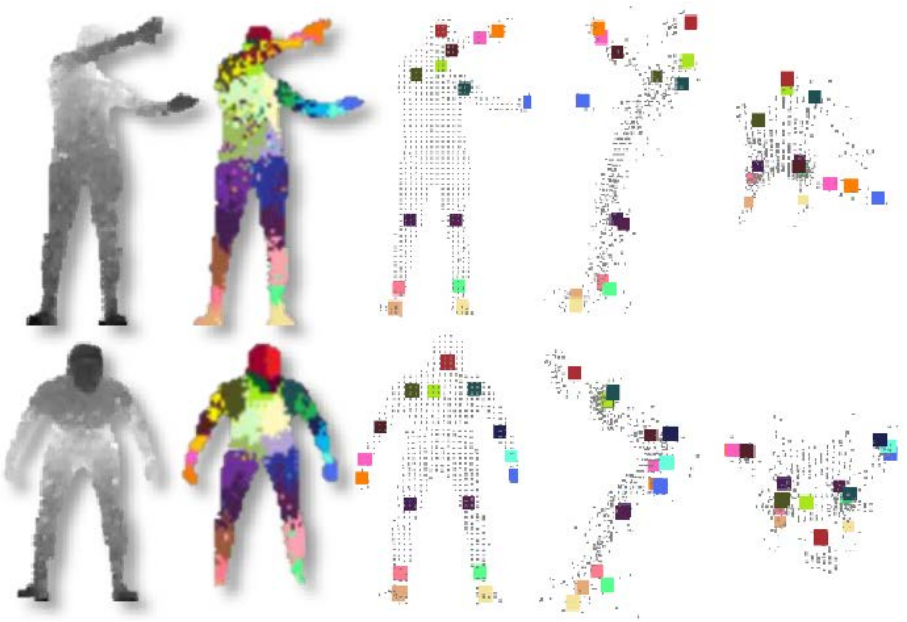
Features from different modalities

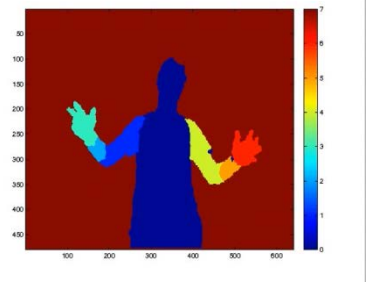
B) Describe points of interest to match features

Depth-based features

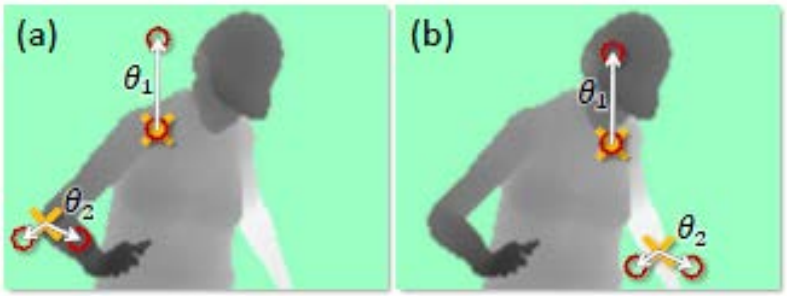
Real-Time Human Pose Recognition in Parts from Single Depth Images

Jamie Shotton Andrew Fitzgibbon Mat Cook Toby Sharp Mark Finocchio
Richard Moore Alex Kipman Andrew Blake
Microsoft Research Cambridge & Xbox Incubation

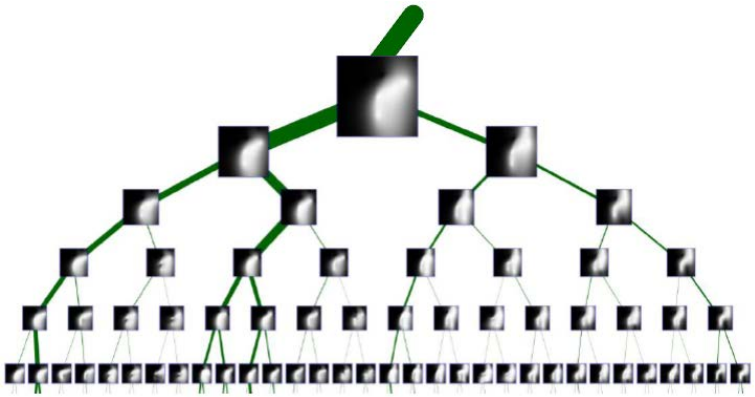




$$f_\theta(I,\mathbf{x}) = d_I\left(\mathbf{x} + \frac{\mathbf{u}}{d_I(\mathbf{x})}\right) - d_I\left(\mathbf{x} + \frac{\mathbf{v}}{d_I(\mathbf{x})}\right)$$



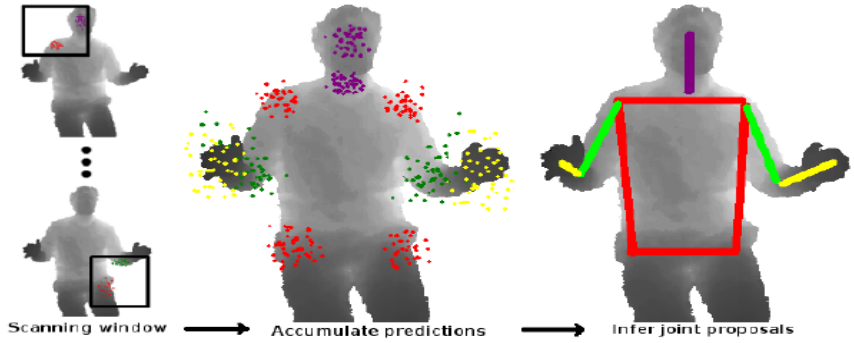
$$P(c|I,\mathbf{x}) = \frac{1}{T}\sum_{t=1}^{T} P_t(c|I,\mathbf{x})$$

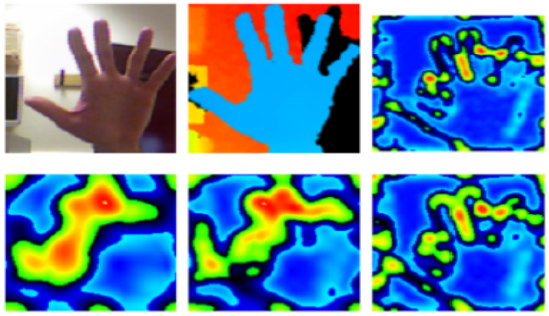


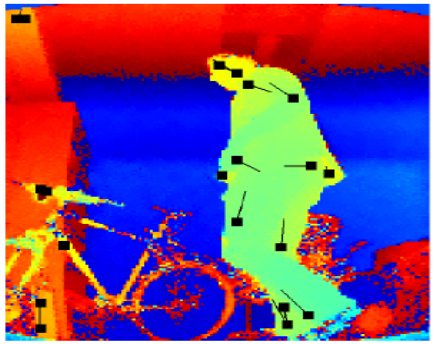
Features from different modalities

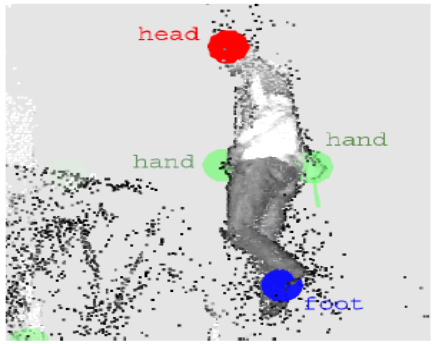
B) Describe points of interest to match features

Depth-based features









B. Holt, E.-J. Ong, H. Cooper, R. Bowden, Putting the pieces together: Connected poselets for human pose estimation, in: ICCV, 2011. **SVM Votes from depth derivatives**

N. Pugeault, R. Bowden, Spelling it out: Real-time asl fingerspelling recognition, in: ICCV, 2011. **Gabor filters on depth maps**

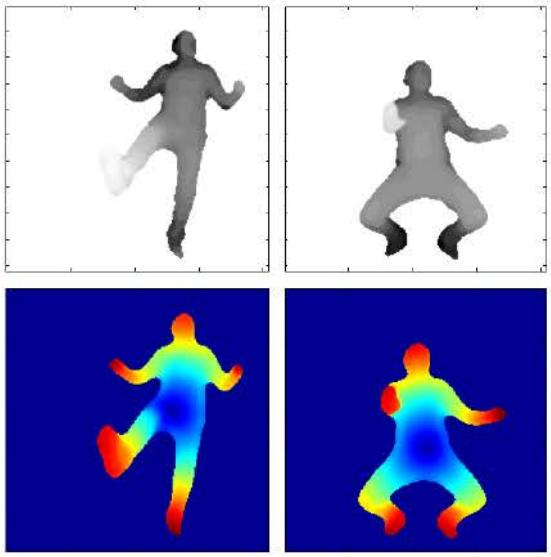
C. Plagemann, V. Ganapathi, D. Koller, S. Thrun, Real-time identication and localization of body parts from depth images, in: ICCV, 2011, pp. 3108{3113. **Extrema of geodesic maps over depth maps including orientation**

A. Clapes, M. Reyes, S. Escalera, User identication and object recognition in clutter scenes based on rgb-depth analysis, in: Articulated Motion of Deformable Objects, 2012. **FPFH descriptors**

Features from different modalities

B) Describe points of interest to match features

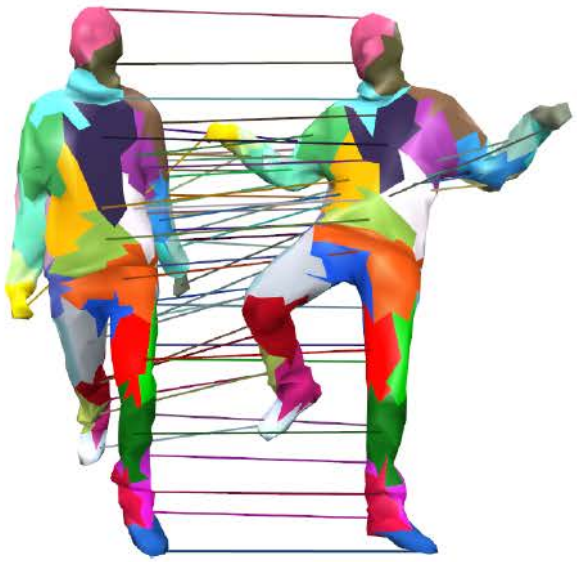
Depth-based features



[1]

[2]

[3,4]



[5]

[1] L. Schwarz, A. Mkhitaryan, D. Mateus, N. Navab, Estimating human 3d pose from time-of-flight images based on geodesic distances and optical flow, in: IEEE Conference on Automatic Face and Gesture Recognition (FG), 2011.

[2] V. Ganapathiand, C. Plagemann, D. Koller, S. Thrun, Real time motion capture using a single time-of-flight camera, in: CVPR, 2010, pp. 755–762.

[3] C. Keskin, F.racC, Y. Kara, L. Akarun, Real time hand pose estimation using depth sensors, in: ICCV, 2011.

[4] D. Minnen, Z. Zafrulla, Towards robust cross-user hand tracking and shape recognition, in: ICCV, 2011, pp. 1235–1241.

[5] T. Windheuser, U. Schlickewei, F. R. Schmidt, Geometrically consistent elastic matching of 3d shapes: A linear programming solution, in: ICCV, 2011.

Geodesic maps

Surface points

Random Forest

Sillouhete

Linear programming methods

Several hybrid **discriminative** **generative** methods taking benefit from **multi-modal** **RGBDT** representation

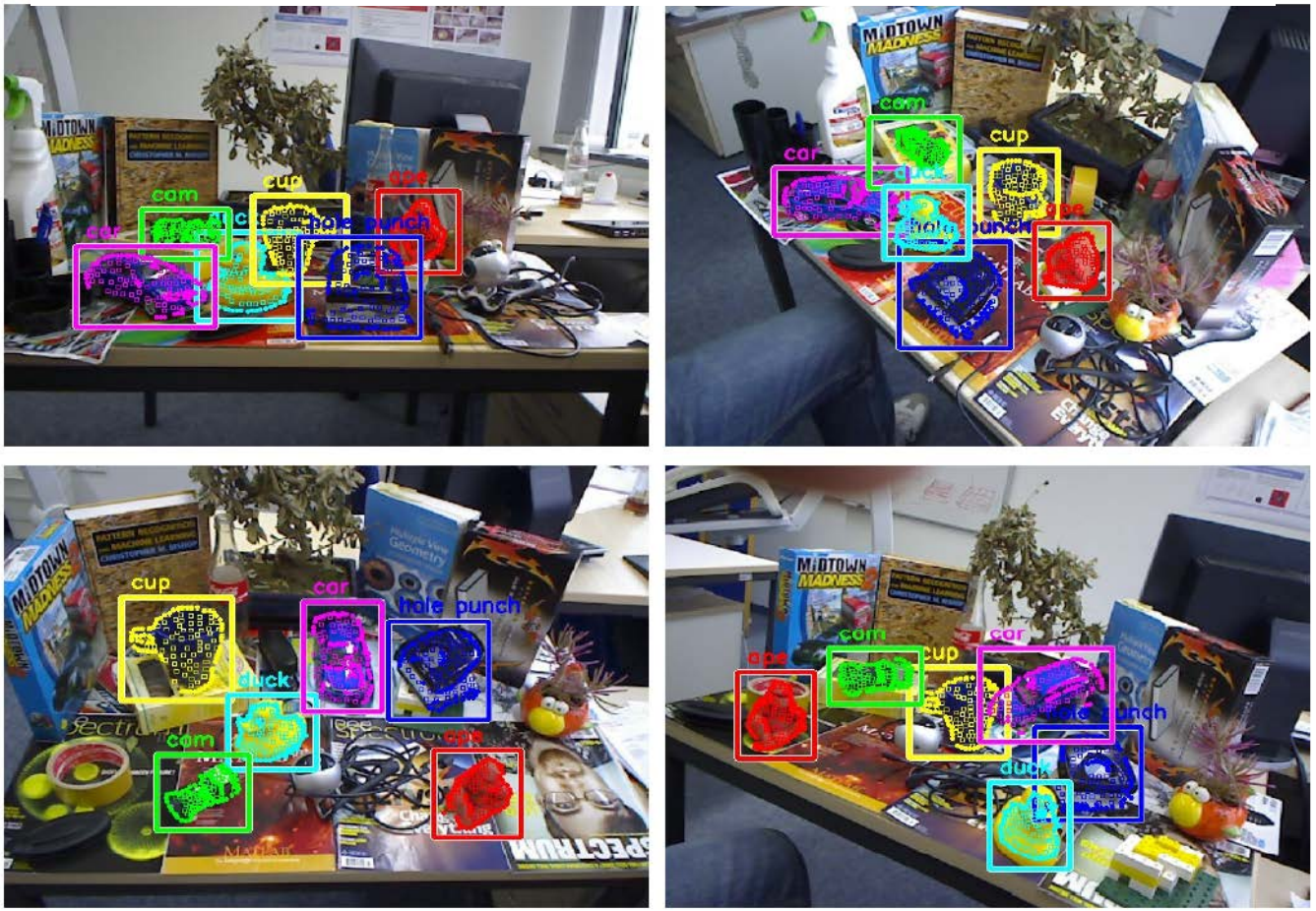
Features from different modalities

B) Describe points of interest to match features

Depth-based features

Multimodal Templates for Real-Time Detection of Texture-less Objects in Heavily Cluttered Scenes

Stefan Hinterstoisser[1], Stefan Holzer[1], Cedric Cagniart[1], Slobodan Ilic[1],
Kurt Konolige[2], Nassir Navab[1], Vincent Lepetit[3]



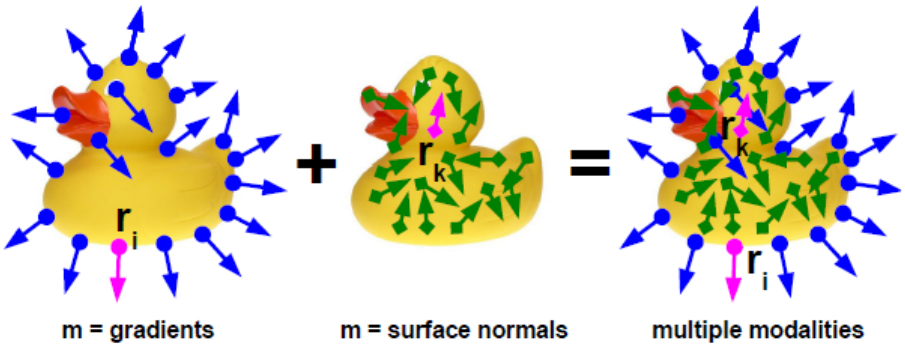
Features from different modalities

B) Describe points of interest to match features

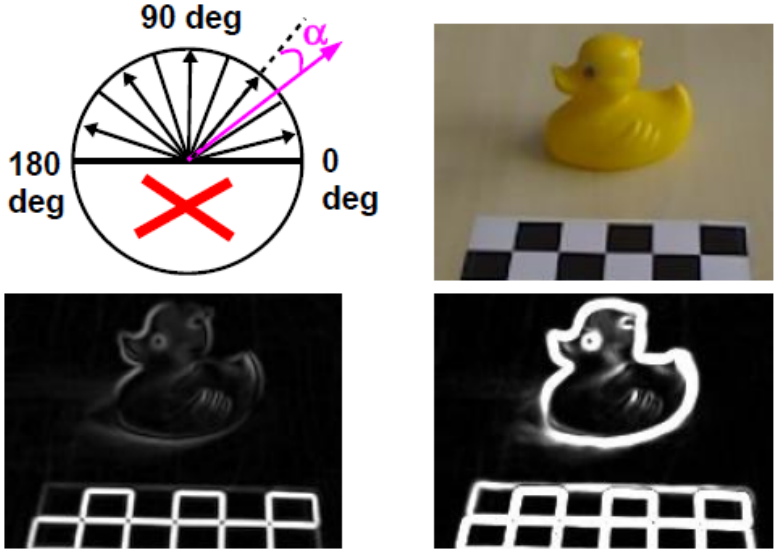
Depth-based features

Multimodal Templates for Real-Time Detection of Texture-less Objects in Heavily Cluttered Scenes

Stefan Hinterstoisser[1], Stefan Holzer[1], Cedric Cagniart[1], Slobodan Ilic[1], Kurt Konolige[2], Nassir Navab[1], Vincent Lepetit[3]



A toy duck with different modalities. **Left:** Image gradients are mainly found on the contour. The gradient location r_i is displayed in pink. **Middle:** Surface normals are found on the body of the duck. The normal location r_k is displayed in pink. **Right:** our approach combines multiple cues which are complementary: gradients are usually found on the object contour while surface normals are found on the object interior



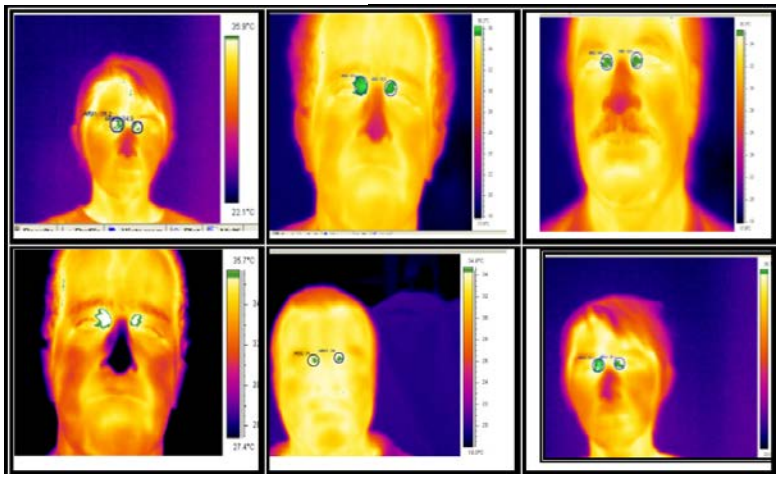
Upper Left: Quantizing the gradient orientations: the pink orientation is closest to the second bin. **Upper right:** A toy duck with a calibration pattern. **Lower Left:** The gradient image computed on a gray value image. The object contour is hardly visible. **Lower right:** Gradients computed with our approach. Details of the object contours are clearly visible.

Features from different modalities

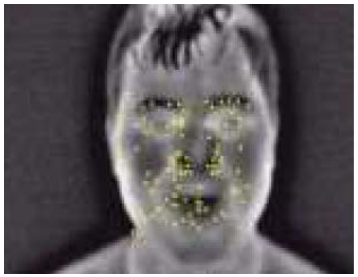
B) Describe points of interest to match features

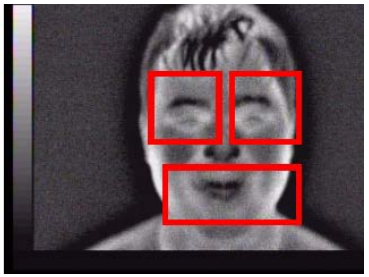
Thermal features

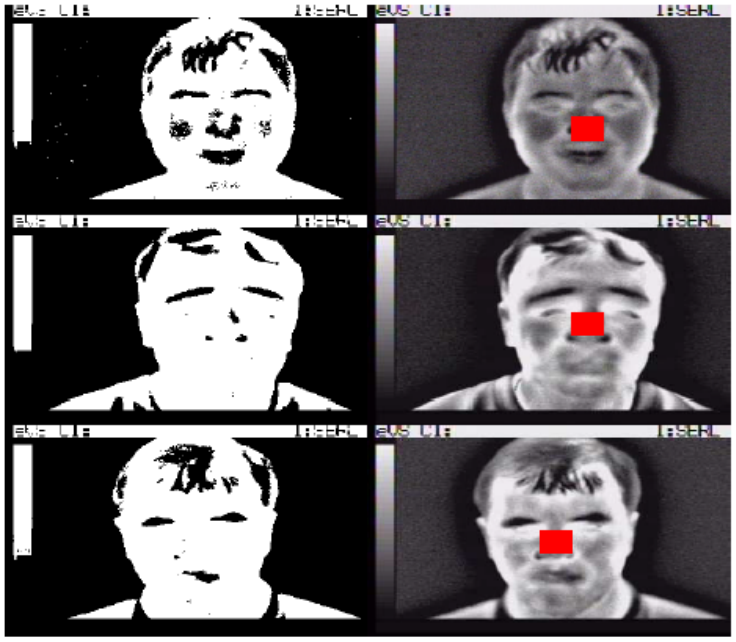
Mainly for particular objects detection/segmentation:
Use to be **intensity & gradient**



Eigenfaces from thermal

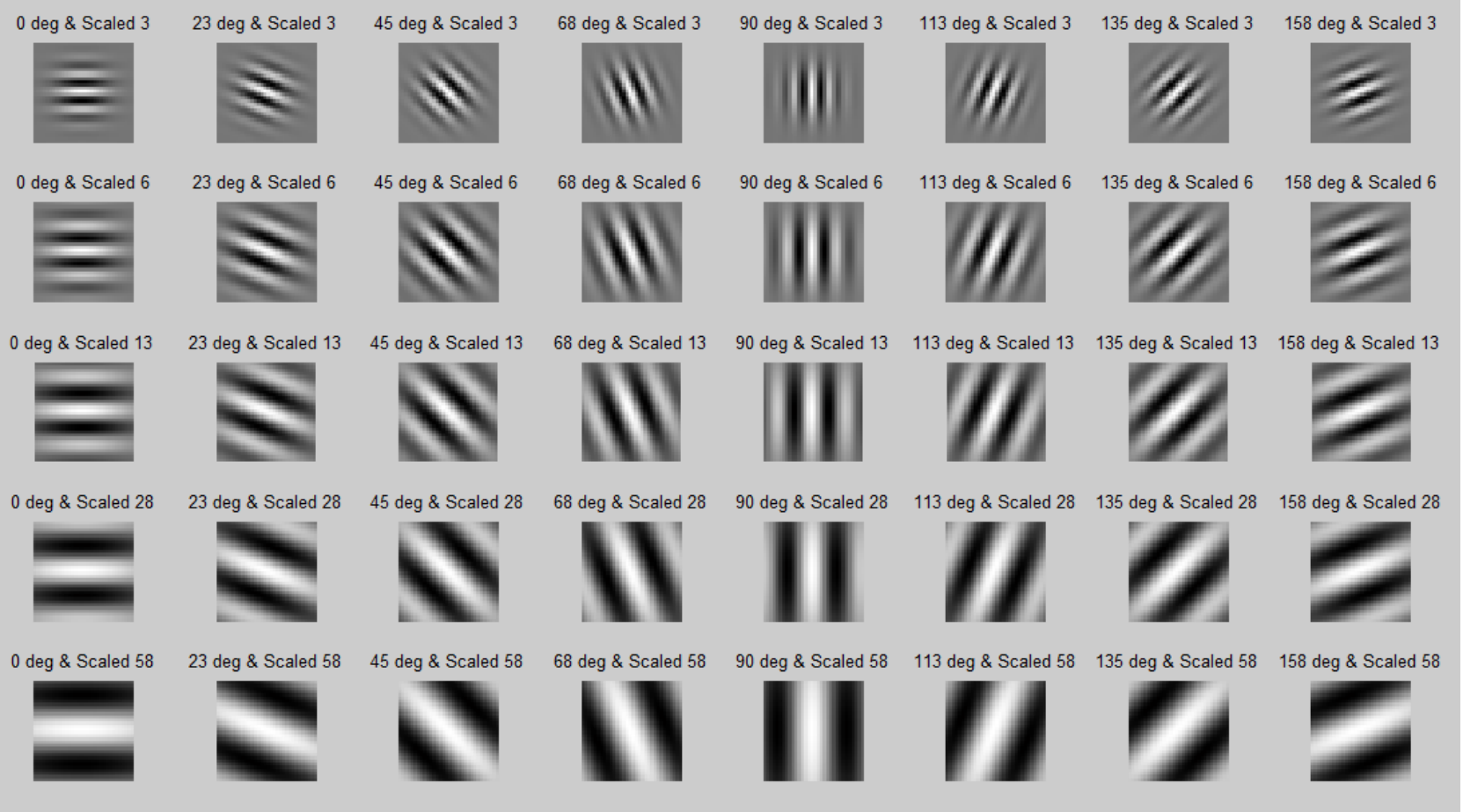






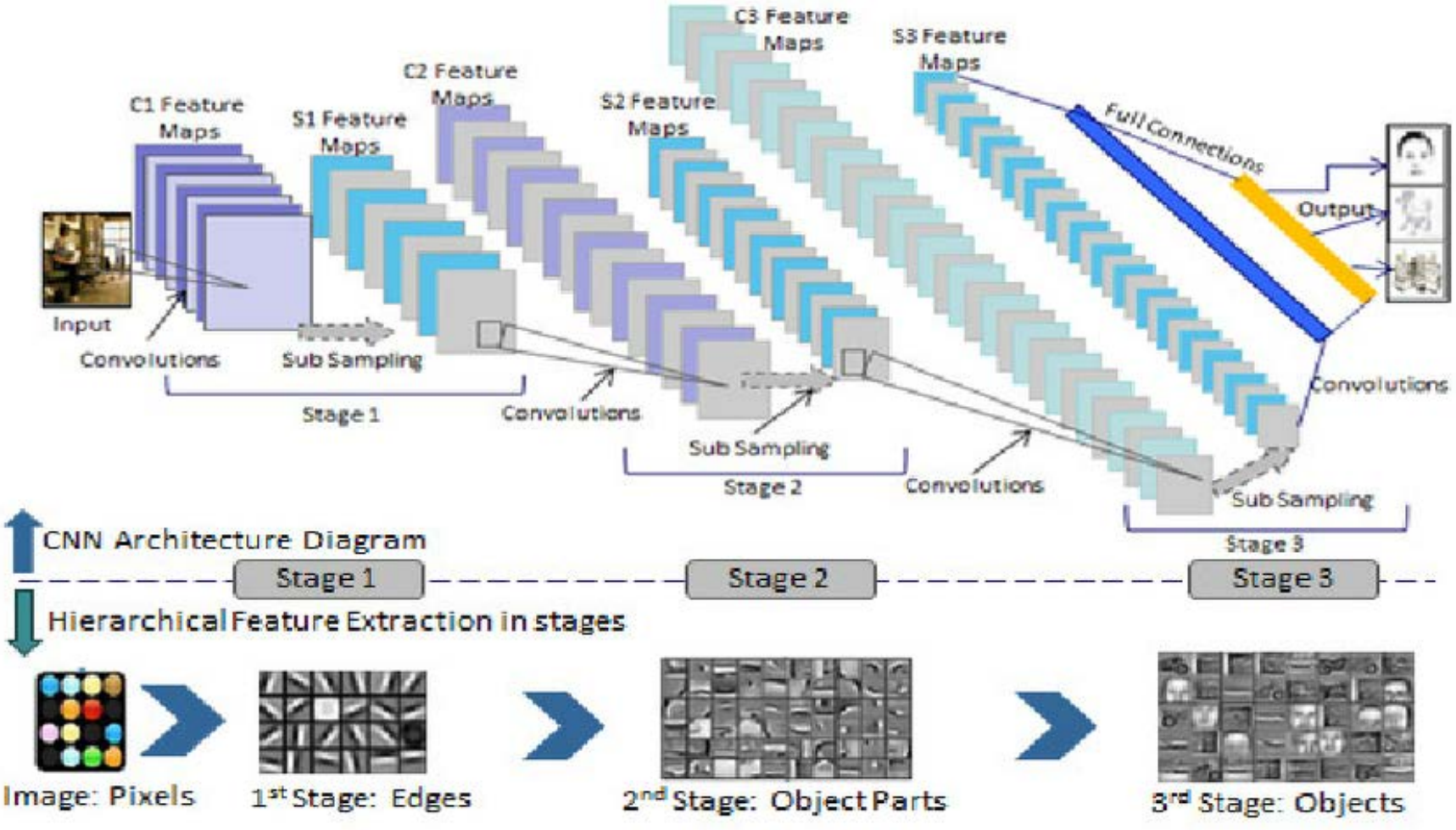
Features from different modalities

Gabor filters and CNN filters



Features from different modalities

Gabor filters and CNN filters



Fusion strategies

How to fuse features extracted from different modalities?

-This mostly depends on the “classification” strategy considered.

-Most times “concatenation” as fusion is just enough **(Early fusion).**

-A weighted output of modalities is sometimes a better choice **(Late fusion).**

-Middle fusion can also be considered by including complementary features/modalities in the middle steps of the model.

-Most state-of-the-art classification strategies can deal with both early and late fusion modalities.

-Iterative application of different feature modalities are known as **multi-modal** systems though they **do not perform data fusion**.