CS294-43: Visual Object and Activity Recognition

Prof. Trevor Darrell

Feb 3rd: Global Features

Today

- Demo from last week

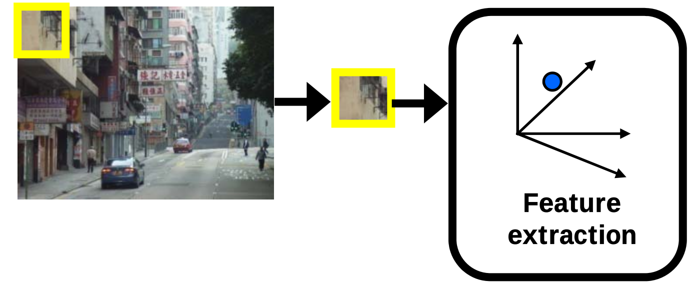
  reminder: email me with desired paper demo or pres.

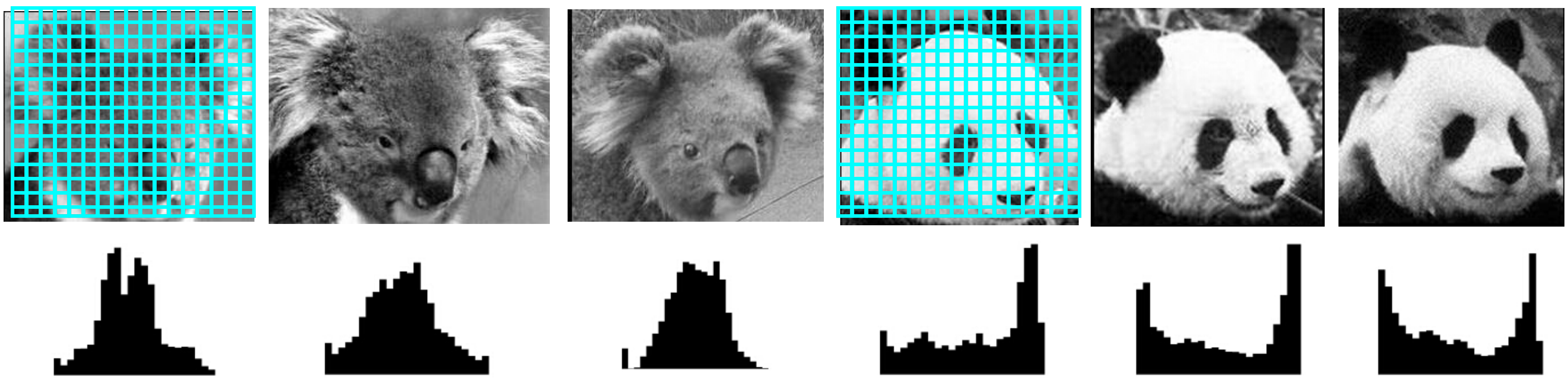
- Background / Overview

- Histograms of edges (Schiele)
- Windowed spectral analysis (Oliva)
- Tiled histograms of edges (Triggs)

- Motion History Images (Bobick)
- Rectified Flow Descriptors (Efros)
- Differential Geometry Signatures (Shah)

Feature extraction: global appearance





Simple holistic descriptions of image content

- **grayscale / color histogram**
- **vector of pixel intensities**

Slide credit: K. Grauman, B. Leibe

Eigenfaces: global appearance description

An early appearance-based approach to face recognition

Generate low-dimensional representation of appearance with a linear subspace.

Project new images to “face space”.

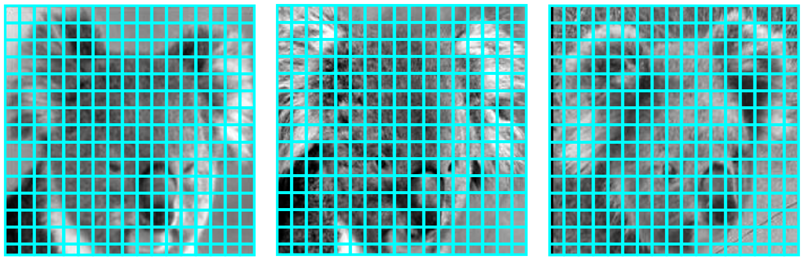
Recognition via nearest neighbors in face space

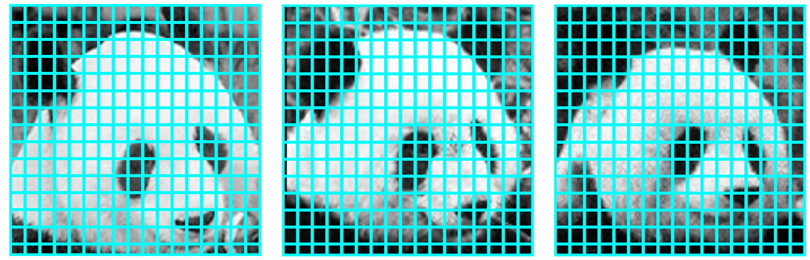
Turk & Pentland, 1991

Slide credit: K. Grauman, B. Leibe

Feature extraction: global appearance

- **Pixel-based representations sensitive to small shifts**



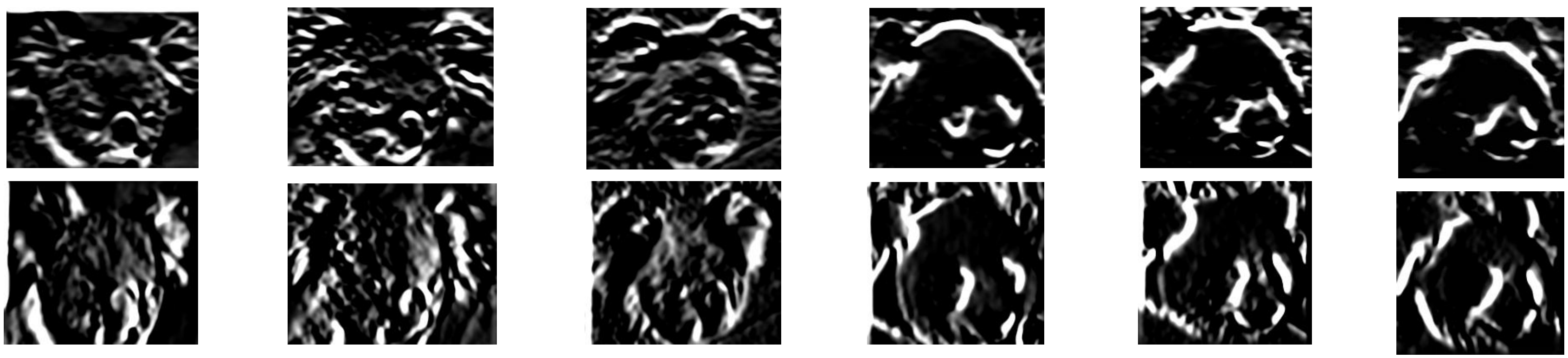


- **Color or grayscale-based appearance description can be sensitive to illumination and intra-class appearance variation**

Cartoon example: an albino koala

Gradient-based representations

- **Consider edges, contours, and (oriented) intensity gradients**



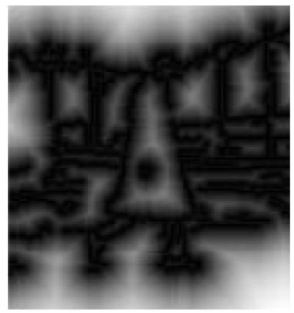
Slide credit: K. Grauman, B. Leibe

Gradient-based representations: Matching edge templates

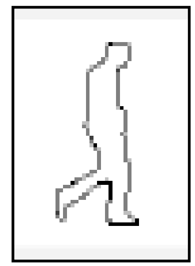
- **Example: Chamfer matching**

Input image

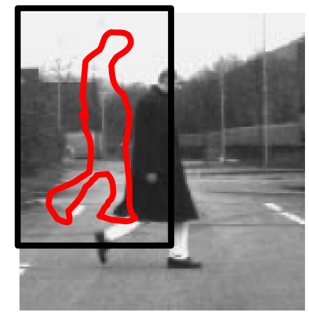
Edges detected



Distance transform



Template shape



Best match

At each window position, compute average min distance between points on template (T) and input (I).

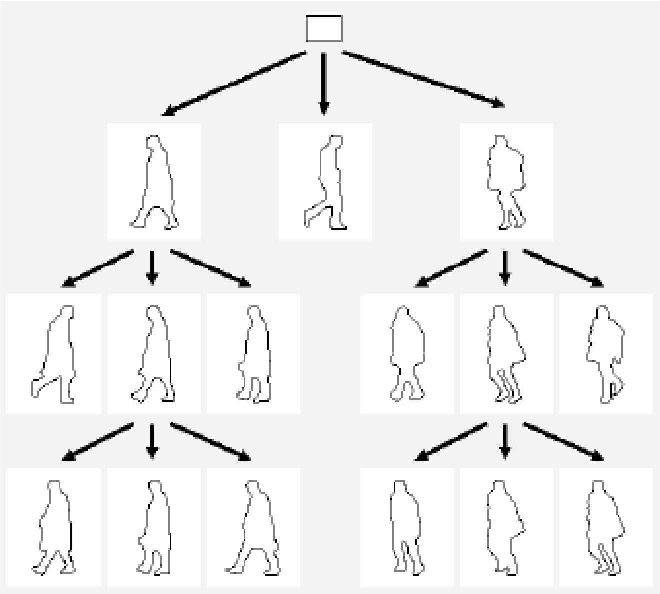
$$D_{chamfer}(T, I) \equiv \frac{1}{|T|} \sum_{t \in T} d_I(t)$$

Gavrila & Philomin ICCV 1999

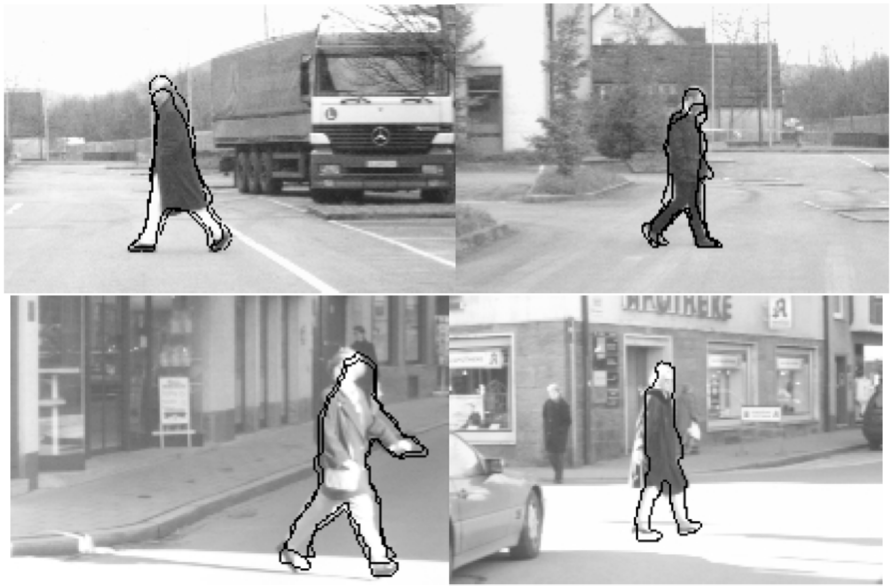
Slide credit: K. Grauman, B. Leibe

Gradient-based representations: Matching edge templates

- **Chamfer matching**

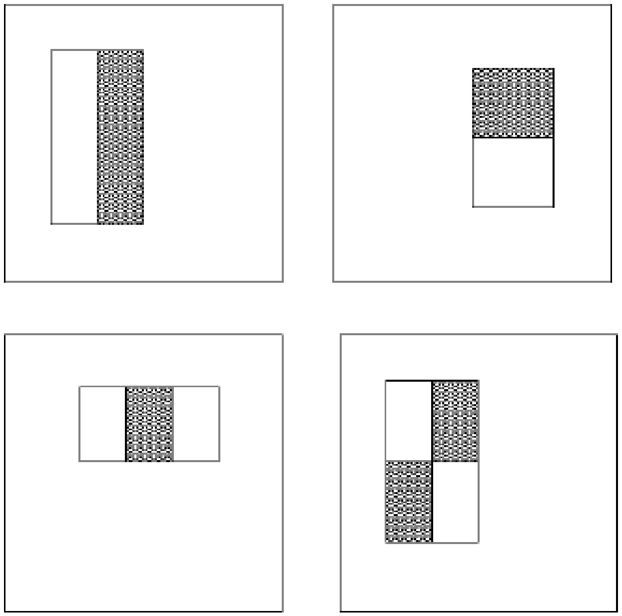


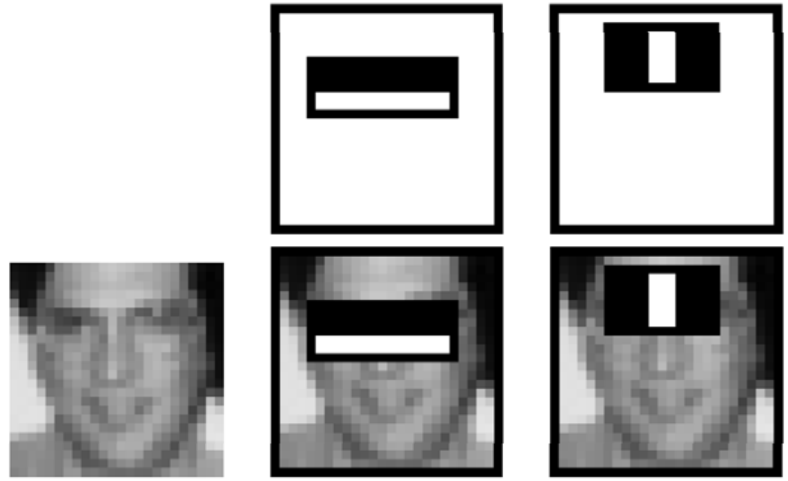
Hierarchy of templates



Gavrila & Philomin ICCV 1999

Gradient-based representations: Rectangular features





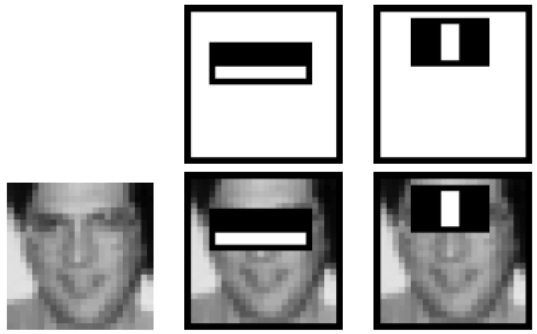
Compute differences between sums of pixels in rectangles

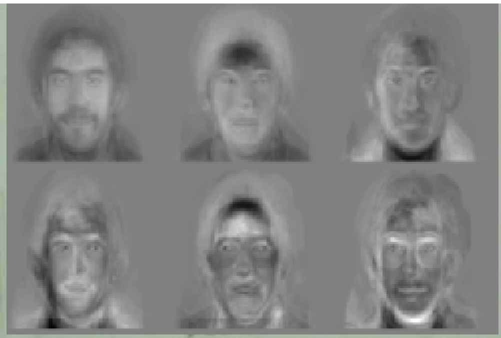
Captures contrast in adjacent spatial regions

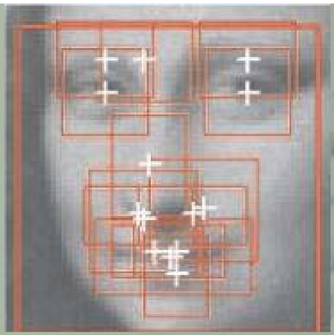
Similar to Haar wavelets, efficient to compute

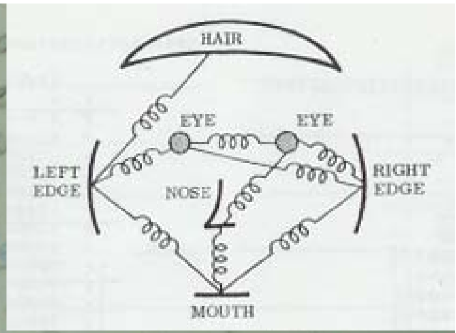
Viola & Jones, CVPR 2001

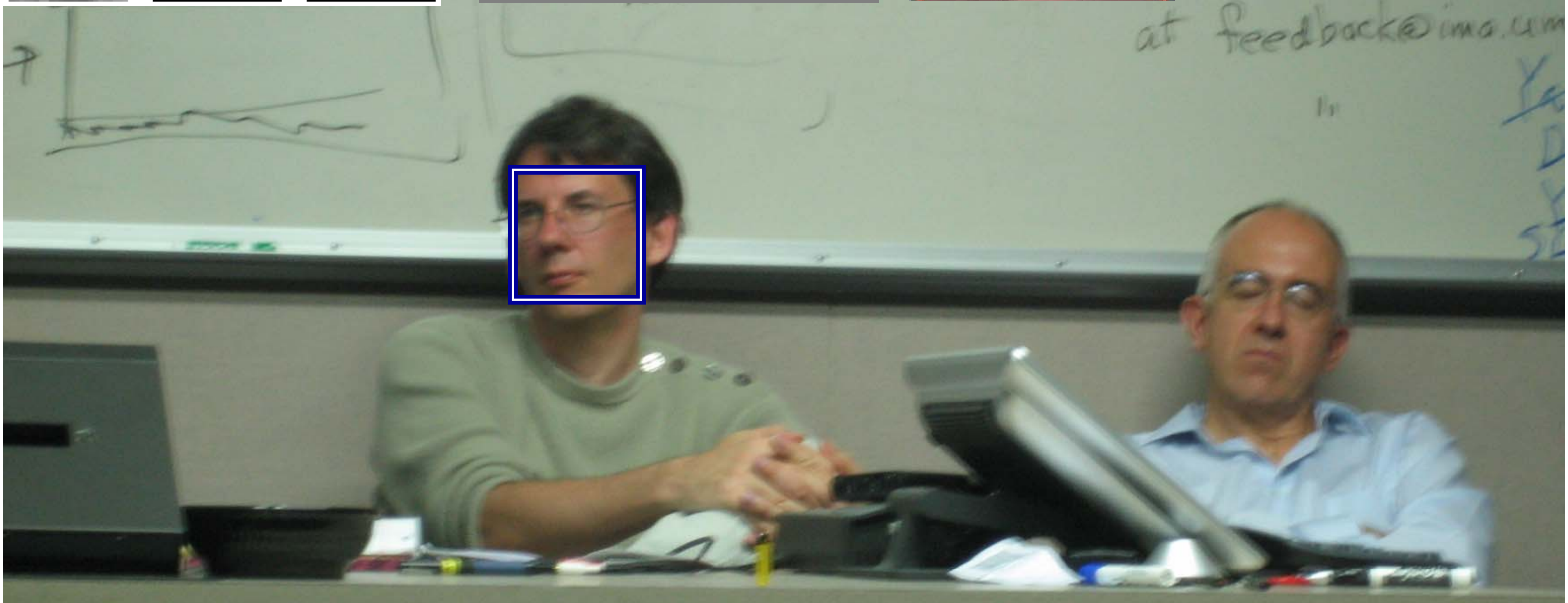
Slide credit: K. Grauman, B. Leibe









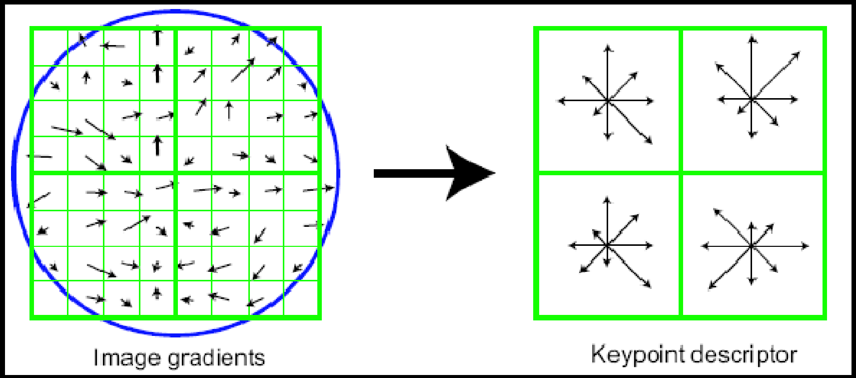


- The representation and matching of pictorial structures Fischler, Elschlager (1973)
- Face recognition using eigenfaces M. Turk and A. Pentland (1991).
- Human Face Detection in Visual Scenes - Rowley, Baluja, Kanade (1995)
- Graded Learning for Object Detection - Fleuret, Geman (1999)
- Robust Real-time Object Detection - Viola, Jones (2001)
- Feature Reduction and Hierarchy of Classifiers for Fast Object Detection in Video Images - Heisele, Serre, Mukherjee, Poggio (2001)
- ….

Slide credit: A. Torralba

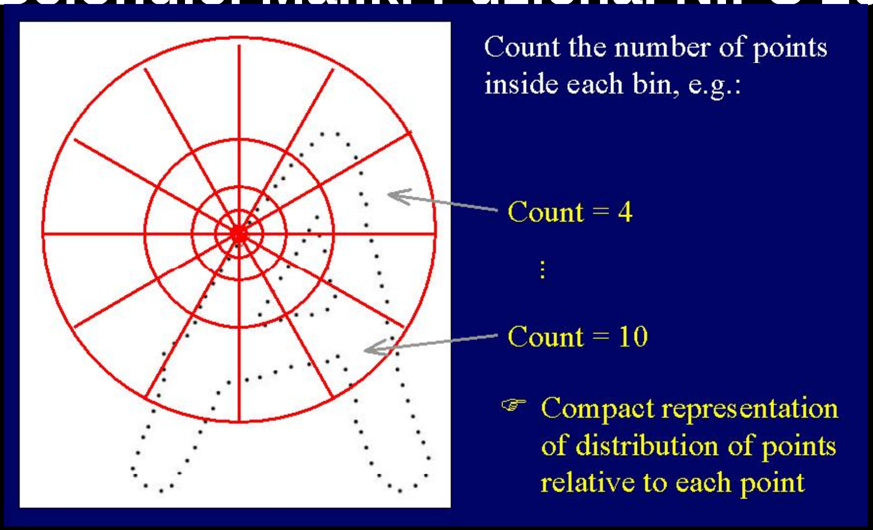
Histograms of oriented gradients

- **SIFT, D. Lowe, ICCV 1999**

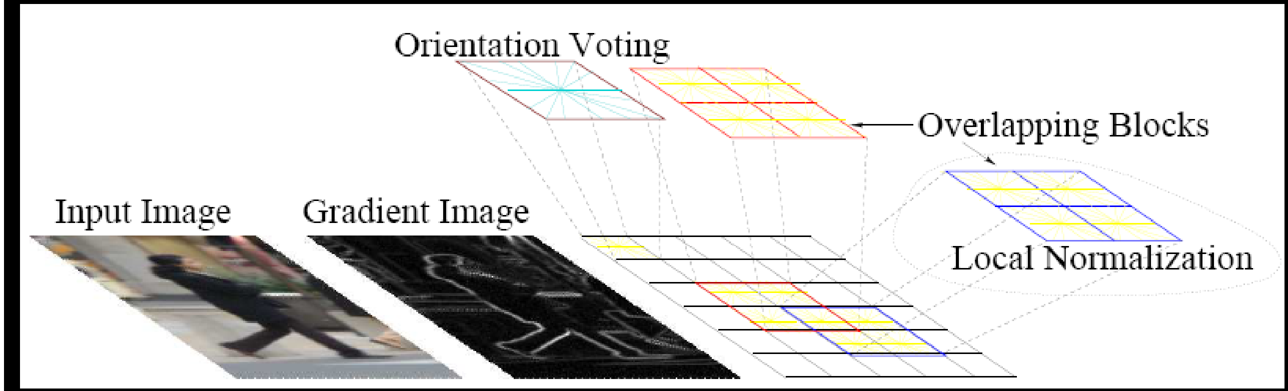


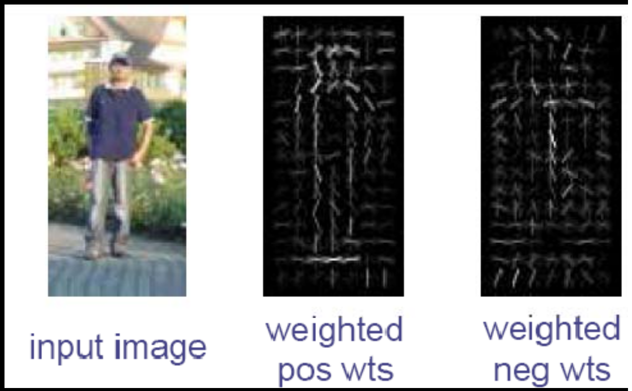
- **Shape context**

Belongie, Malik, Puzicha, NIPS 2000



- **Dalal & Trigs, 2006**





Slide credit: A. Torralba

Histograms of Gradients, ca. 1996

- Schiele and Crowley

Object Recognition using Multidimensional Receptive Field Histograms

Bernt Schiele and James L. Crowley

LIFIA/GRAVIR, 46 Ave Félix Viallet, 38031 Grenoble, France

Abstract. This paper presents a technique to determine the identity of objects in a scene using histograms of the responses of a vector of local linear neighborhood operators (receptive fields). This technique can be used to determine the most probable objects in a scene, independent of the object's position, image-plane orientation and scale. In this paper we describe the mathematical foundations of the technique and present the results of experiments which compare robustness and recognition rates for different local neighborhood operators and histogram similarity measurements.

- Freeman and Roth

Orientation Histograms for Hand Gesture Recognition

William T. Freeman and Michal Roth
Mitsubishi Electric Research Labs
201 Broadway
Cambridge, MA 02139 USA
e-mail: {freeman, roth}@merl.com

From: IEEE Intl. Wkshp. on Automatic Face and Gesture Recognition, Zurich, June, 1995.

Abstract

We present a method to recognize hand gestures, based on a pattern recognition technique developed by McConnell [16] employing histograms of local orientation. We use the orientation histogram as a feature vector for gesture classfication and interpolation.

This method is simple and fast to compute, and

the special glove. We seek a visually based method which will be free of gloves and wires.

Relying on visual markings on the hands, previous researchers have recognized sign language and pointing gestures [24, 5, 8]. However, these methods require the placement of markers on the hands.

The marking-free systems of [12, 21] can recognize specific finger or pointing events, but not general gestures. Employing special hardware or off-

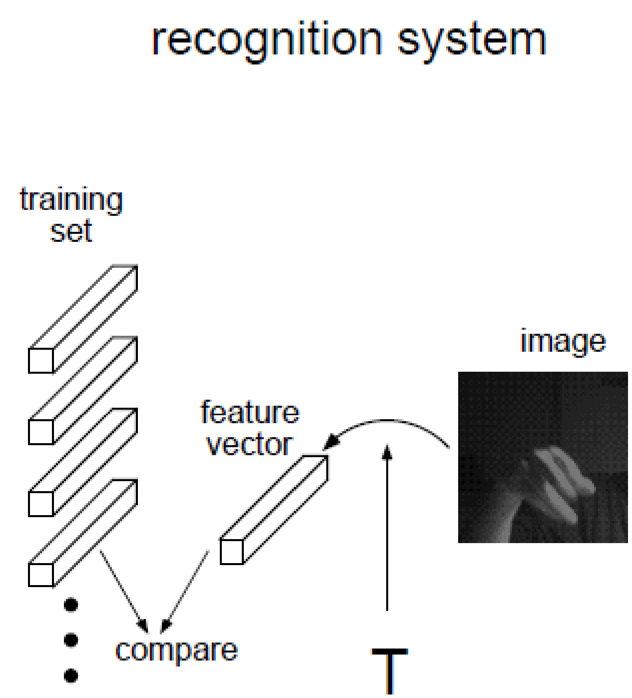


Figure 1: Outline of the recognition system. We apply some transformation T to the image data to form a feature vector which represents that particular gesture. To classify the gesture, we compare the feature fector with the feature vectors from a previously generated training set. For dynamic gesture recognition, the input would be a sequence of images.

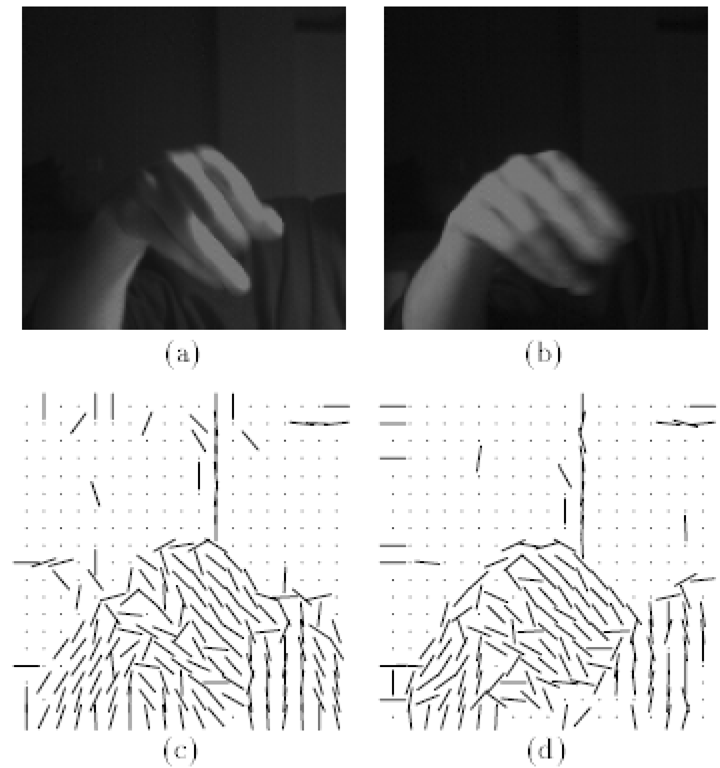
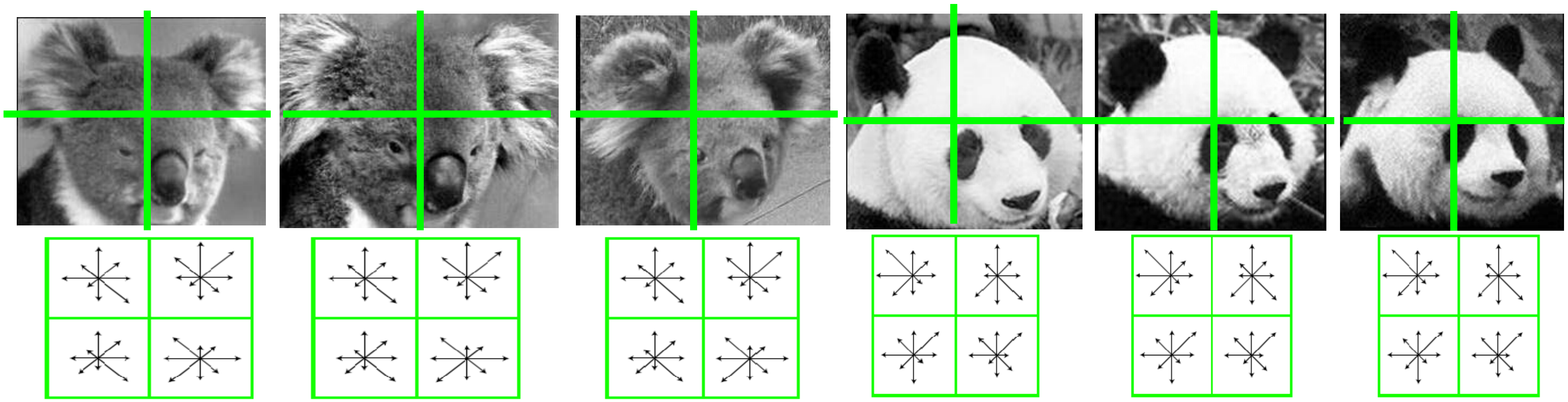


Figure 2: Showing the robustness of local orientation to lighting changes. Pixel intensities are sensitive to lighting change. (a) and (b) show the same hand gesture illuminated under two different lighting conditions. The pixel intensities change significantly as the lighting changes. Maps of local orientation, (c) and (d), are more stable. (The orientation maps were computed using steerable filters [10]. Orientation bars below a contrast threshold are suppressed.)

Gradient-based representations

- **Consider edges, contours, and (oriented) intensity gradients**



- **Summarize local distribution of gradients with histogram**
  - **Locally orderless: offers invariance to small shifts and rotations**
  - **Contrast-normalization: try to correct for variable illumination**

Slide credit: K. Grauman, B. Leibe

Today

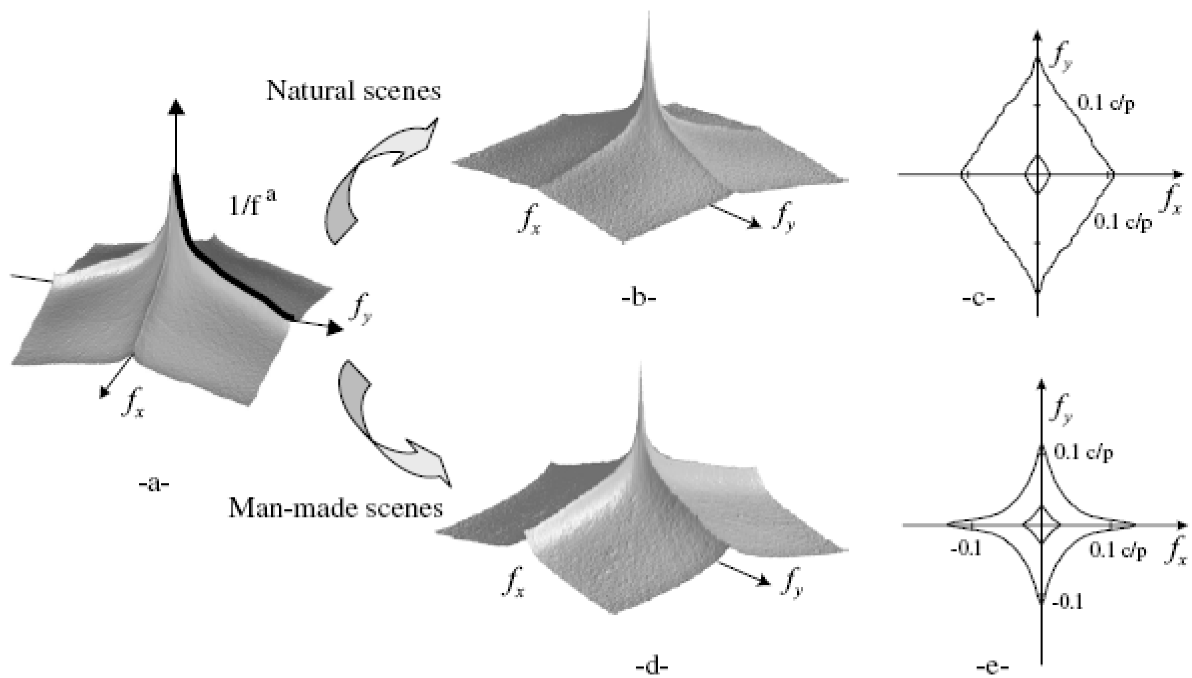
- Background / Overview

- Histograms of edges (Schiele)
- **Windowed spectral analysis (GIST)**
- Tiled histograms of edges (HOG)

- Motion History Images (Bobick)
- Rectified Flow Descriptors (Efros)
- Differential Geometry Signatures (Shah)

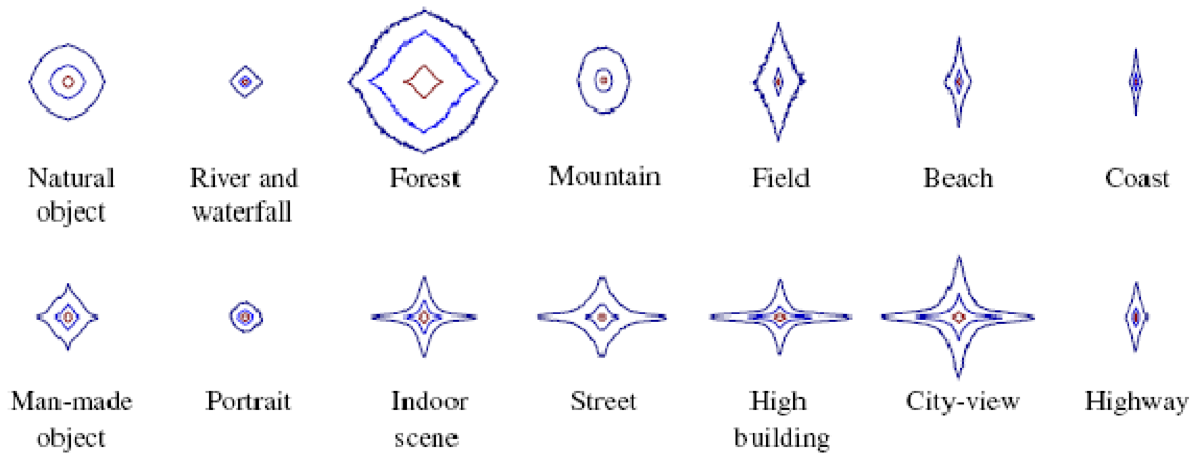
Key Point of Torralba/Oliva Papers

Natural Image statistics depend on the interaction between the observer and the world:



Slide Credit: Torralba, Olivia, J. Huang

Spectral Signatures



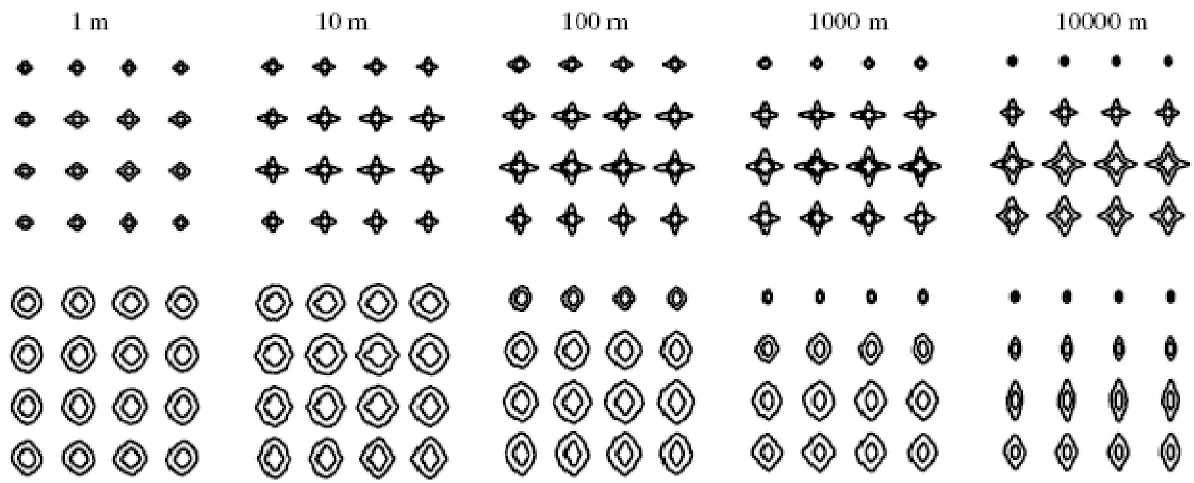
$$E[A(f)|S]$$

- **Why are Fields, Beaches and Coasts less isotropic than other natural environments?**

Slide Credit: Torralba, Olivia, J. Huang

Spatially Localized Statistics

- Windowed FFT



Top Row: Man-made environments
Bottom Row: Natural environments

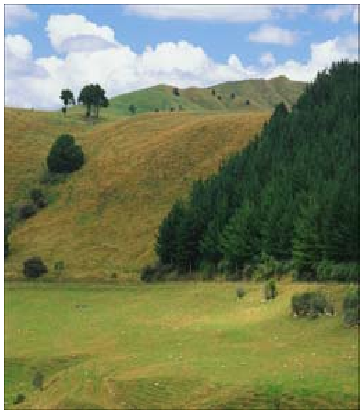
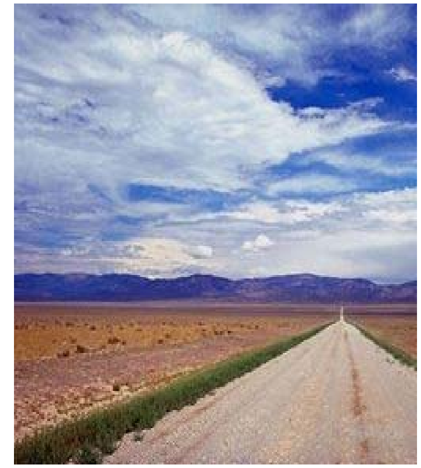
- **Image statistics become non-stationary as scene scale increases.**

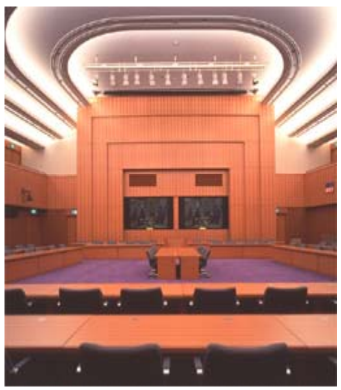
Slide Credit: Torralba, Olivia, J. Huang

The Spaial Envelope

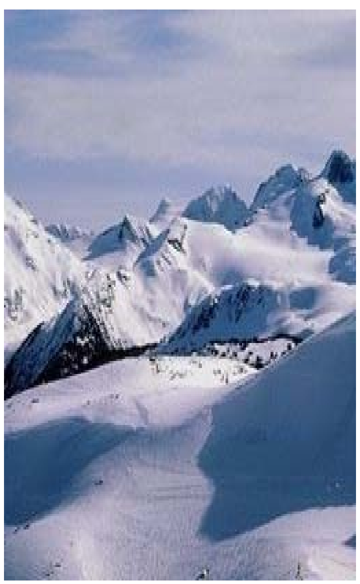
Aude Oliva

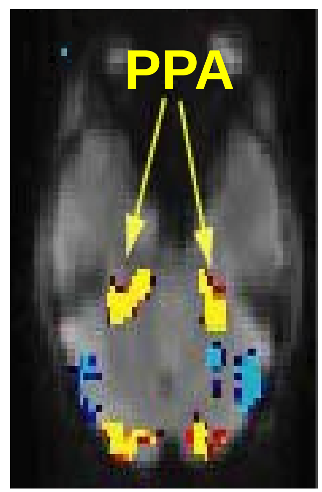
Brain & Cognitive Sciences
Massachusetts Institute of Technology
Email: oliva@mit.edu   http://cvcl.mit.edu







Spatial Envelope Theory

As a scene is inherently a 3D entity, initial scene recognition might be based on properties *diagnostic of the space* that the scene subtends and not necessarily the objects the scene contains

"Street"



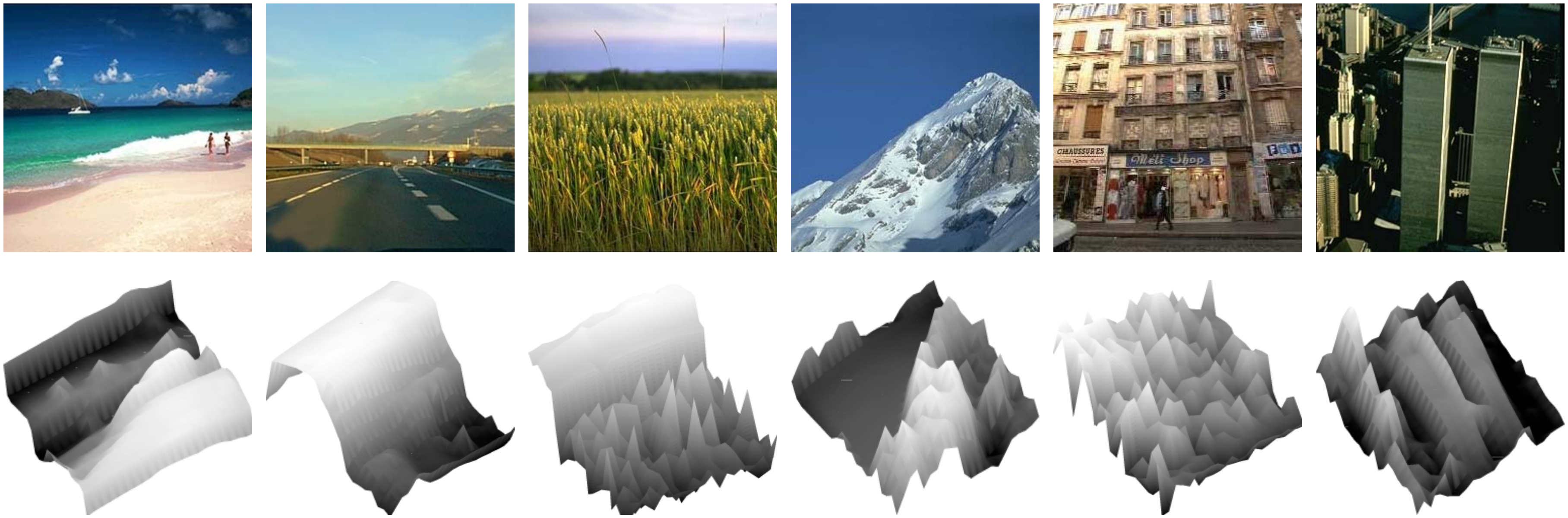

Degree of clutter, openness, perspective, roughness, etc …

Oliva et al (1999); Oliva & Torralba (2001, 2002, 2006); Torralba & Oliva (2002,2003); Greene & Oliva (2006, 2008, 2009)

Slide Credit: Olivia

Spatial Envelope Theory of Scene Representation

Oliva & Torralba (2001)



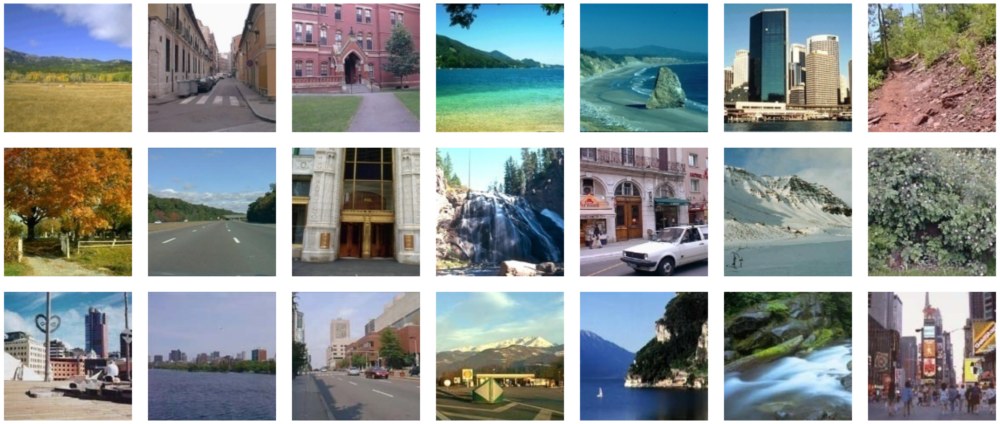
A scene is a single surface that can be represented by global (statistical) descriptors

Slide Credit: Olivia

Scene Perceptual Dimensions

Like a *texture*, a scene could be represented by a set of structural dimensions, but describing surface properties of a *space*.

<u>We use a classification task:</u> observers were given a set of scene pictures and were asked to organize them into groups of similar shape, similar global aspect, similar spatial structure.



They were explicitly told to not use a criteria related to the objects or a scene semantic group.

Slide Credit: Olivia

Scene Perceptual Dimensions

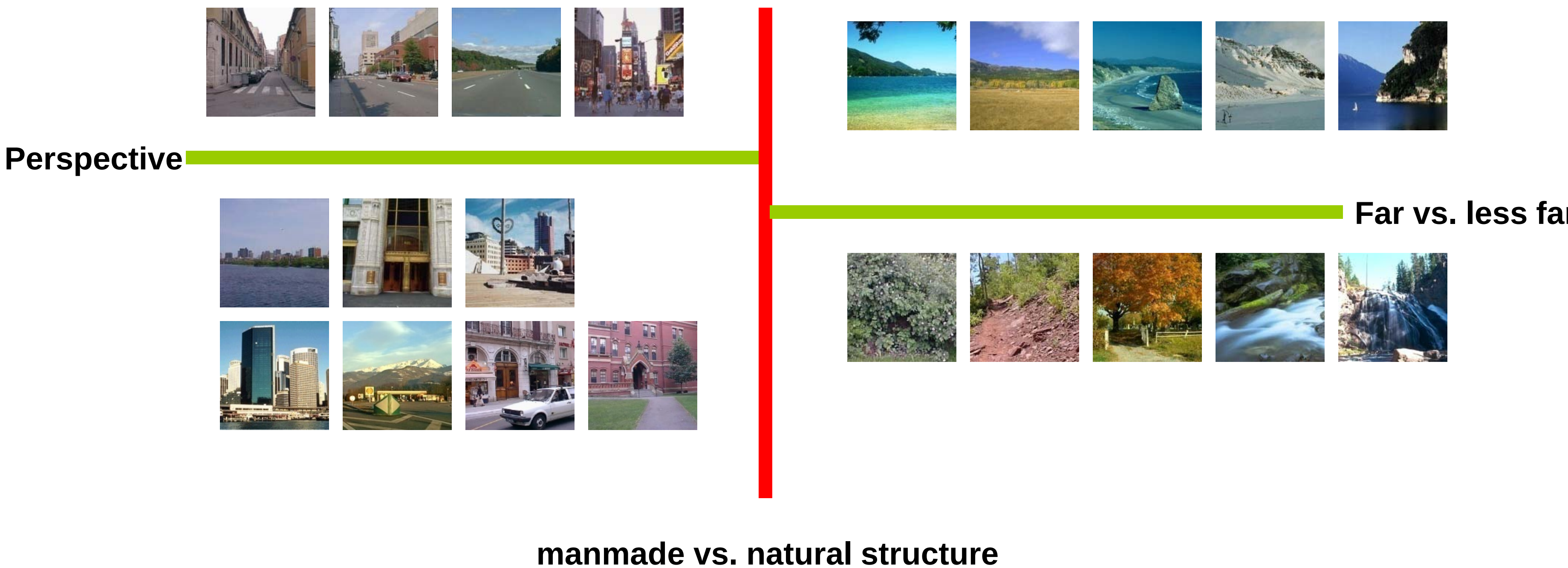
<u>Task:</u> The task consisted in 3 steps: the first step was to divide the pictures into 2 groups of similar shape.

Example: manmade vs. natural structure

Slide Credit: Olivia

Scene Perceptual Dimensions

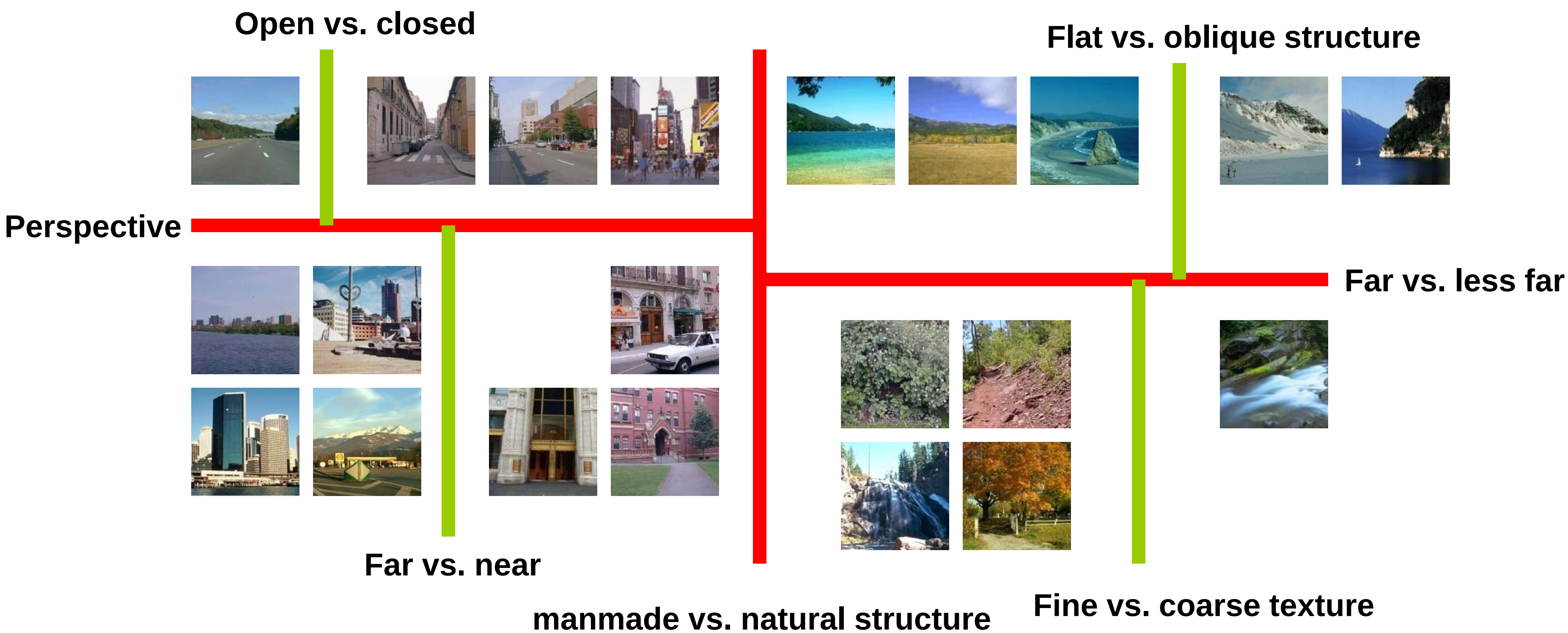
<u>Task:</u> The second step was to split each of the 2 groups in two more subdivisions.



Slide Credit: Olivia

Scene Perceptual Dimensions

<u>Task:</u> **In the third step, participants split the 4 groups in two more groups.**



Slide Credit: Olivia

Perceptual Dimensions

Dimensions	%
Naturalness	77
Openness	83
Perspective	53
Size (roughness)	47
Diagonal planes	41
Depth	59
Symmetry	29
Verticalness	18

Oliva & Torralba (01), Torralba & Oliva (02)

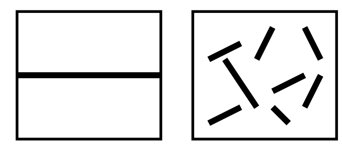
Slide Credit: Olivia

A vocabulary of global properties

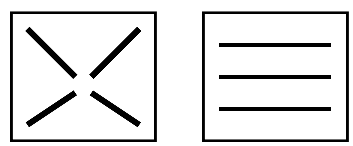
The Spatial Envelope is a combination of global properties describing the scene structure as a whole

<u>Naturalness:</u> principal structure of building blocks

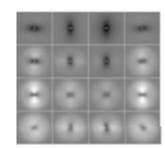
<u>Openness</u>: the sense of enclosure of the space



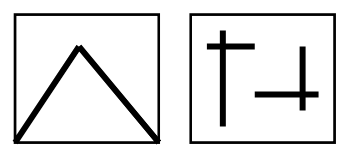
<u>Expansion</u>: the perspective



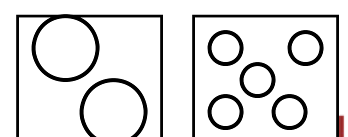
<u>Mean depth</u>: the scale of the space



<u>Ruggedness</u>: deviation of the ground plane



<u>Roughness</u>: size of the building blocks



Slide Credit: Olivia

Global Properties: Structure of space

Mean depth

Small volume **large volum**

Openness

Expansion

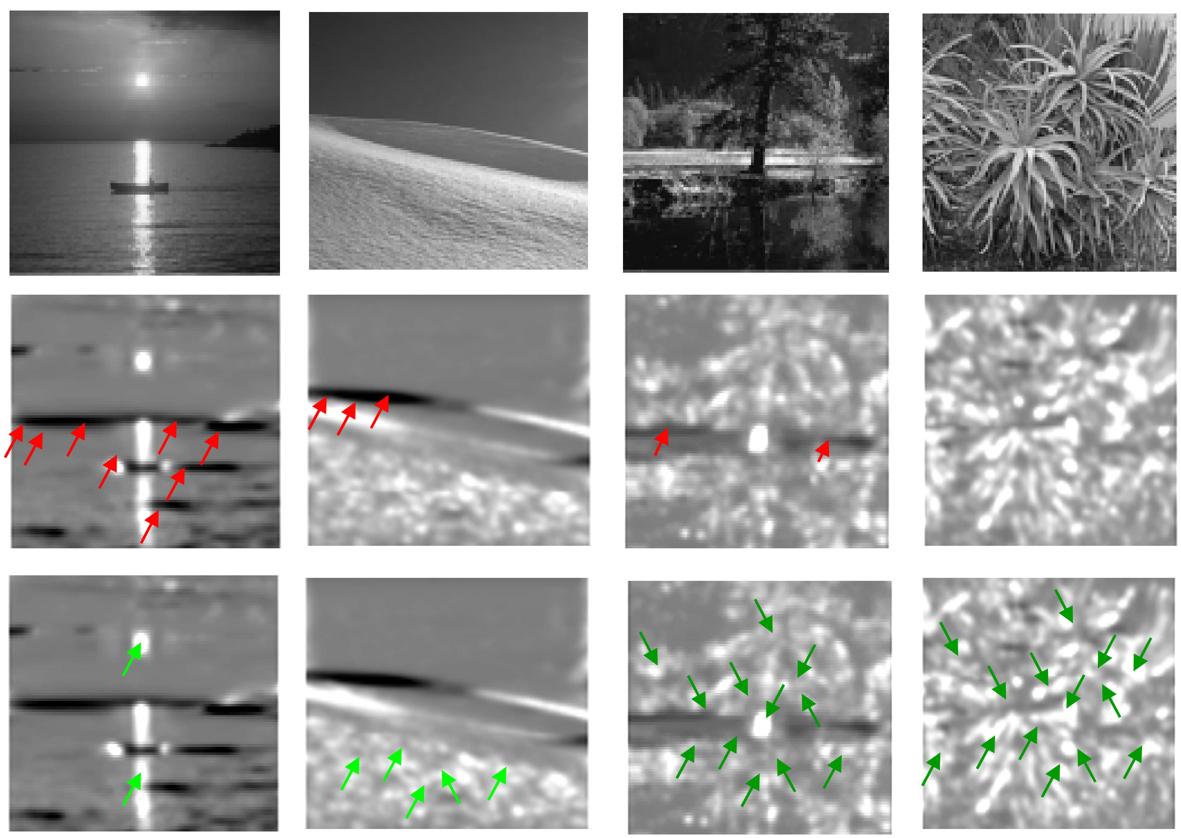
Slide Credit: Olivia

Oliva & Torralba, 2001

Diagnostic features of Openness & Closedness

Open scenes

Closed scenes



Slide Credit: Olivia

“Openness” Diagnostic Features

High degree of Openness

Lack of texture

Low frequency horizontal

High frequency isotropic texture

Low degree of openness

Slide Credit: Olivia

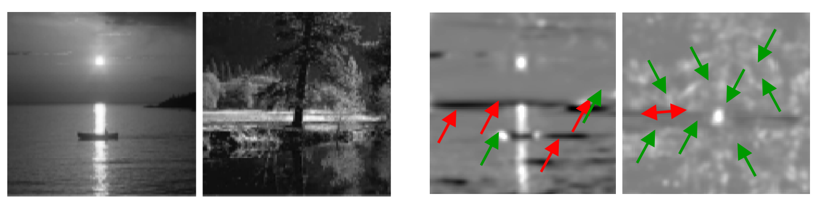
Learning Diagnostic Features

- **Any scene image has a value along each global property.**

From open scenes → closed scenes

V$_G$ V$_G$ V$_G$ V$_G$ V$_G$

- **Can we find a set of features that would represent adequately each global property ?**



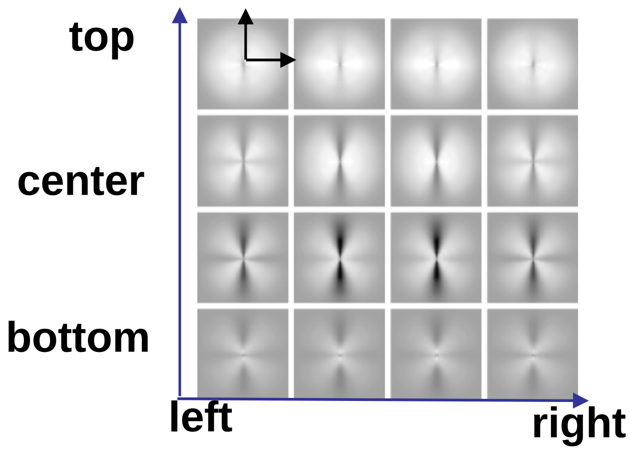
"open" "close"

Slide Credit: Olivia

Oliva & Torralba (2001)

Learning Diagnostic Features

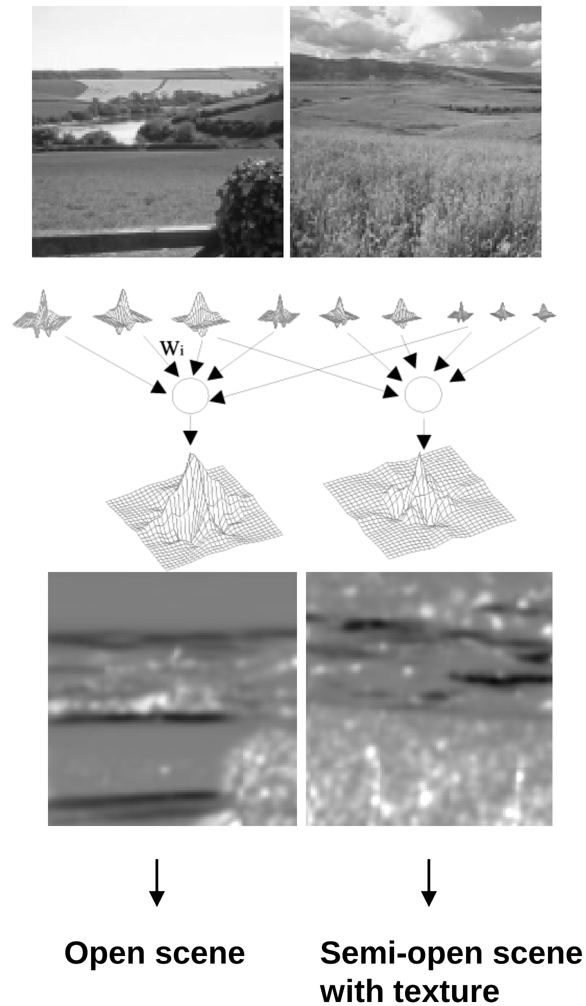
<u>Method</u>: <u>Learning stage</u>: Knowing the rank of 200-500 images along each global property, we learn the linear regression between V_G and rank.



The template (here shown in the spectral domain) is the result of the regression:

it illustrates how each spectral component contributes to a global property.

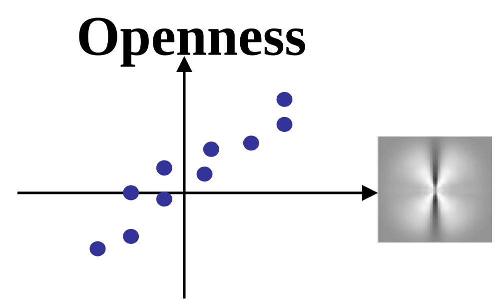
Diagnostic features of Openness



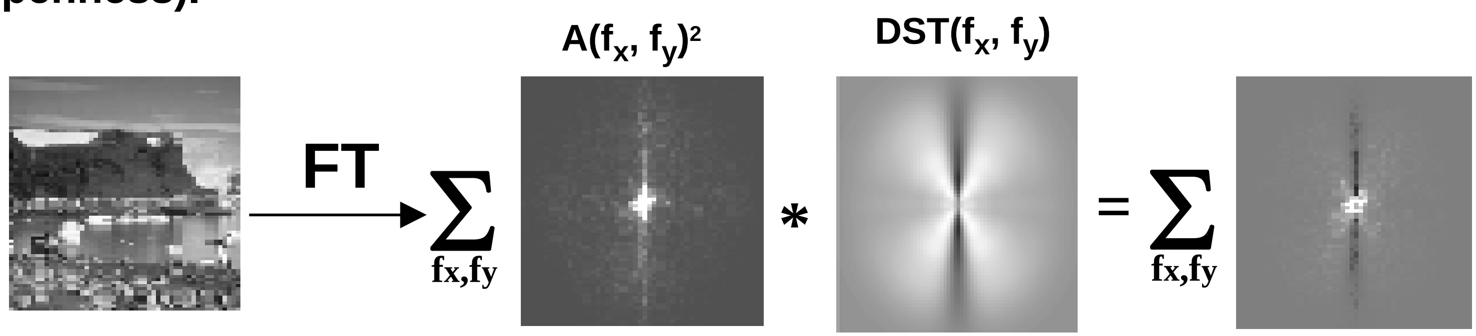
Slide Credit: Olivia

Estimation of Space descriptors

- **Method: Linear Regression Analysis: for each space property, we look for a weighting of the spectral components so that we can reproduce the same ordinal ranking as the subjects.**



- **The spatial envelope property is estimated by a a dot product between the energy spectrum and a template (Discriminant Spectral Template). The DST describes how *each spectral component contributes to a space property* (e.g. openness).**



Estimation of Space descriptor

- **Spatial envelope properties are *continuous* perceptual dimensions**

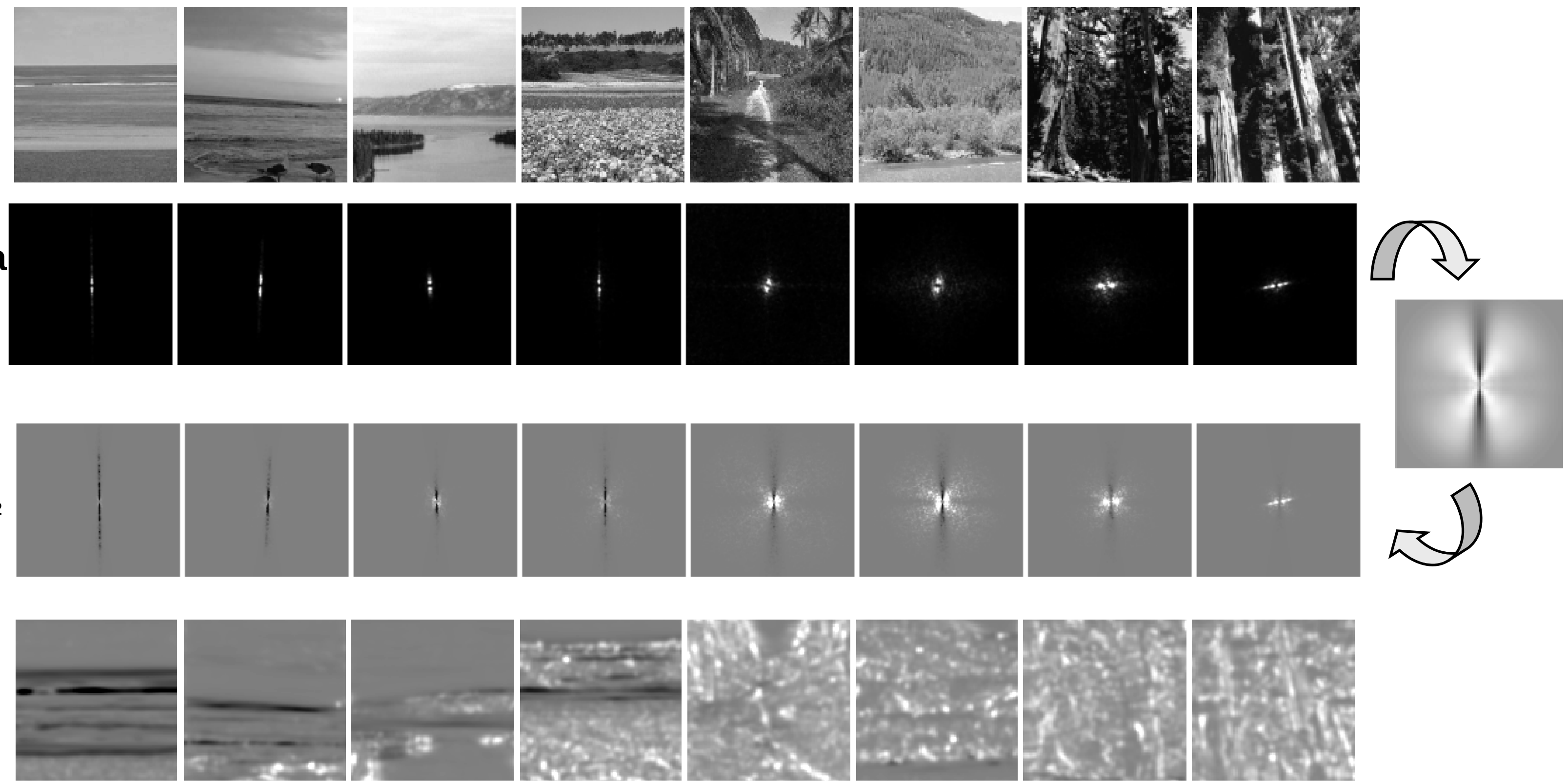
From open scenes to …. **….. closed scenes**

Original images

Power spectra $A(f_x, f_y)^2$

Product $DST(f_x, f_y)\ A(f_x, f_y)^2$

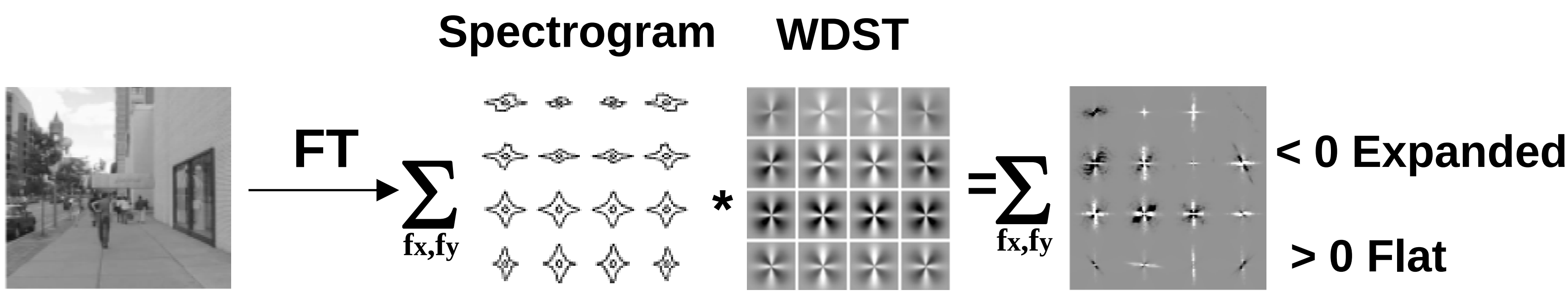
Opponent energy image



Dark: *Openness* features **Light: *closeness* feature**

Slide Credit: Olivia

Windowed Discriminant Template

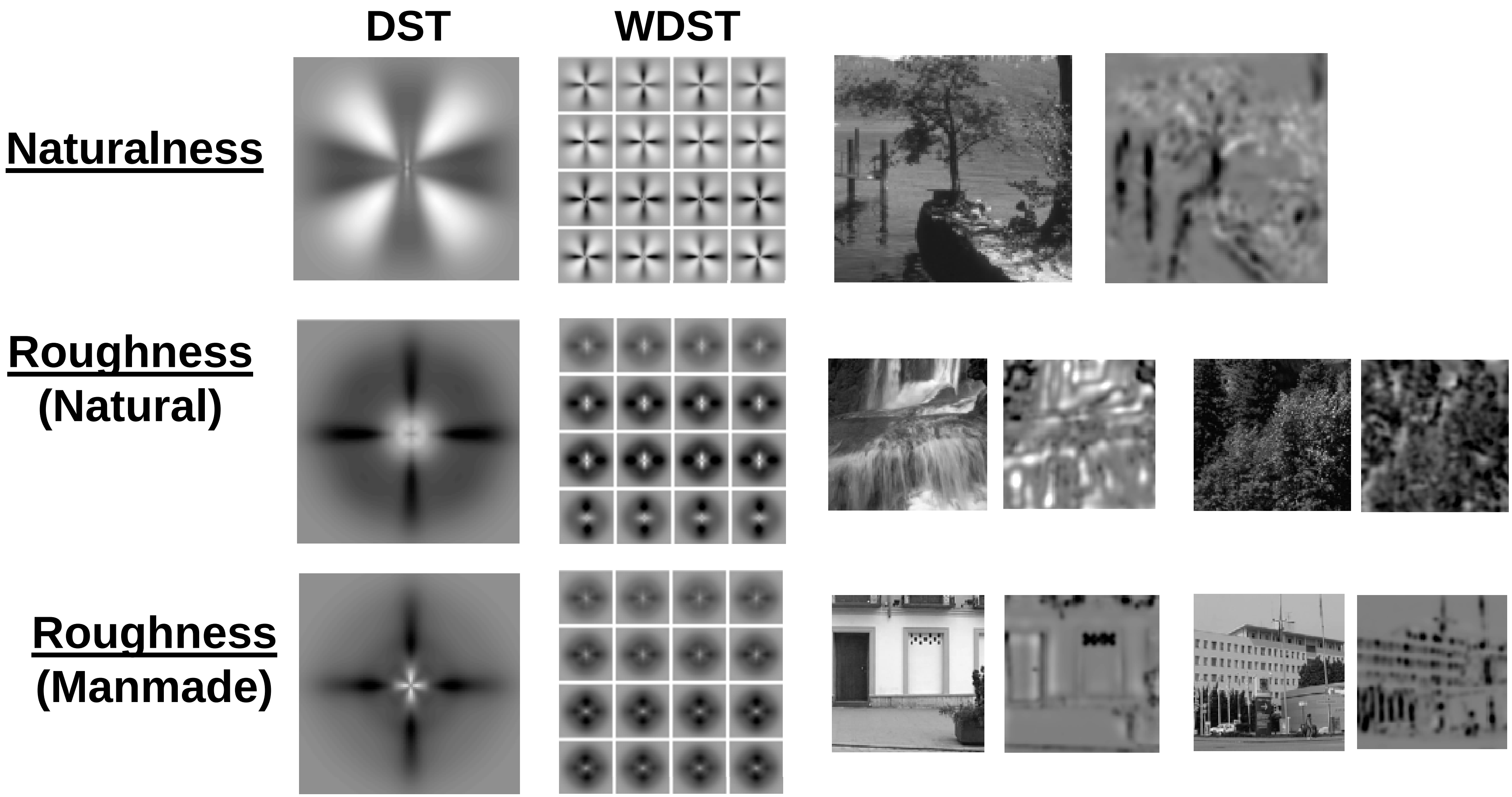


Expansion

Flat

Slide Credit: Olivia

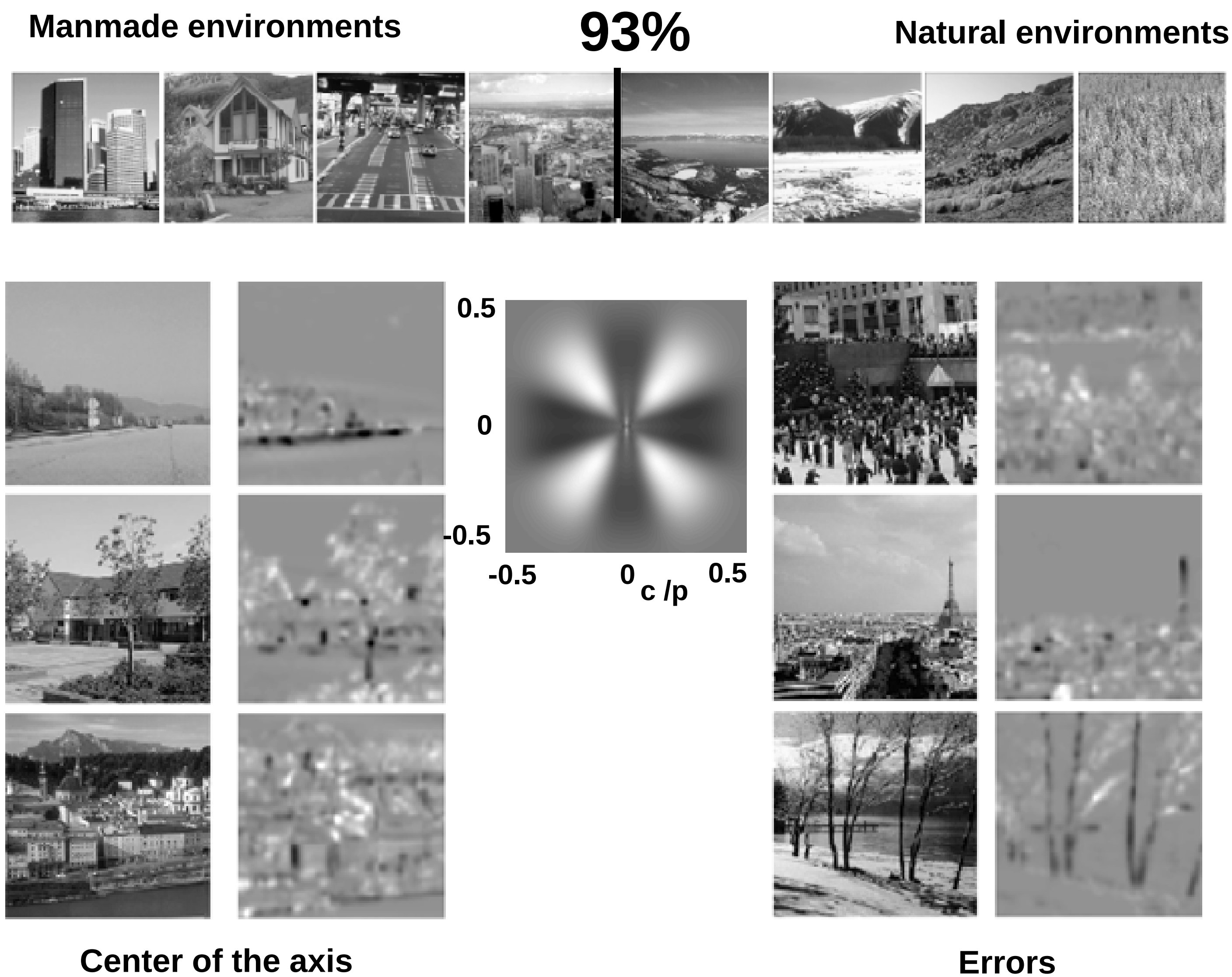
Space Properties of the *Content* of the scene



Stationary distribution of features

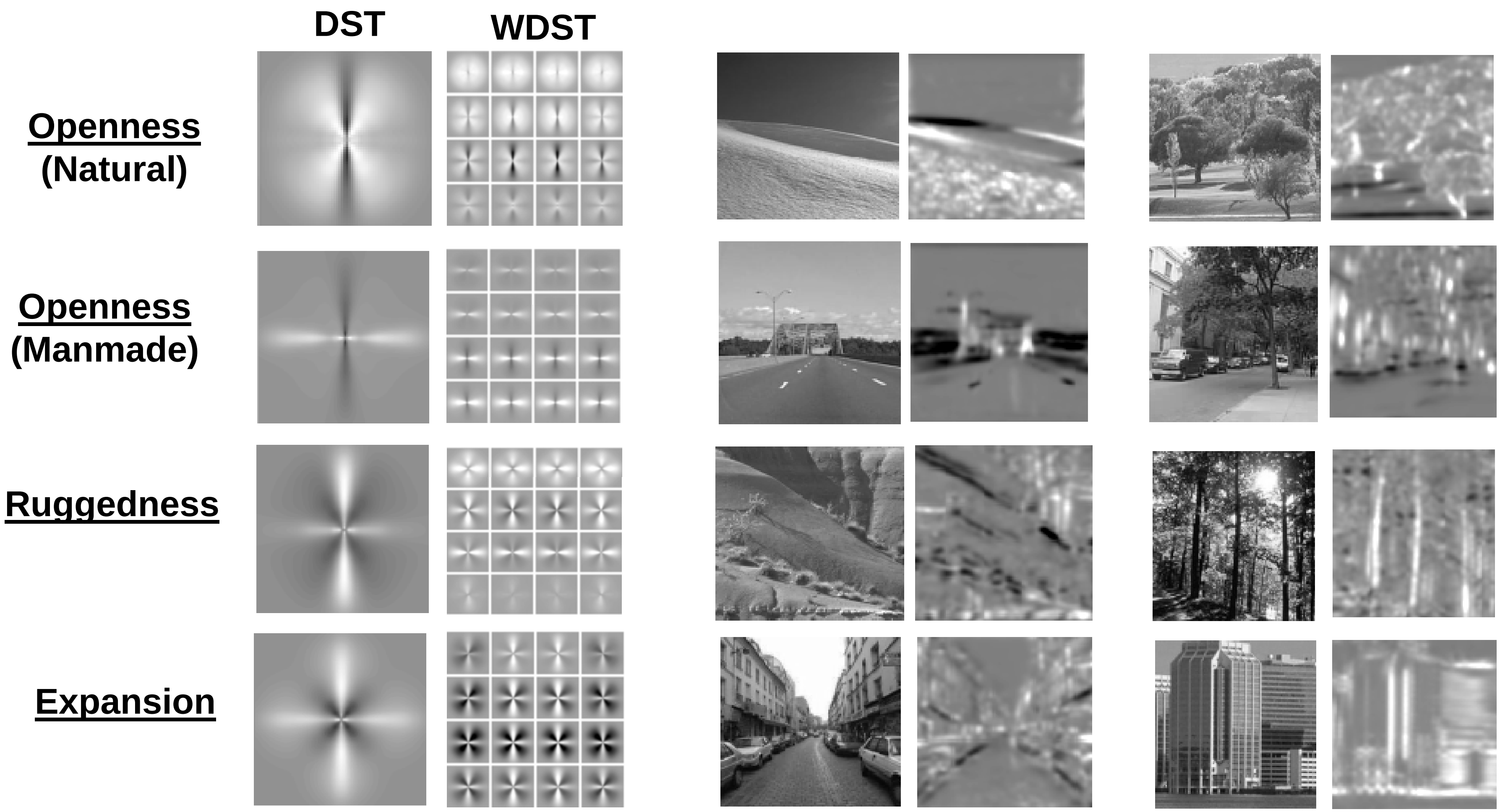
Slide Credit: Olivia

Naturalness descriptor



Slide Credit: Olivia

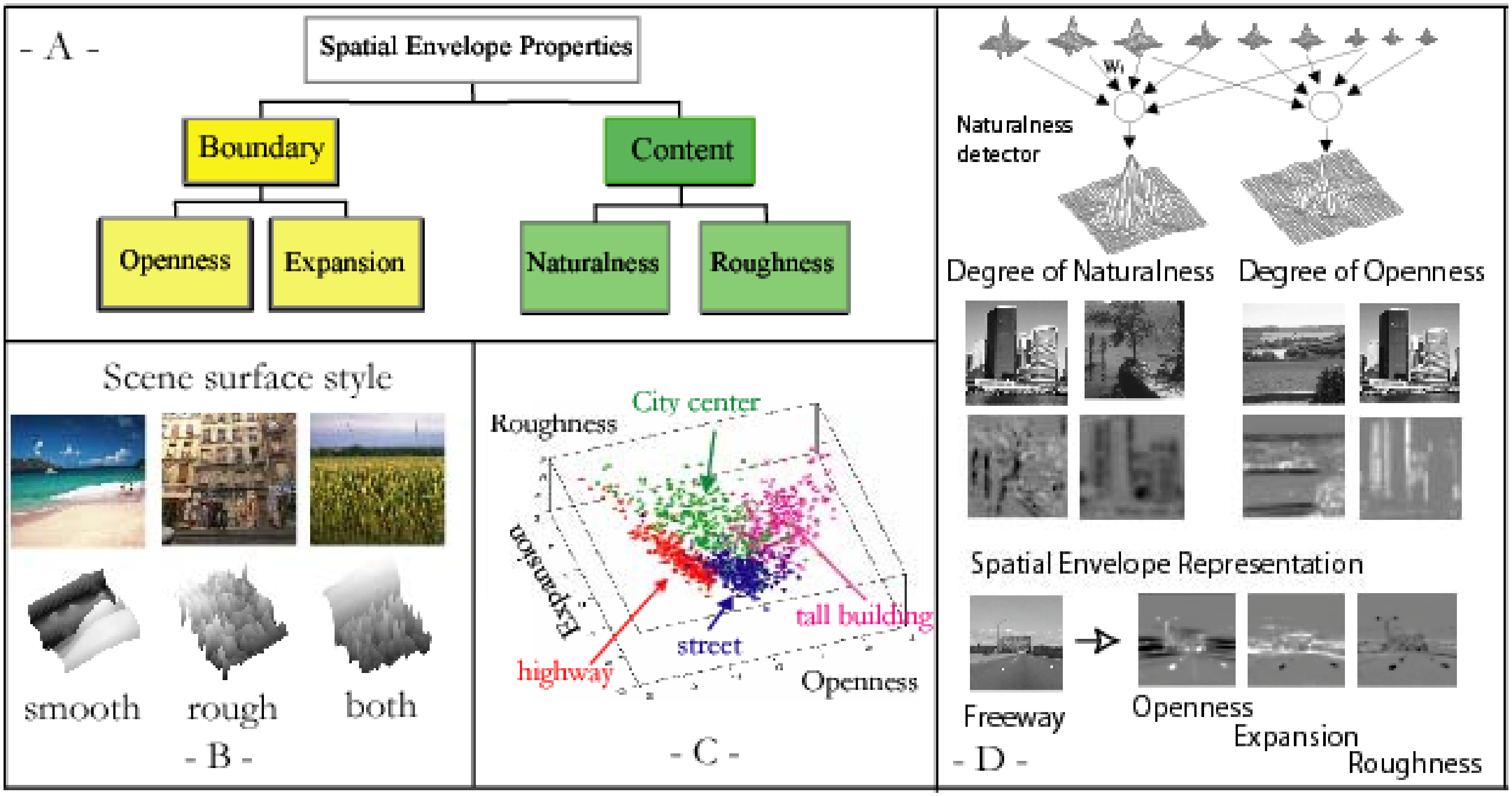
Space Properties of the *Shape* of the scene



Non-stationary distribution of features

Slide Credit: Olivia

Spatial Envelope Theory of Scene Recognition



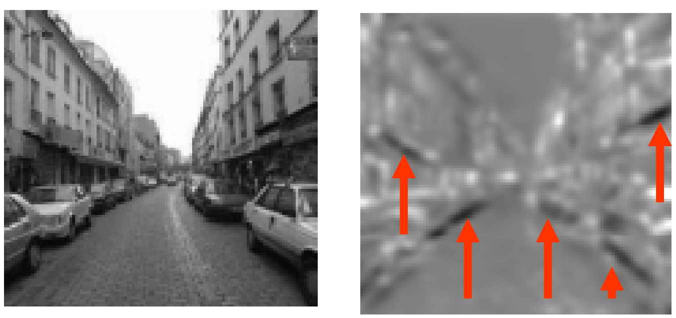
Oliva & Torralba (2001). International Journal of Computer Vision.

Slide Credit: Olivia

Global Scene Structure

- **<u>Hypothesis</u>**

  Scenes of the same category membership share similar global spatial layout properties

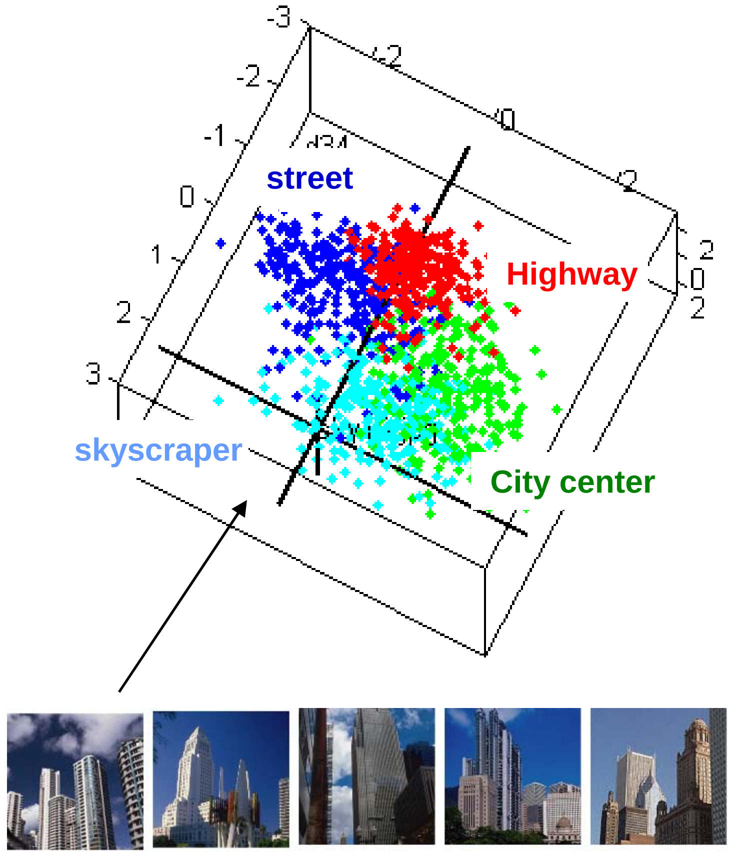


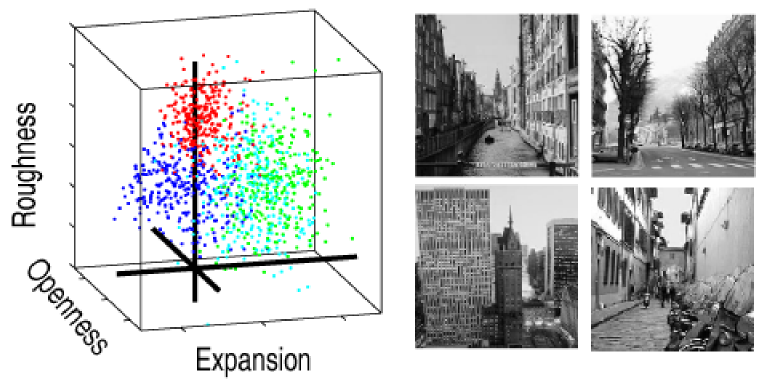
- **<u>Hypothesis</u>**

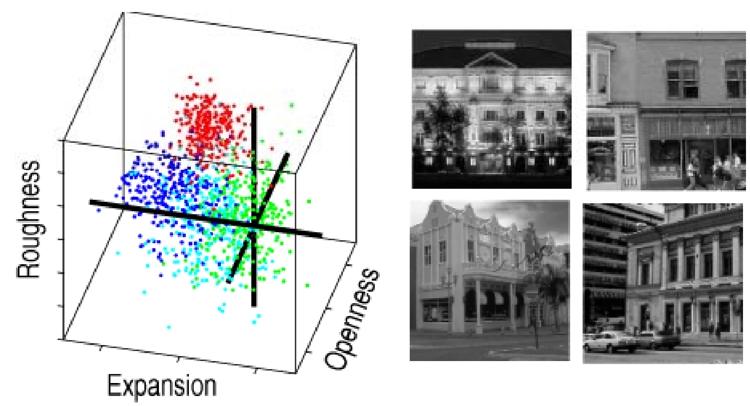
  Low level features are correlated with spatial properties (e.g. perspective)

Modeling Scene Gist

Scenes from the same category share similar global properties







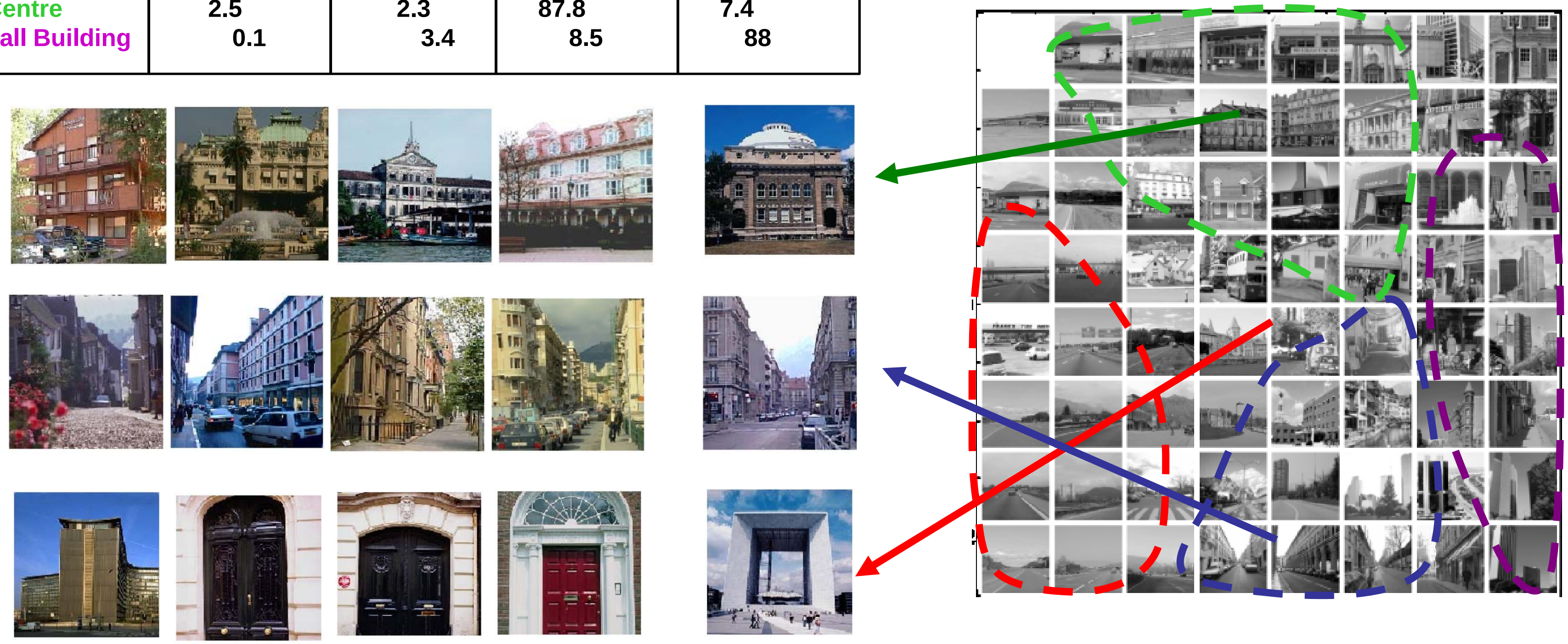
Slide Credit: Olivia

Categorization of Manmade Scenes

Confusion Matrix (in % using Layout template) :
Classification of prototypical scenes (400 / category)

	Highway	Street	City centre	tall building
Highway	91.6	4.8	2.7	0.9
Street	4.7	89.6	1.8	3.4
Centre	2.5	2.3	87.8	7.4
Tall Building	0.1	3.4	8.5	88

Local organization:
correct for 86 % images
(4 similar images on 7 K-NN)



Slide Credit: Olivia

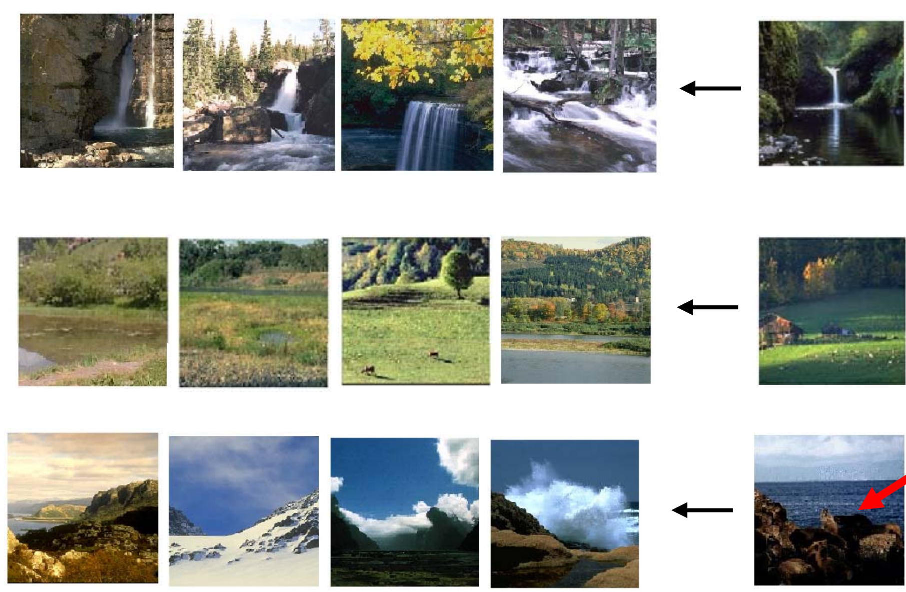
Categorization of Natural Scenes

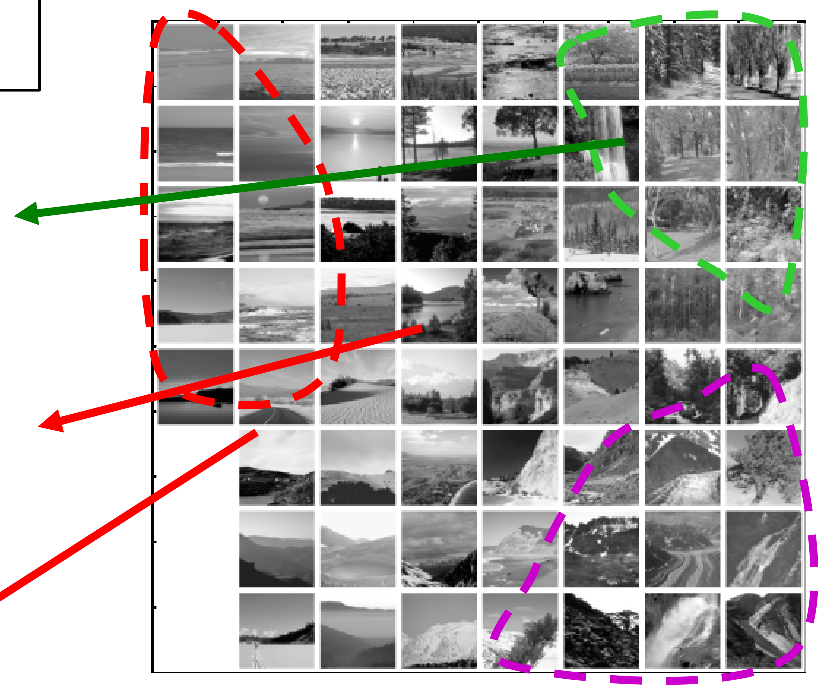
Confusion Matrix (in % using Layout template) :
Classification of prototypical scenes (400 / category)

	Coast	Countryside	Forest	Mountain
Coast	88.6	8.9	1.2	1.3
Countryside	9.8	85.2	3.7	1.3
Forest	0.4	3.6	91.5	4.5
Mountain	0.4	4.6	3.8	91.2

Local organization: correct for 92 % images

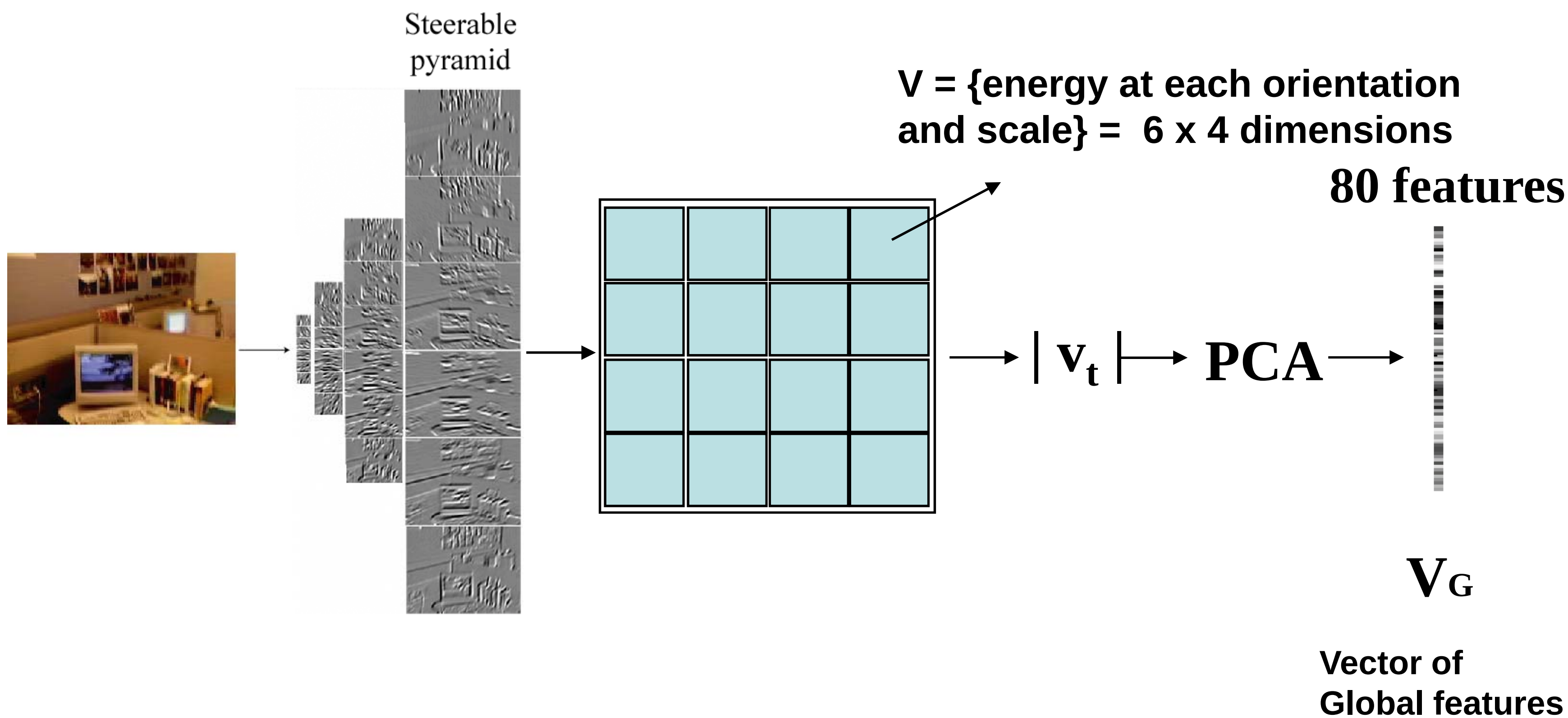
(4 similar images on 7 K-NN)





Slide Credit: Olivia

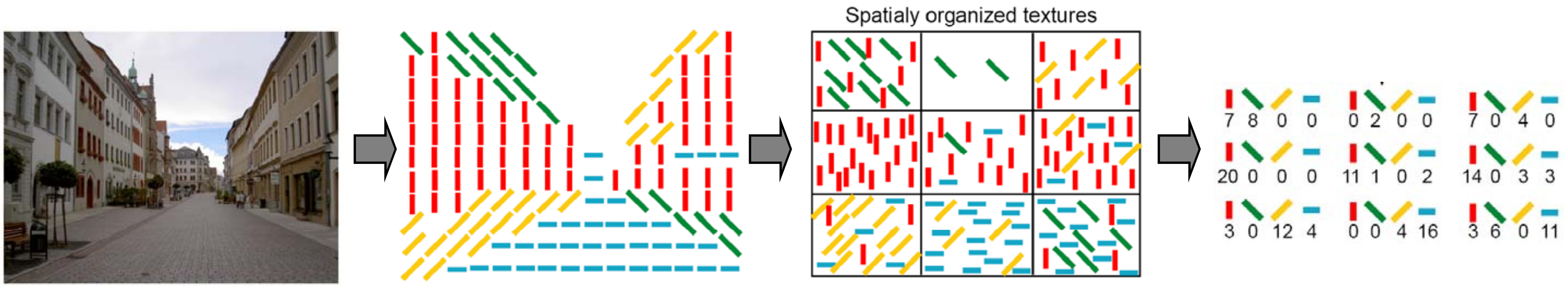
Representing Image Structure



Oliva & Torralba (2001,2002, 2006)

Slide Credit: Olivia

Scene Recognition via *texture surface*



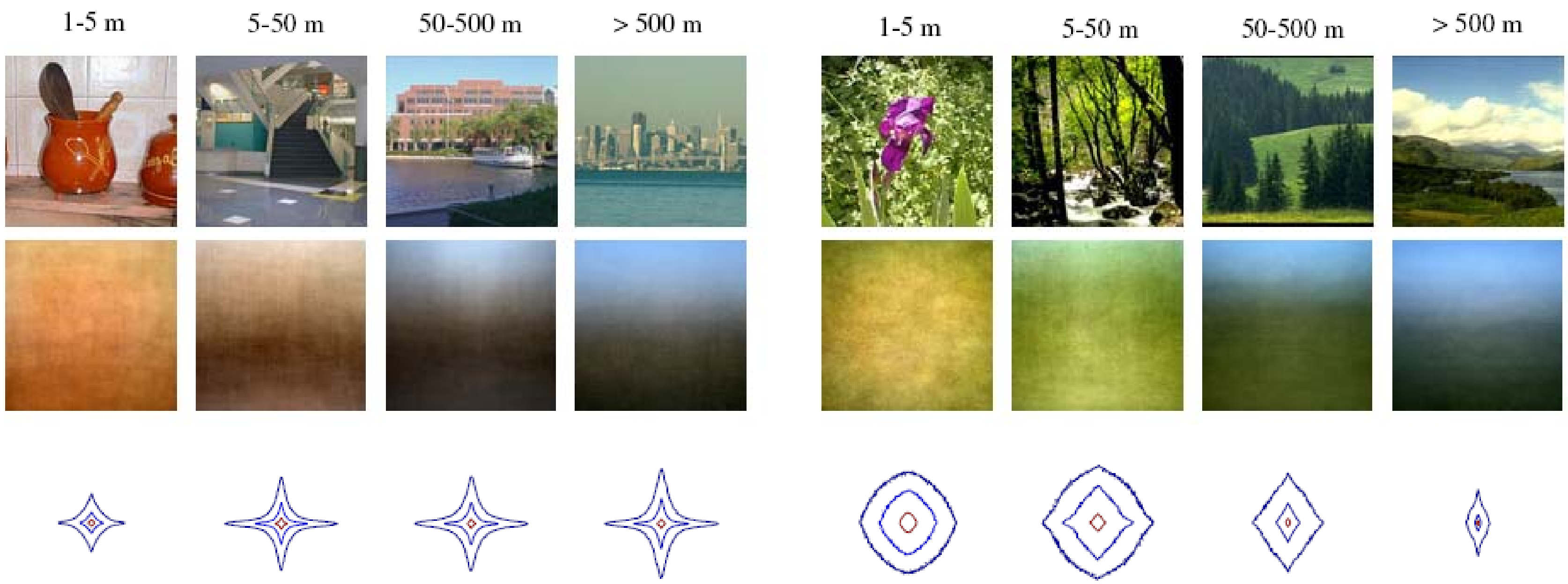
Slide Credit: Olivia

Scene Classification from “Texture”

Oliva & Torralba (2001,2006)

Slide Credit: Olivia

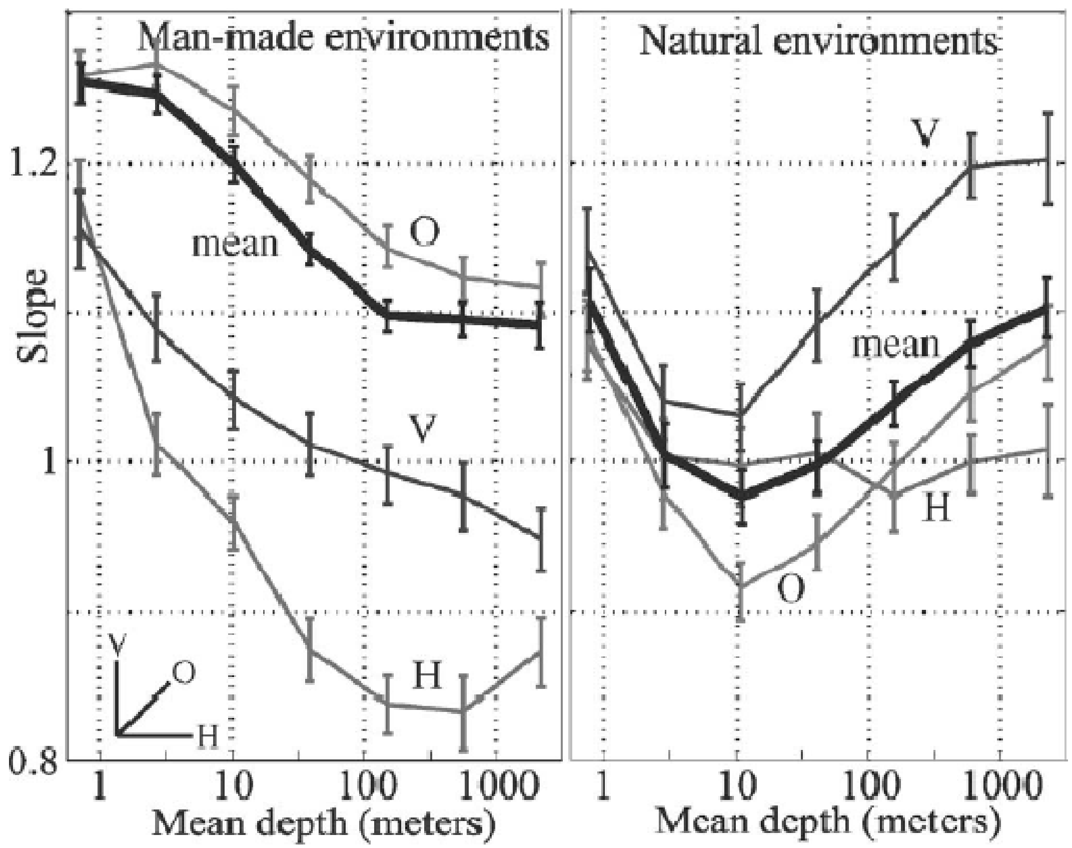
Scene Scale



- “The point of view that any given observer adopts on a specific scene is constrained by the volume of the scene.”
- How does the amount of clutter vary against scene scale in man-made environments? In natural environments?

Slide Credit: Torralba, Olivia, J. Huang

What do Images Statistics say about Depth?



V: Vertical
H: Horizontal
O: Oblique

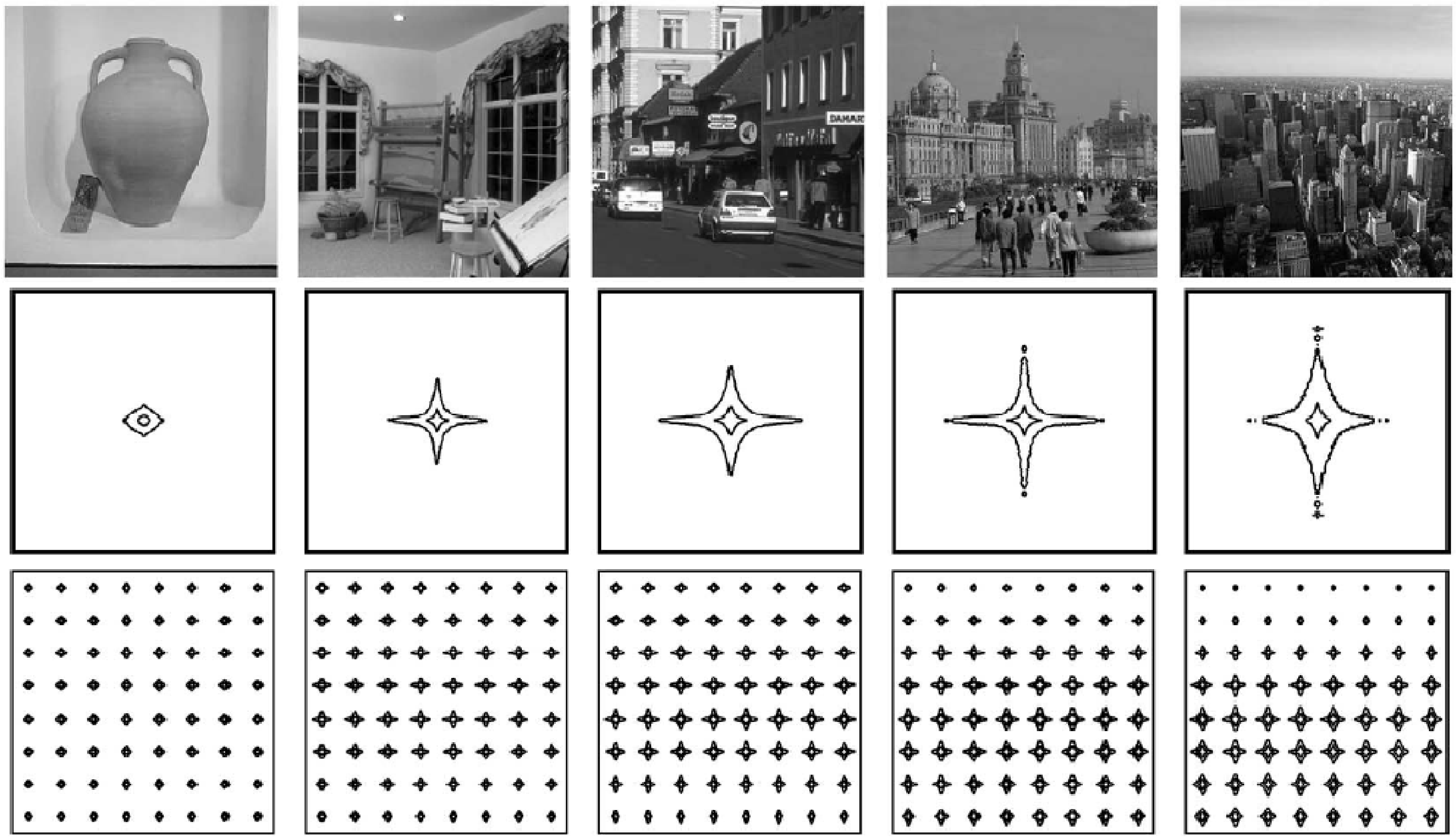
Slide Credit: Torralba, Olivia, J. Huang

Comparing Localized Spectral Signatures and Depth

- With increasing depth comes:
  - An increase in global roughness for man-made structures
  - A decrease in global roughness for natural structures
  - Nonuniformity in spatially localized spectral signatures

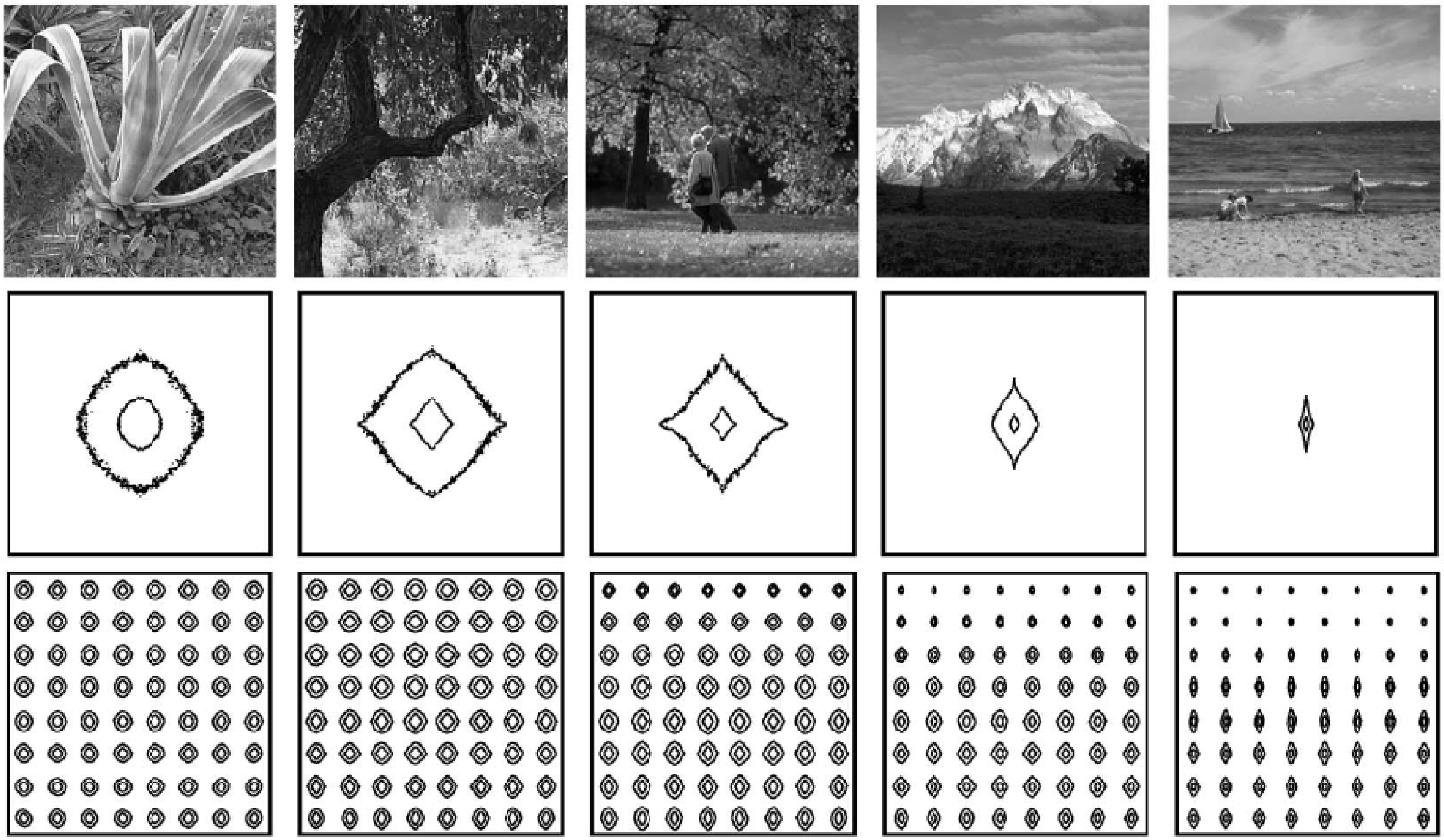
Slide Credit: Torralba, Olivia, J. Huang

Examples (man-made)



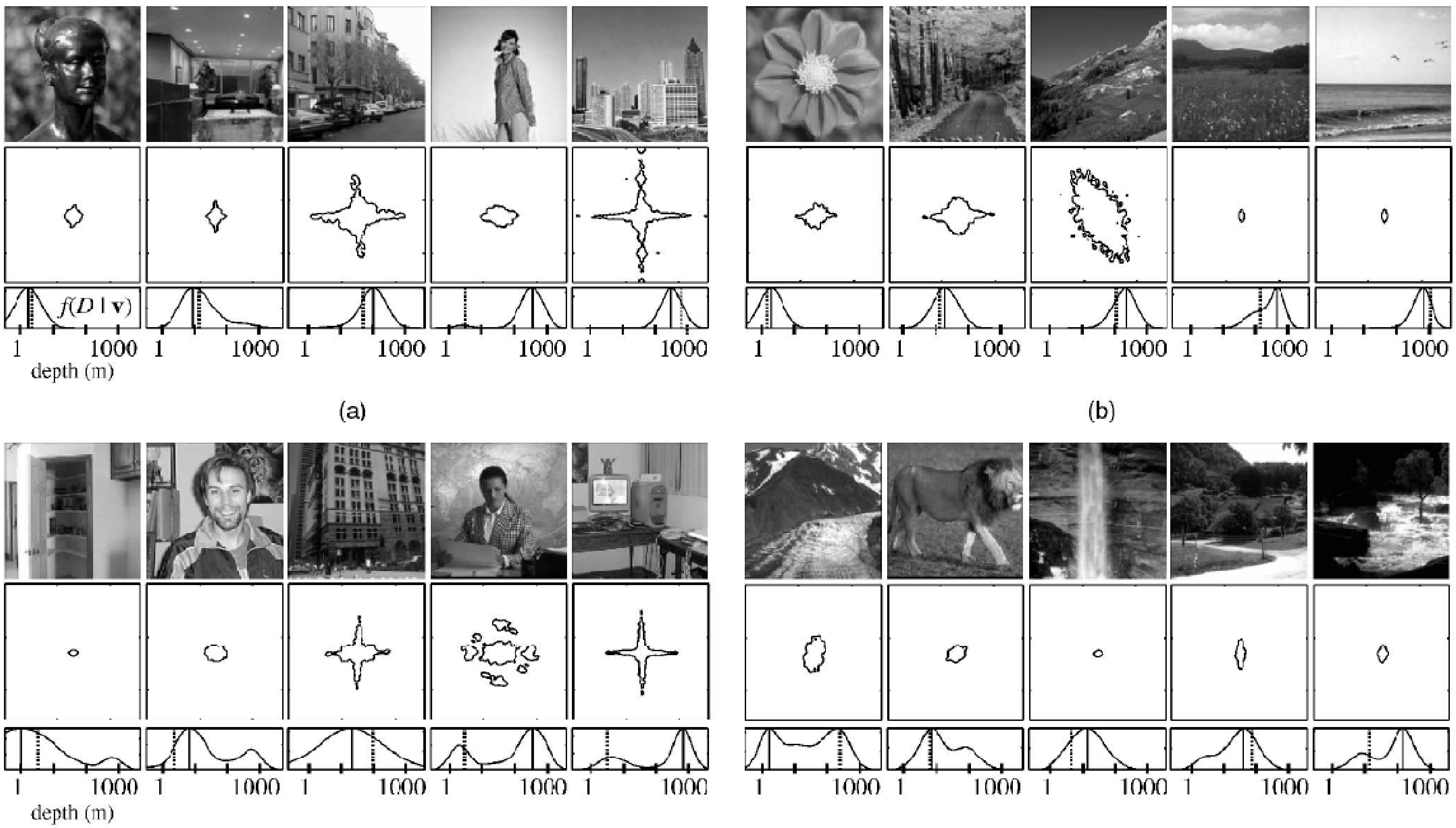
Slide Credit: Torralba, Olivia, J. Huang

Examples (Natural)

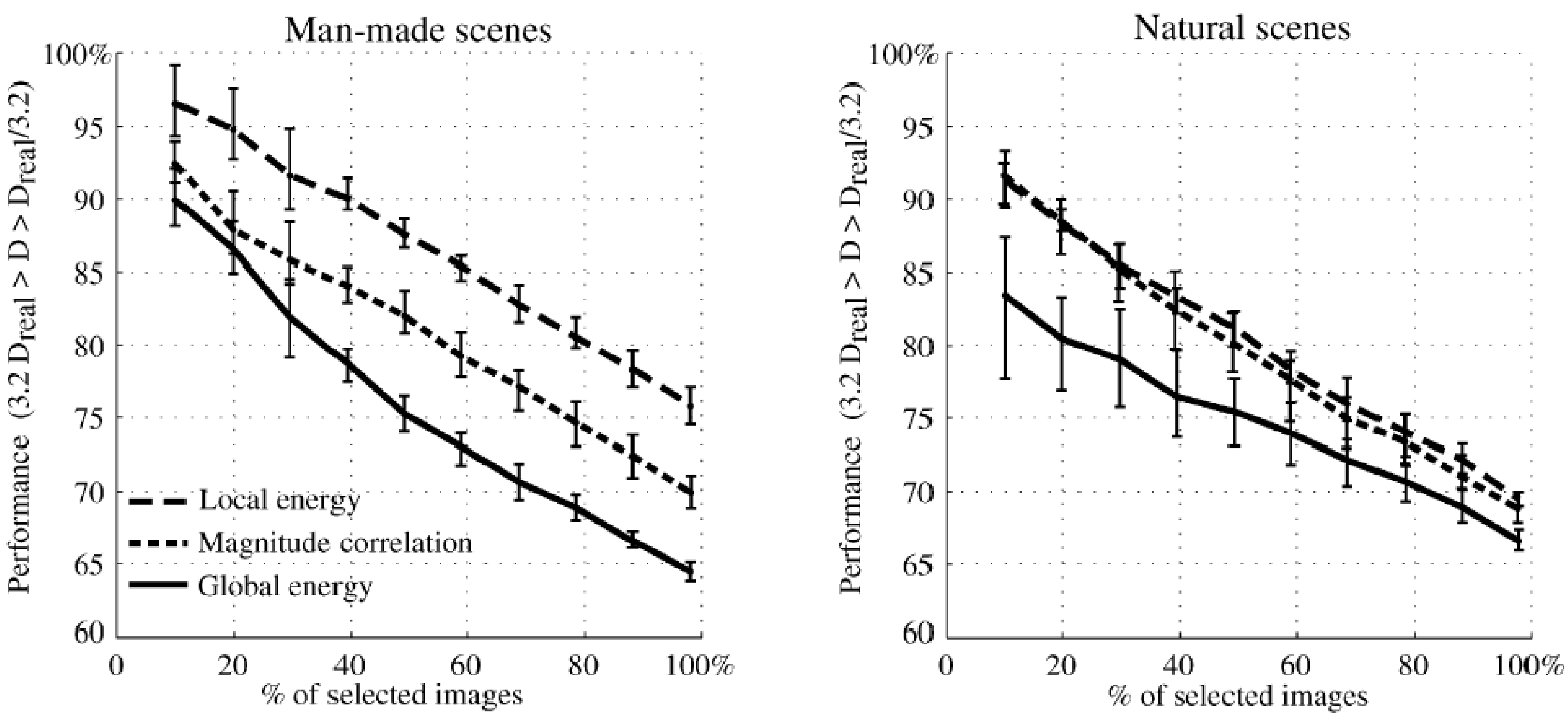


Slide Credit: Torralba, Olivia, J. Huang

Some Results

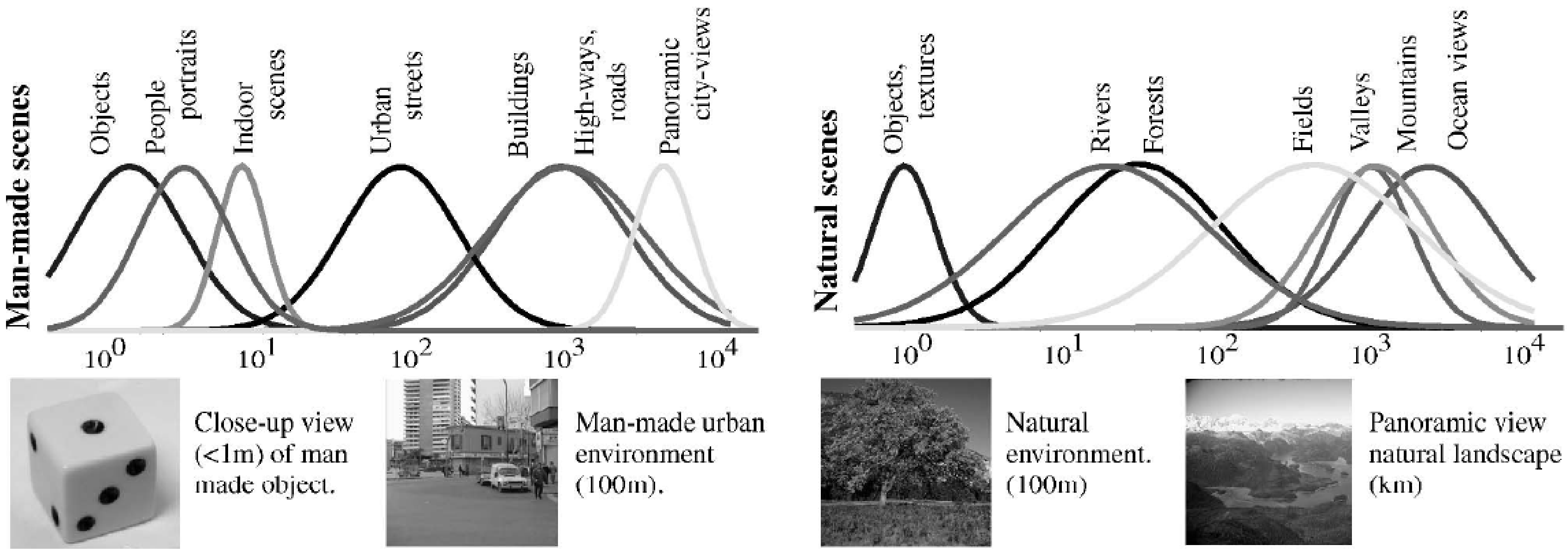


Slide Credit: Torralba, Olivia, J. Huang



Slide Credit: Torralba, Olivia, J. Huang

$f(D \mid category)$



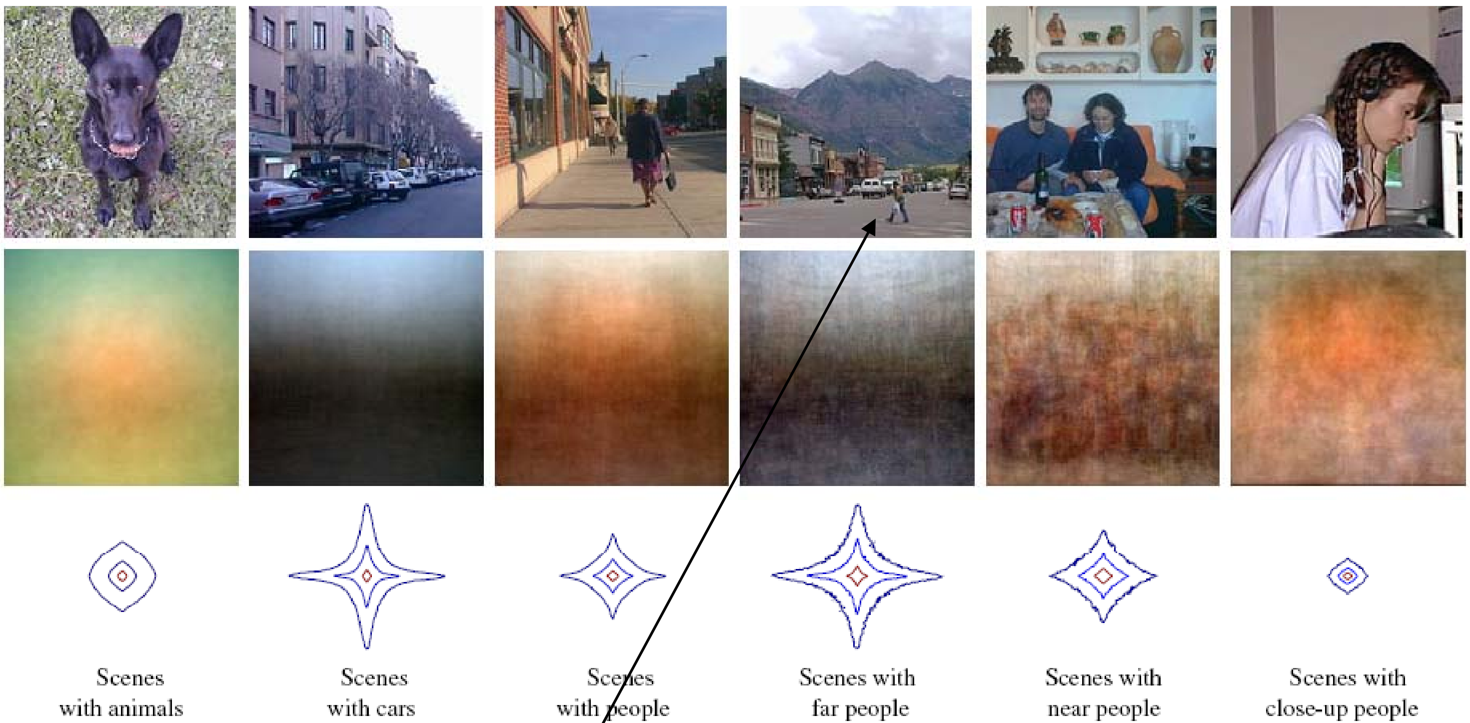
Distribution of Scene Categories as a function of mean depth.

Slide Credit: Torralba, Olivia, J. Huang

Application: Scale Selection

Slide Credit: Torralba, Olivia, J. Huang

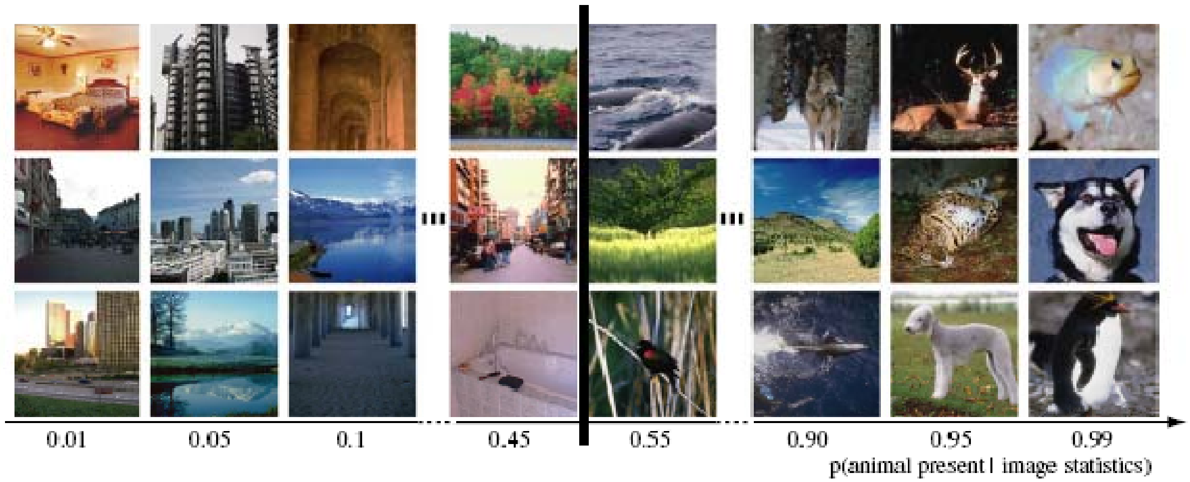
Context in Images

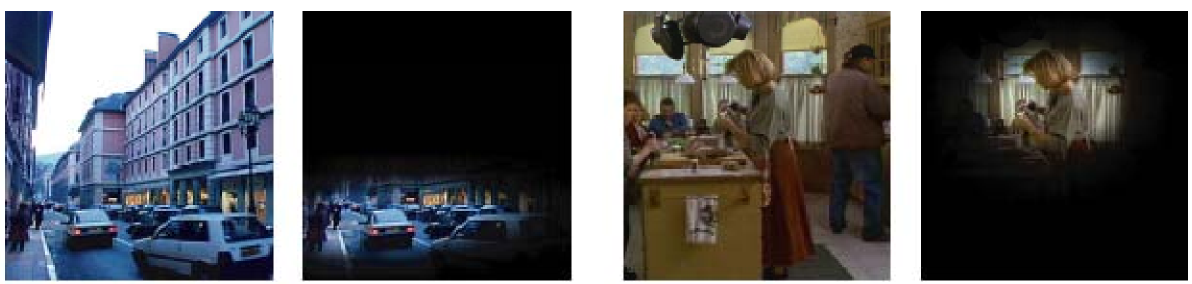


- **Question: How can these small people possibly affect the image statistics in any significant way??**

Slide Credit: Torralba, Olivia, J. Huang

Object Detection





Slide Credit: Torralba, Olivia, J. Huang

References

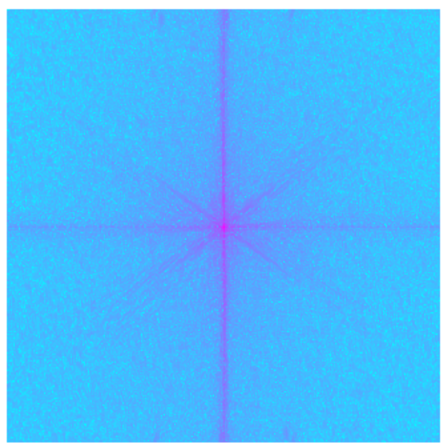
- Torralba and Oliva, *Statistics of Natural Image Categories*. Network: Computation in Neural Systems 14 (2003) 391-412.
- Torralba and Oliva, *Depth Estimation from Image Structure*. IEEE PAMI Vol 14, No. 9 (2002).
- Oliva and Torralba, *Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope*. IJCV 42(3), 145-175 (2001).

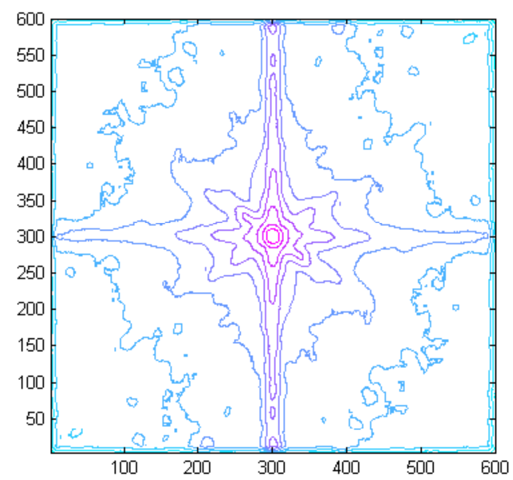
- Srivastava, Lee, Simoncelli, Zhu, *On Advances in Statistical Modeling of Natural Images*. JMIV 18:17-33 (2003)
- Mumford, *Pattern Theory: the Mathematics of Perception*. ICM 2002. Vol III. 1-3

Slide Credit: Torralba, Olivia, J. Huang

"Demo"

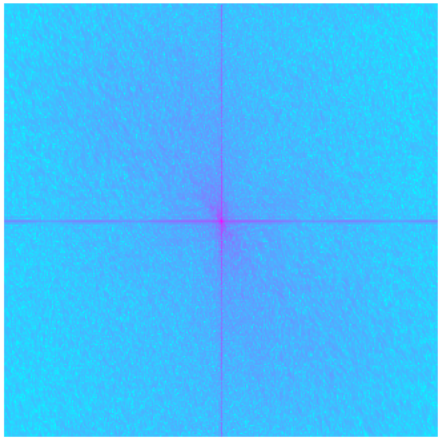
- Computing the Spectrum (Matlab):
  - `Ifft = abs(fftshift(fft2(I,w,h)));`
- Visualization:
  - `imshow(log(Ifft)/max(max(log(Ifft))));`
  - `colormap(cool);`

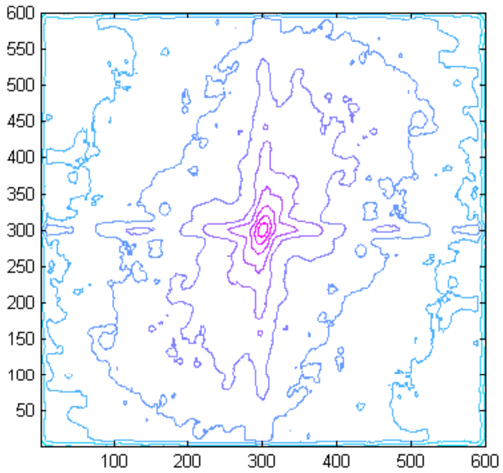




Slide Credit: Torralba, Olivia, J. Huang

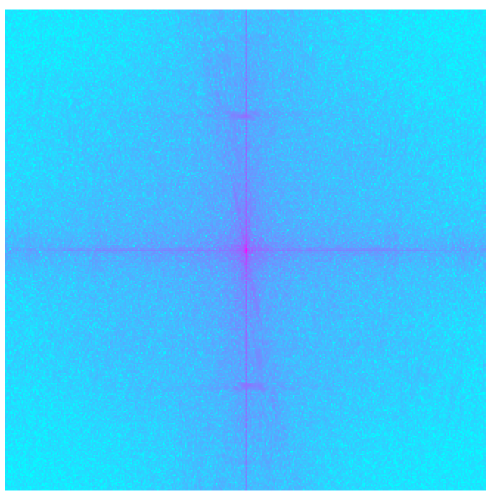
FFT(Beach)

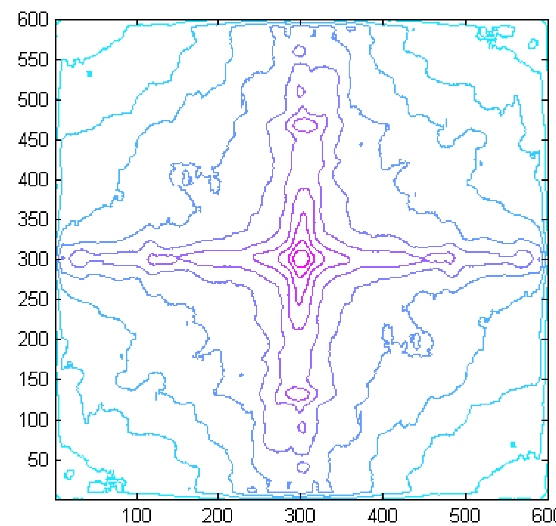




Slide Credit: Torralba, Olivia, J. Huang

FFT(Pittsburgh)





Slide Credit: Torralba, Olivia, J. Huang

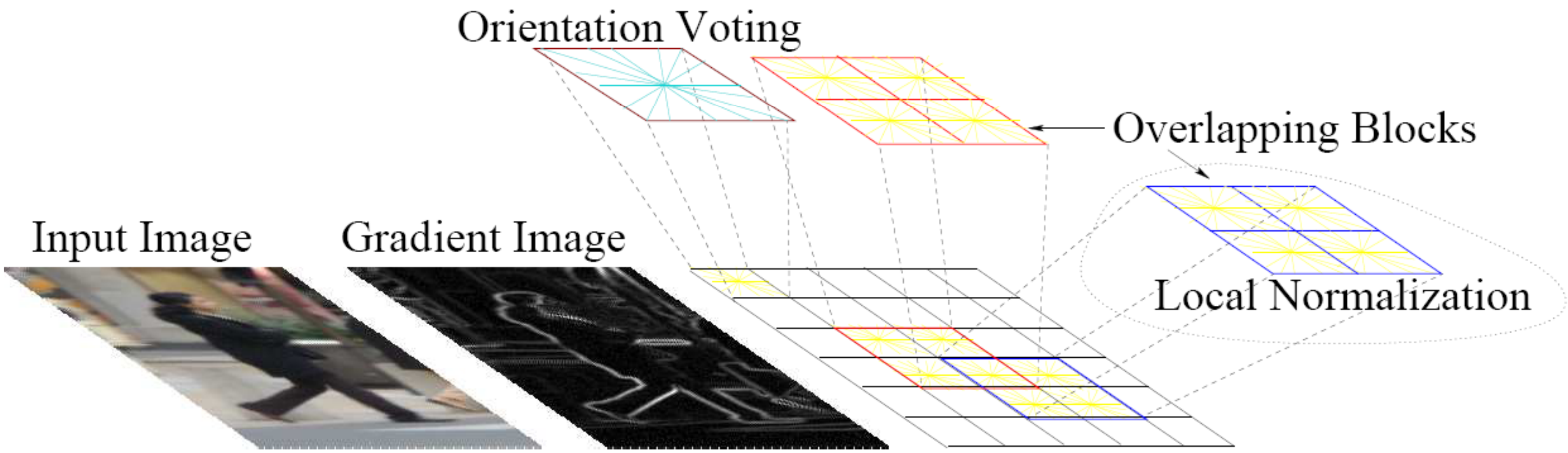
Today

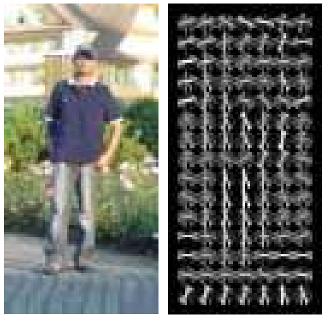
- Background / Overview

- Histograms of edges (Schiele)
- Windowed spectral analysis (GIST)
- **Tiled histograms of edges (HOG)**

- Motion History Images (Bobick)
- Rectified Flow Descriptors (Efros)
- Differential Geometry Signatures (Shah)

Gradient-based representations: Histograms of oriented gradients (HoG)





Map each grid cell in the input window to a histogram counting the gradients per orientation.

Code available:
http://pascal.inrialpes.fr/soft/olt/

Dalal & Triggs, CVPR 2005

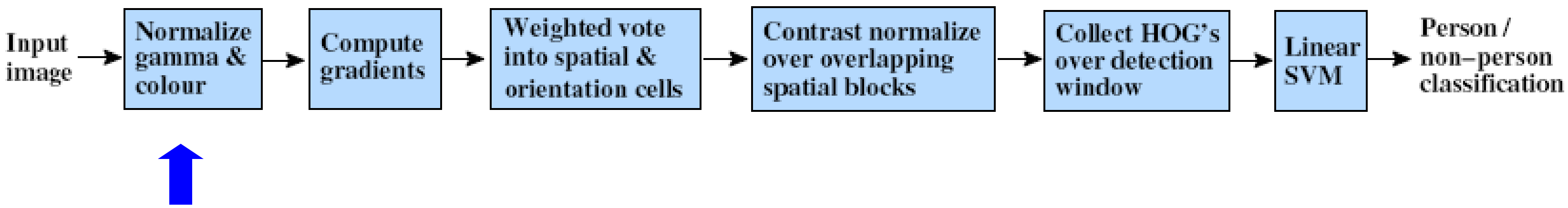
Slide credit: K. Grauman, B. Leibe



Slide credit: Dalal, Triggs, P. Barnum

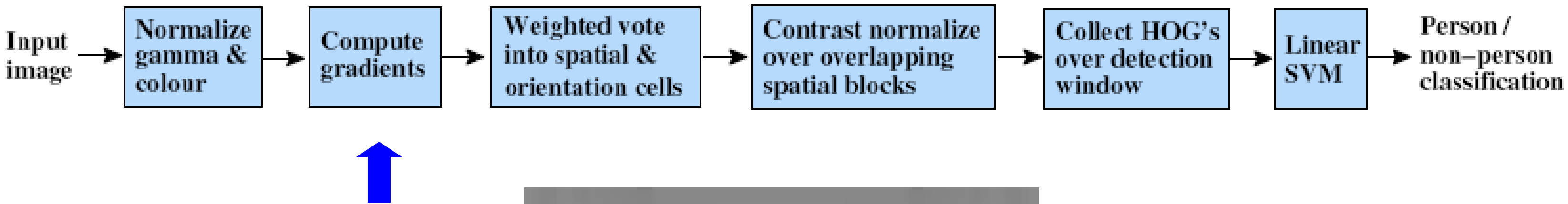
Input image → Normalize gamma & colour → Compute gradients → Weighted vote into spatial & orientation cells → Contrast normalize over overlapping spatial blocks → Collect HOG's over detection window → Linear SVM → Person / non–person classification

Slide credit: Dalal, Triggs, P. Barnum



- Tested with
  - RGB
  - LAB
  - Grayscale
- Gamma Normalization and Compression
  - Square root
  - Log

Slide credit: Dalal, Triggs, P. Barnum



-1	0	1

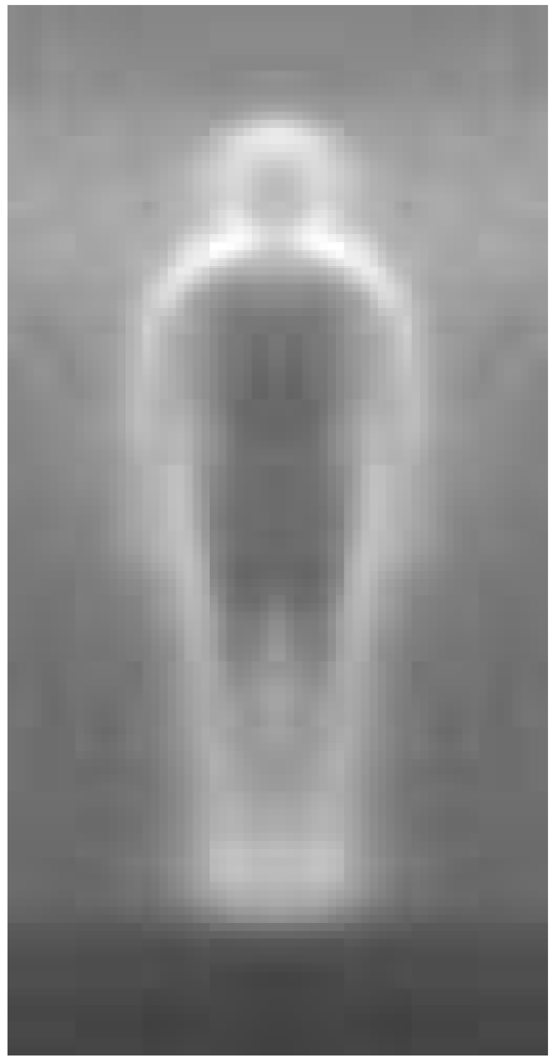
centered

-1	1

uncentered

1	-8	0	8	-1

cubic-corrected



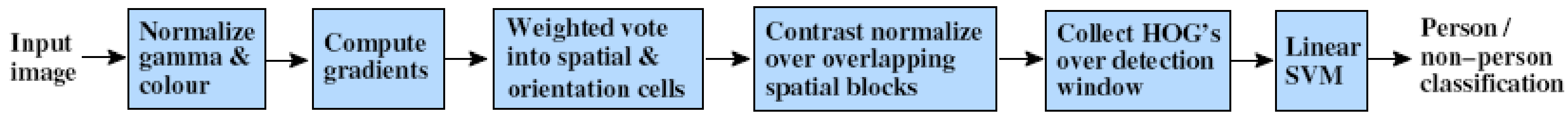
0	1
-1	0

diagonal

-1	0	1
-2	0	2
-1	0	1

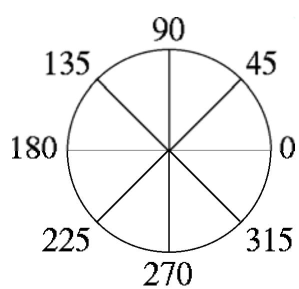
Sobel

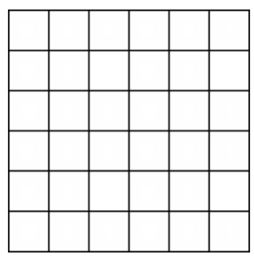
Slide credit: Dalal, Triggs, P. Barnum



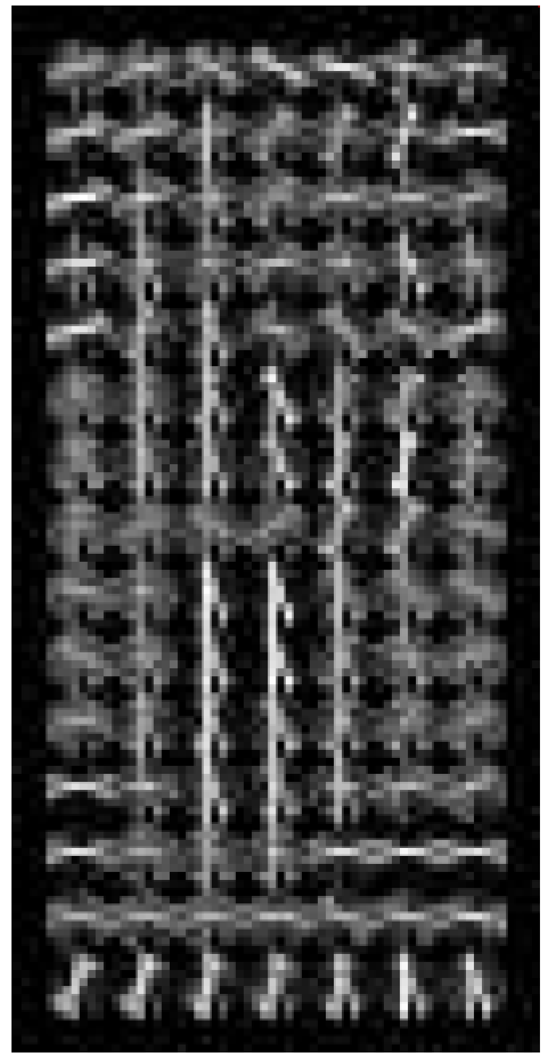
- Histogram of gradient orientations

  -Orientation -Position

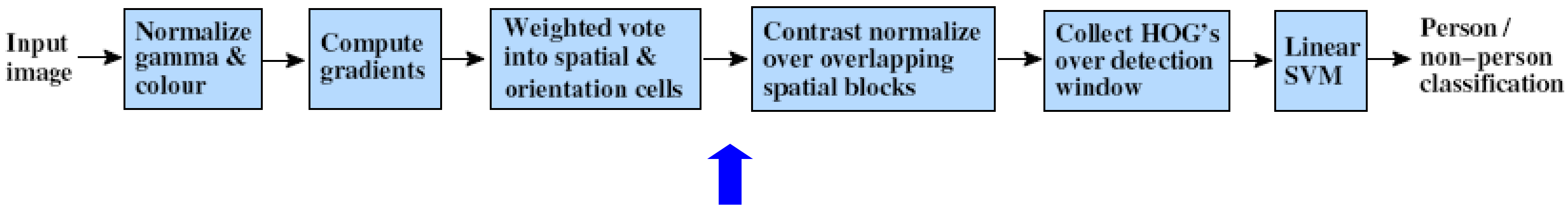


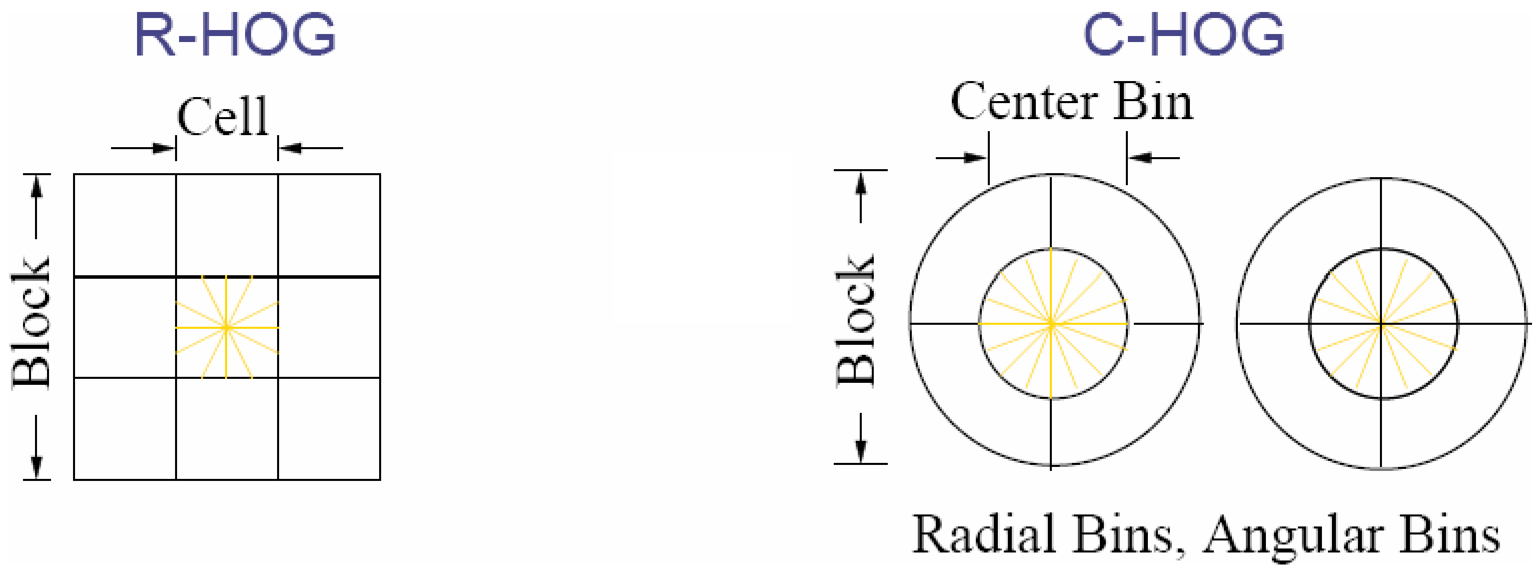


  – Weighted by magnitude

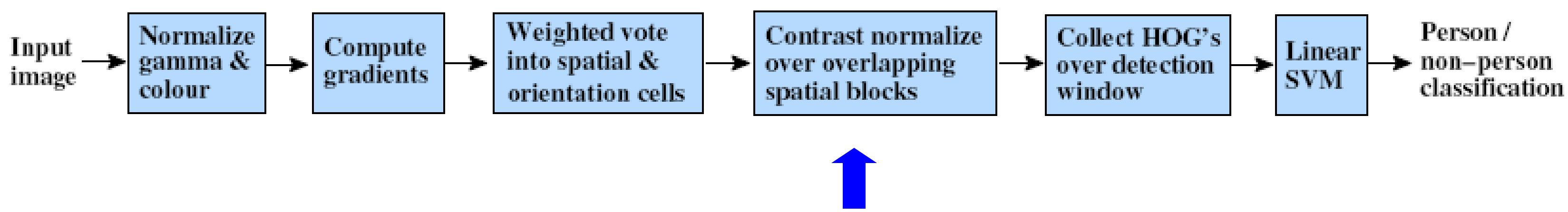


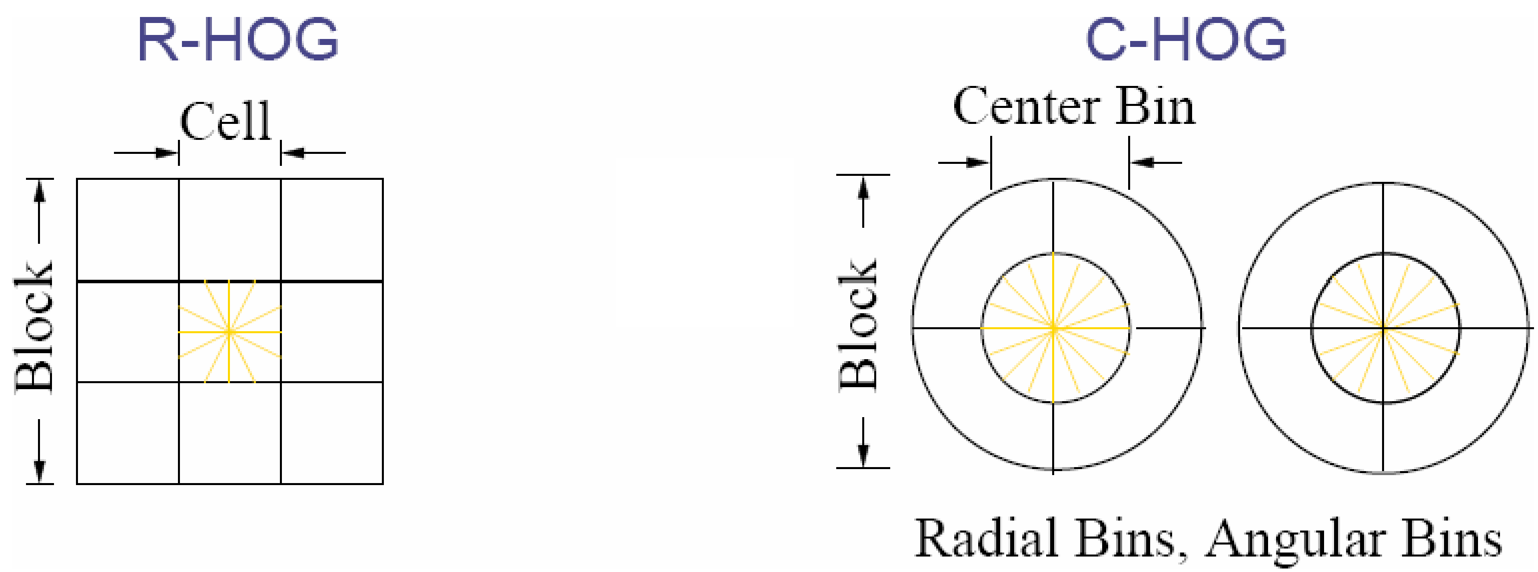
Slide credit: Dalal, Triggs, P. Barnum





Slide credit: Dalal, Triggs, P. Barnum





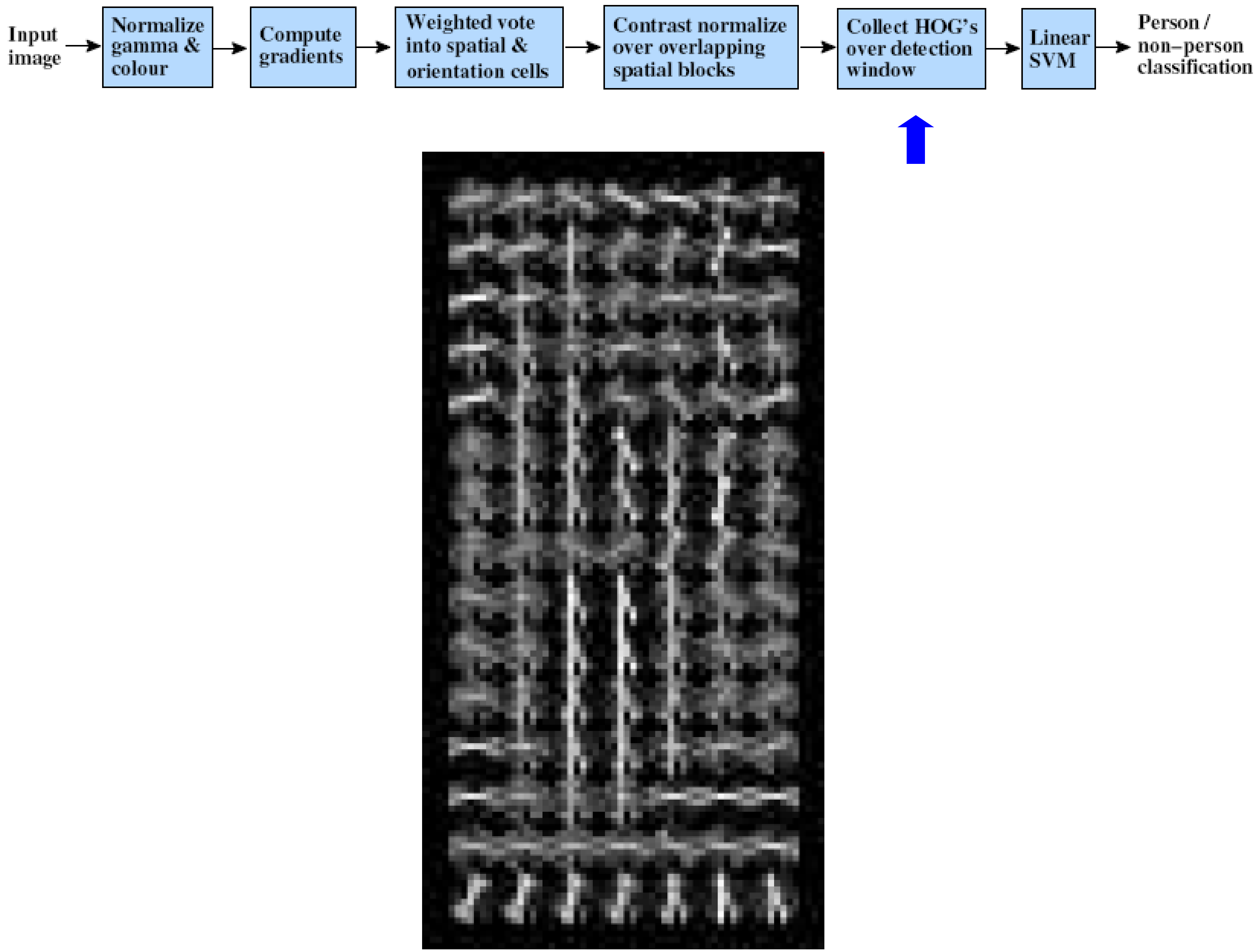
$L1-norm : v \longrightarrow v/(||v||_1+\epsilon)$

$L1-sqrt : v \longrightarrow \sqrt{v/(||v||_1+\epsilon)}$

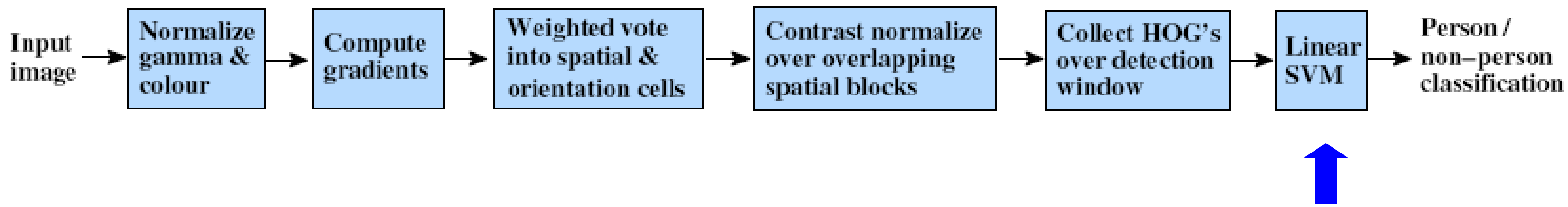
$L2-norm : v \longrightarrow v/\sqrt{||v||_2^2+\epsilon^2}$

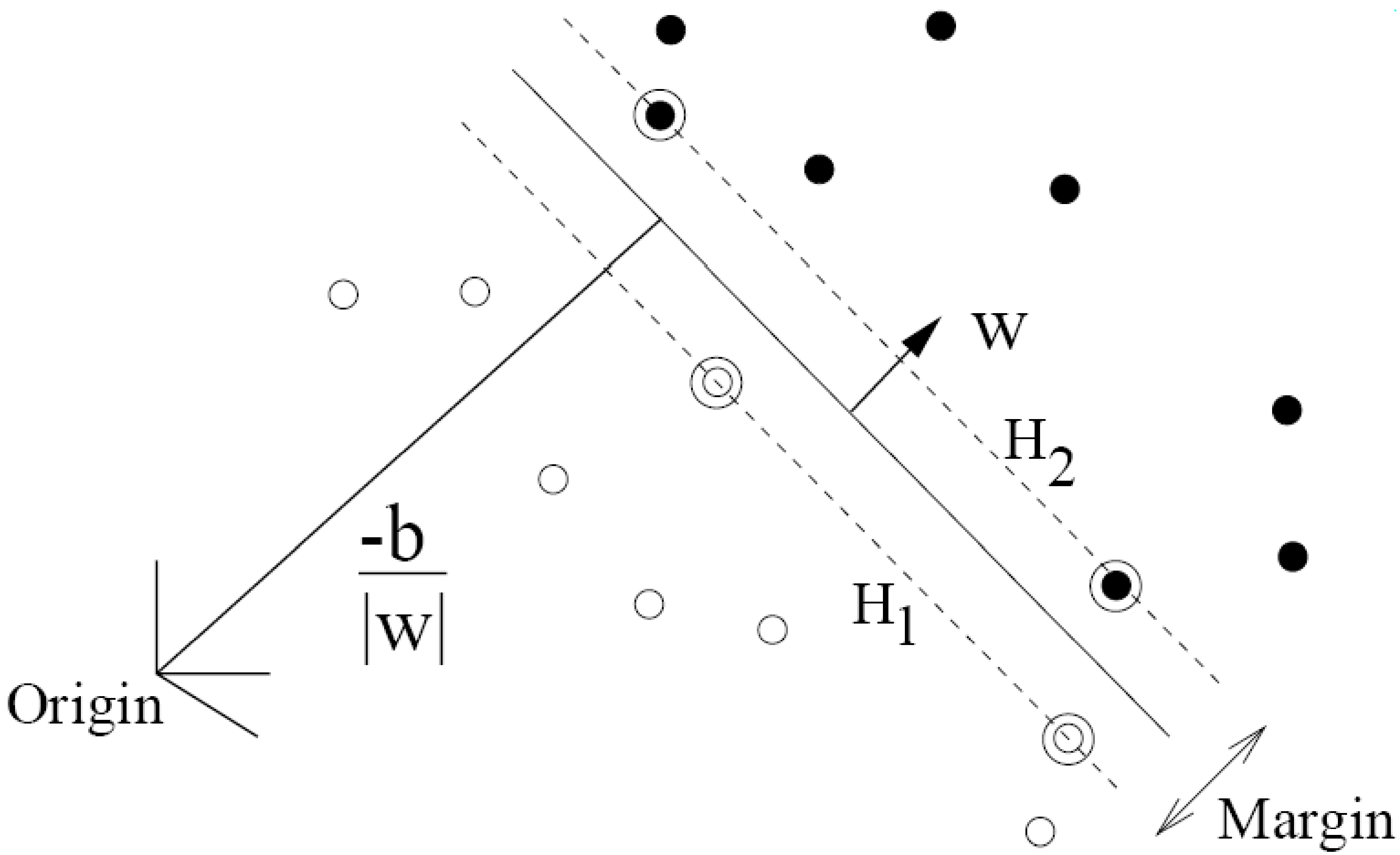
$L2-hys$: L2-norm, plus clipping at .2 and renomalizing

Slide credit: Dalal, Triggs, P. Barnum



Slide credit: Dalal, Triggs, P. Barnum





Slide credit: Dalal, Triggs, P. Barnum

Input image → Normalize gamma & colour → Compute gradients → Weighted vote into spatial & orientation cells → Contrast normalize over overlapping spatial blocks → Collect HOG's over detection window → Linear SVM → Person / non–person classification

Slide credit: Dalal, Triggs, P. Barnum

Slide credit: Dalal, Triggs, P. Barnum

Today

- Background / Overview

- Histograms of edges (Schiele)
- Windowed spectral analysis (GIST)
- Tiled histograms of edges (HOG)

- **Motion History Images (Bobick)**
- Rectified Flow Descriptors (Efros)
- Differential Geometry Signatures (Shah)

Movement: primitive motion

- *Movements* are:
  - atomic, indivisible
  - defined by motion
  - typically a “simple” trajectory in some parameter space
  - temporal variation is at most scaling
  - require almost no knowledge, reasoning, or model of time to recognize

- Examples:

  Baseball: swinging a bat

  Ballet - how do you see a plié?

  Virtual PAT (“temporal templates”)

Slide credit: Davis, Bobick

Strict Appearance: human movements

- Is recognizing movement a 3D or 2D problem? Simple human ~~psychophysics~~ demonstration and computational complexity argue for 2D aspects.
- Temporal templates: Movements are recognized directly from the motion.
- Appearance-based recognition can assist geometric recovery: recognition labels the parts and allows extraction.

Slide credit: Davis, Bobick

Blurry Video

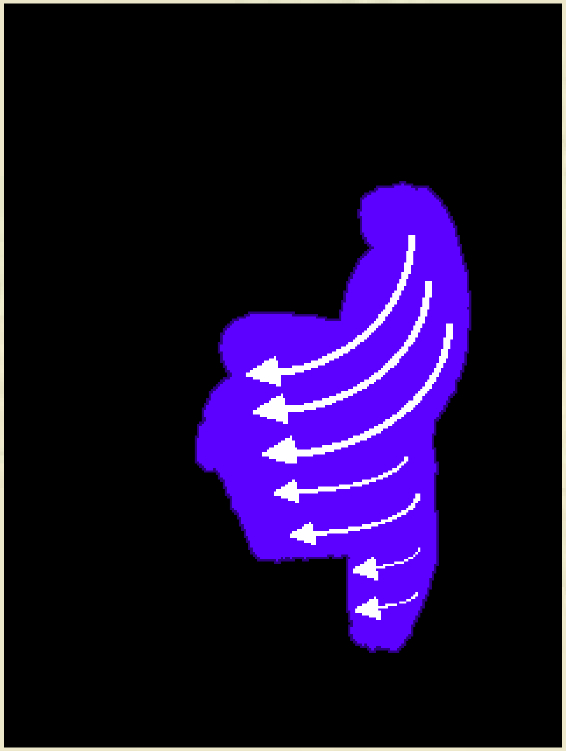
Slide credit: Davis, Bobick

Motivating Example

Slide credit: Davis, Bobick

Shape and motion: view-based

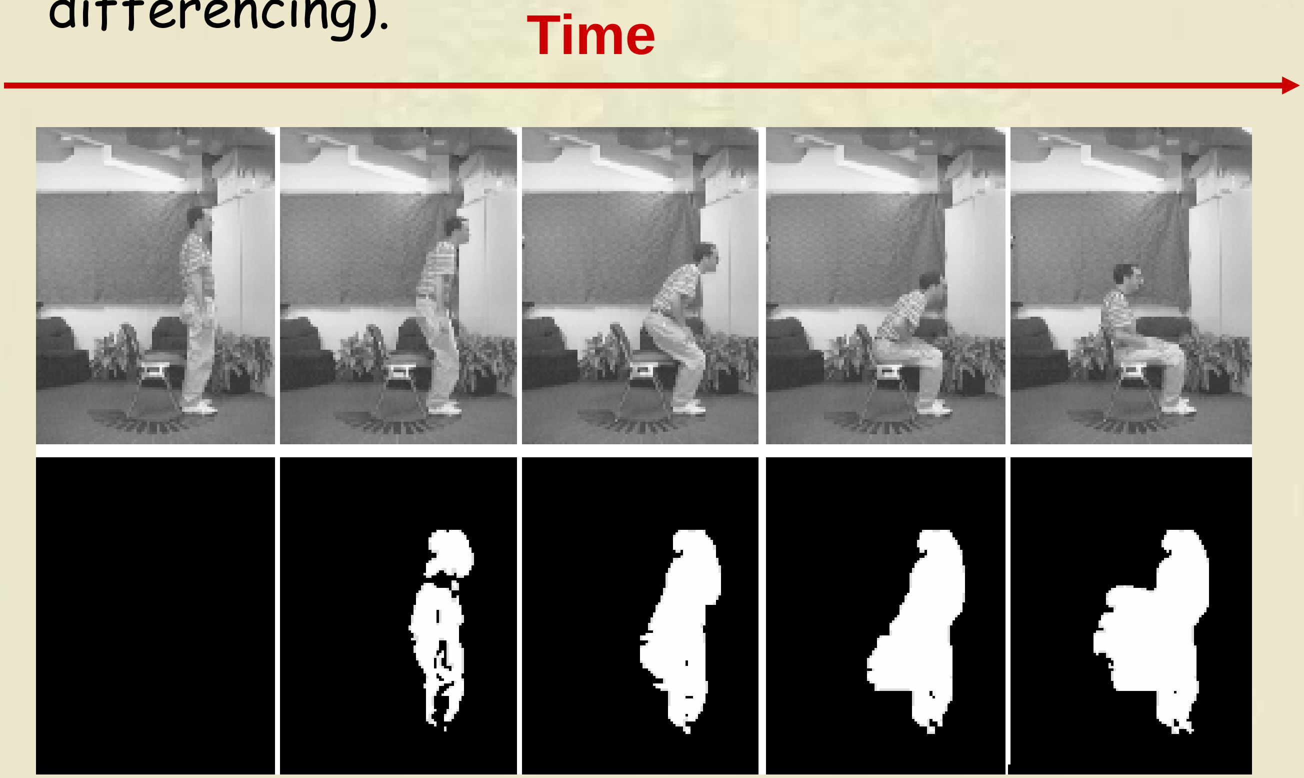
- Schematic representation of sitting at 90°



Slide credit: Davis, Bobick

Motion energy images

- Spatial accumulation of motion.
- Collapse over specific time window.
- Motion measurement method not critical (e.g. motion differencing).



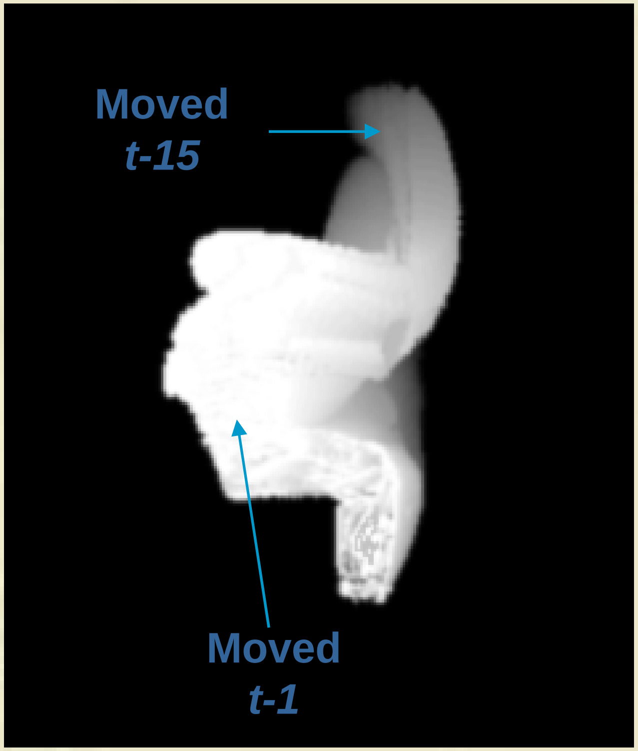
Slide credit: Davis, Bobick

Motion history images

- Motion history images are a different function of temporal volume.
- Pixel operator is replacement decay:

  if moving $I_\tau(x,y,t) = \tau$
  otherwise
  $I_\tau(x,y,t) = \max(I_\tau(x,y,t-1)-1\ ,0)$

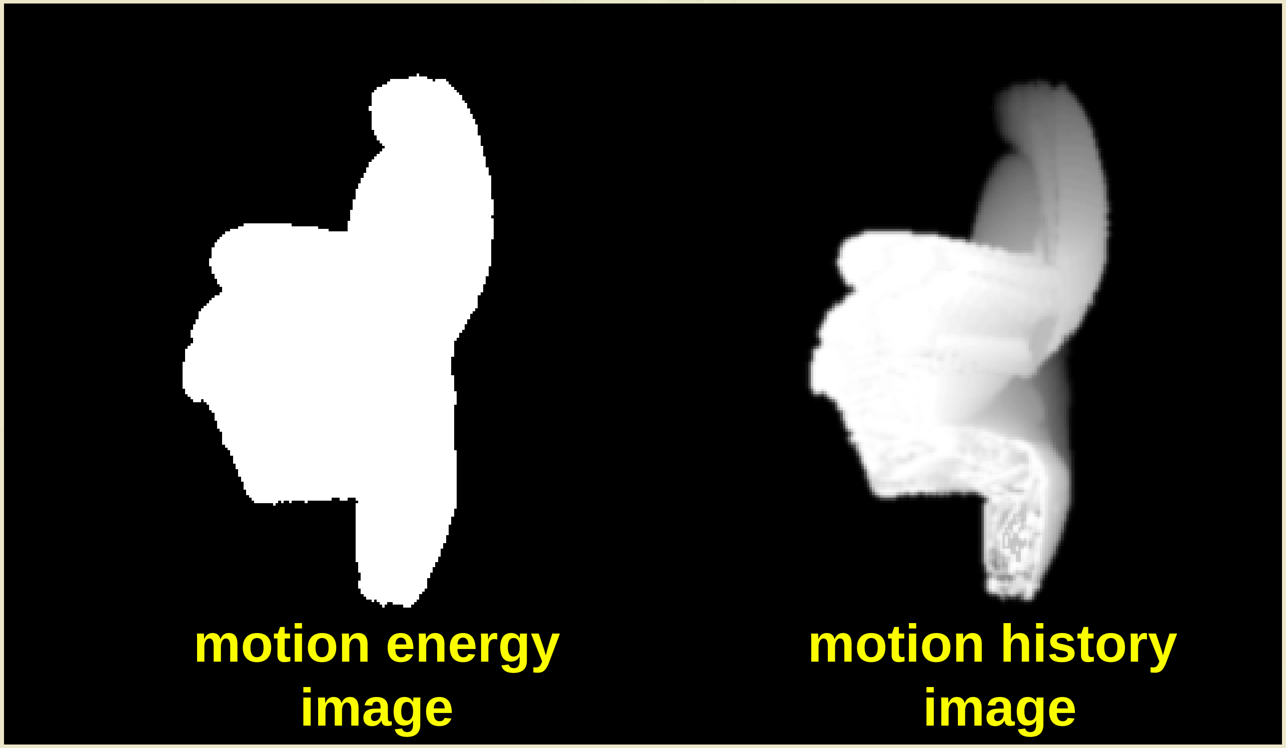
- Trivial to construct $I_{\tau-k}(x,y,t)$ from $I_\tau(x,y,t)$ so can process multiple time window lengths without more search.
- MEI is thresholded MHI



Slide credit: Davis, Bobick

Temporal-templates

- *MEI+ MHI = Temporal template*



Slide credit: Davis, Bobick

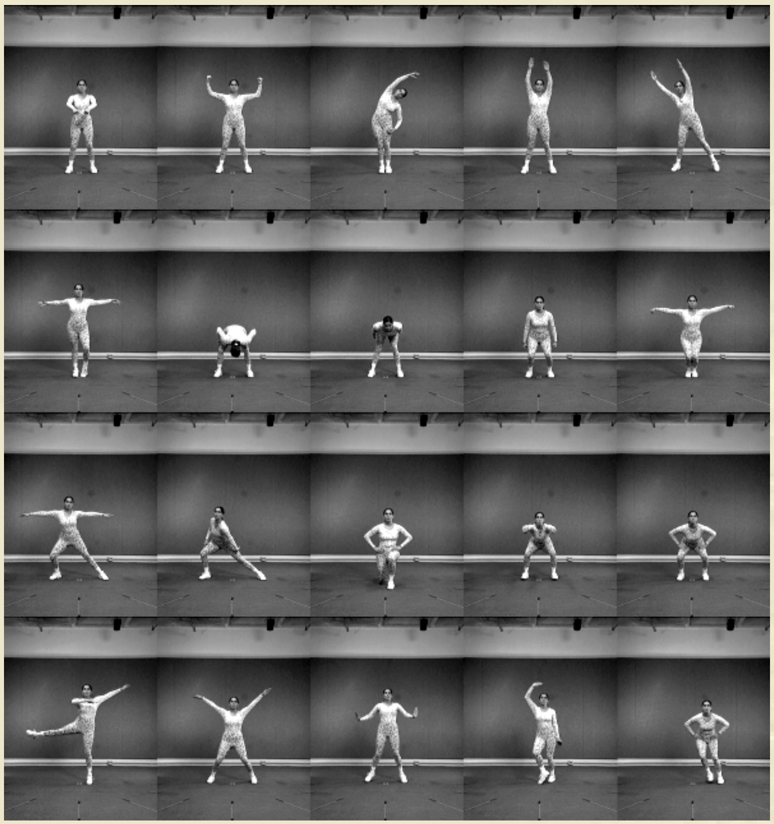
Recognizing temporal templates

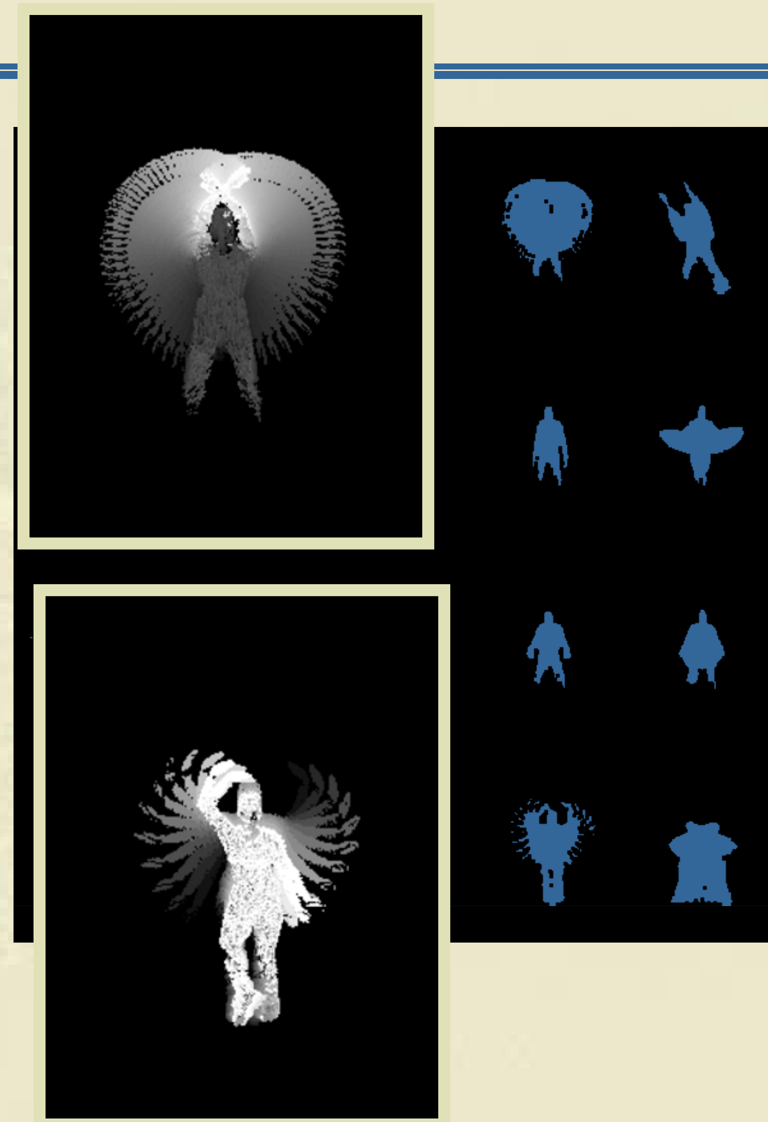
(PAMI 2001, Bobick and Davis)

- For MEI and MHI compute global properties (e.g. Hu moments). Treat both as grayscale images.
- Collect statistics on distribution of those properties over people for each movement.
- At run time, construct MEIs and MHIs backwards in time.
  - Recognizing movements as soon as they complete.
- Linear time scaling.
  - Compute range of τ using the min and max of training data.
- Simple recursive formulation therefore very fast.
- Filter implementation obvious so biologically “relevant”.

Slide credit: Davis, Bobick

Aerobics examples



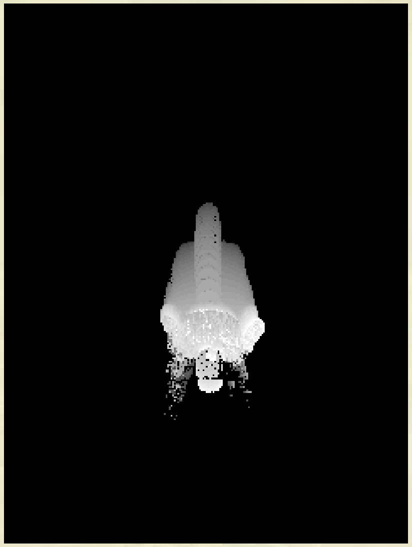


Slide credit: Davis, Bobick

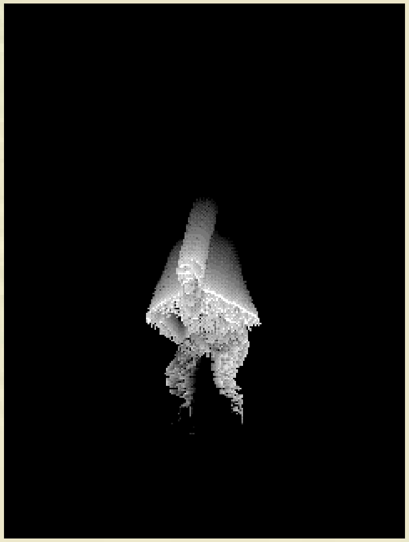
Aerobics with one camera

- With one camera:
  - 12 of 18 moves when viewed at 30° correctly identified.
  - Confusion stems from different views of different moves.

Input



Closest

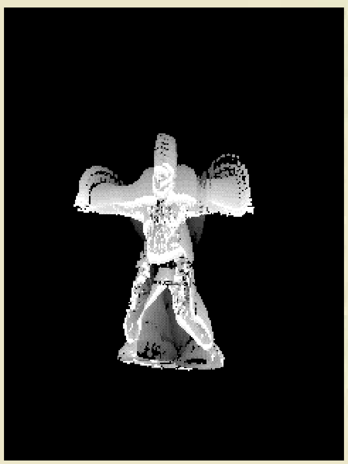


Correct

Slide credit: Davis, Bobick

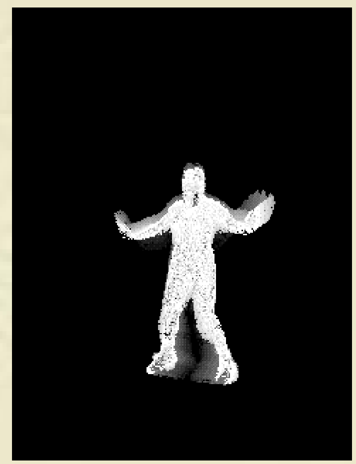
Aerobics with two cameras

- With two cameras:
  - 15 of 18 moves when viewed at 30° correctly identified; others second or third
  - Confusion stems from bad image differencing.



Input

Closest



Correct

Slide credit: Davis, Bobick

Virtual PAT (Personal Aerobics Trainer)

- Uses MHI recognition
- Portable IR background subtraction system (CAPTECH '98)

Slide credit: Davis, Bobick

The KidsRoom

- A narrative, interactive children's playspace.
- Demonstrates computer vision "action" recognition.
- Someitmes, possible because the machine knows the context.
- A kinder, gentler C^3I interface
- Ported to the Millenium Dome, London, 2001
- Summary and critique in *Presence, August 1999.*

Slide credit: Davis, Bobick

Recognizing Movement in the KidsRoom

- First teach the kids, then observe.
- Temporal templates "plus" (but in paper).
- Monsters always do something, *but only speak it when sure.*

Slide credit: Davis, Bobick

Today

- Background / Overview

- Histograms of edges (Schiele)
- Windowed spectral analysis (GIST)
- Tiled histograms of edges (HOG)

- Motion History Images (Bobick)
- **Rectified Flow Descriptors (Efros)**
- Differential Geometry Signatures (Shah)

Recognizing Action at a Distance

A. Efros, A. Berg, G. Mori, J. Malik
UC Berkeley

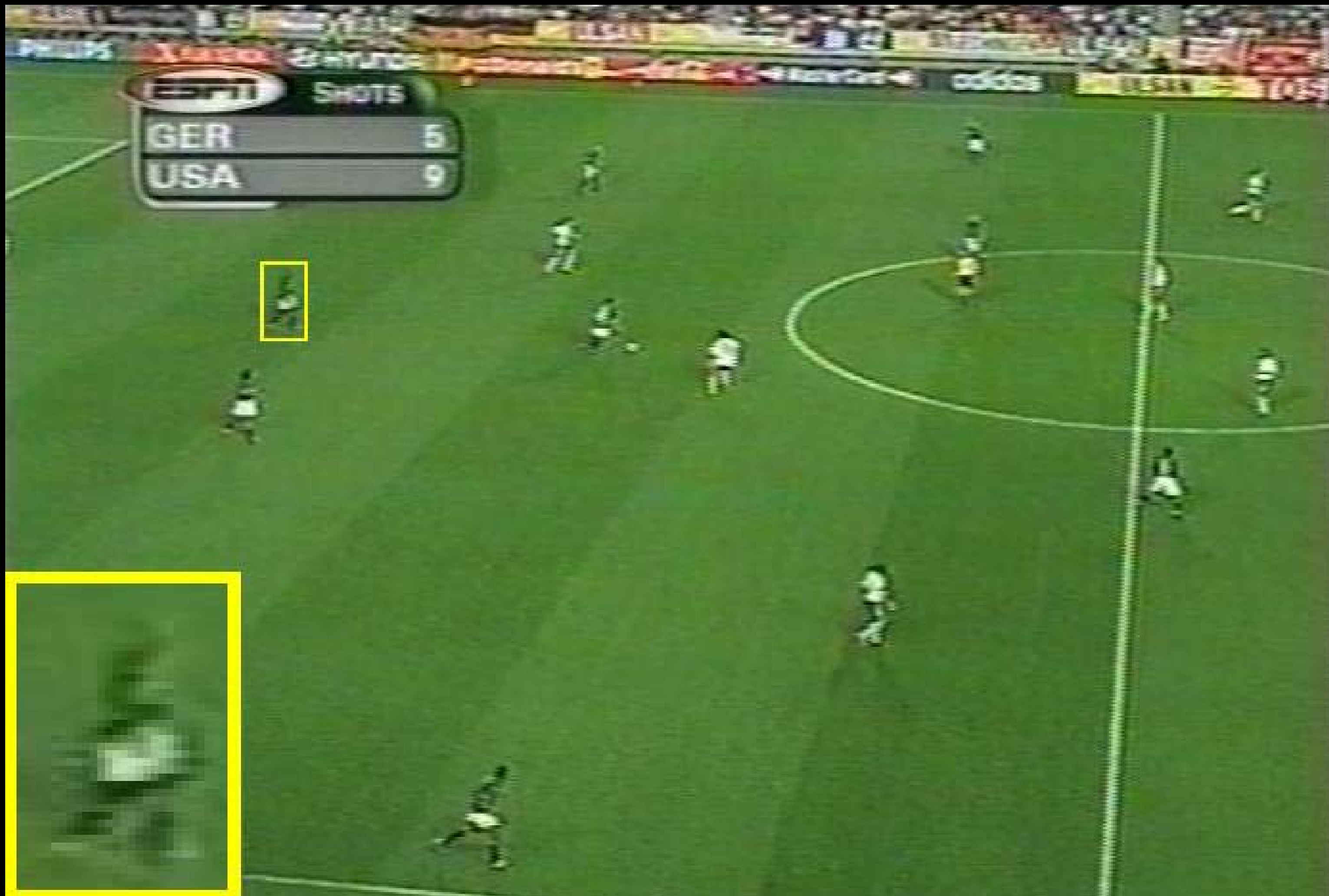
Slide credit: Malik

Medium Field

- Recognize human actions
  - Real-world setting
  - Low resolution, noisy data
  - Moving camera, occlusions

Slide credit: Malik

Medium-field Recognition



The 30-Pixel Man

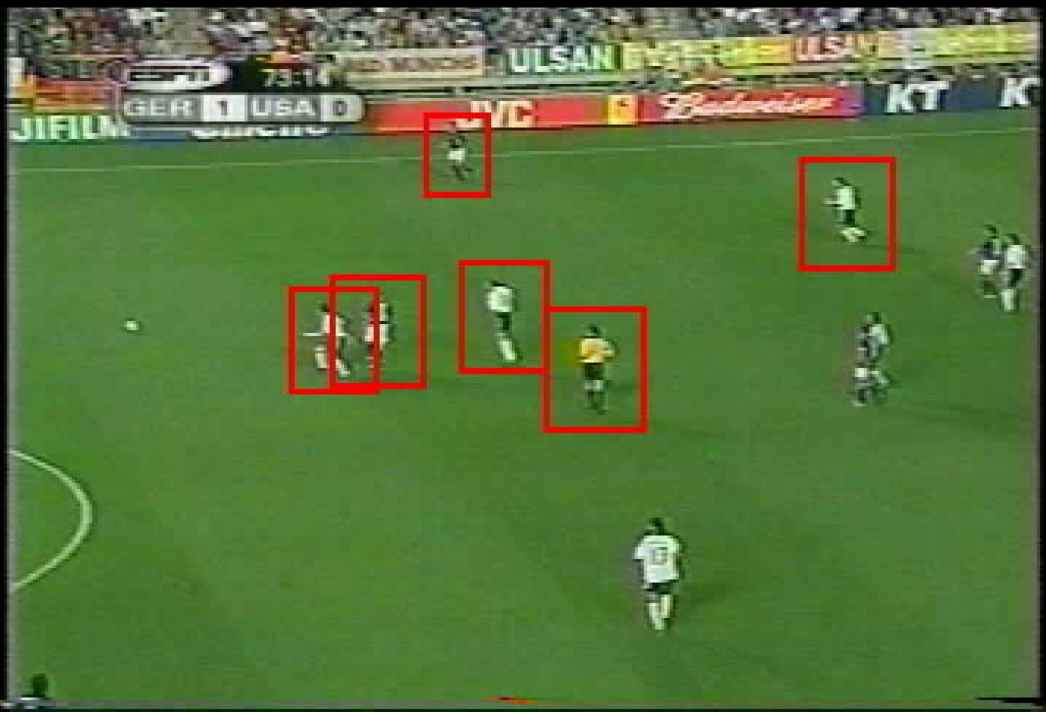
Slide credit: Malik

Our Approach

- Non parametric image-based approach
- Use large amount of data
- Compute motion descriptors
  - Aggregate of low-level motion features
- Classify a novel motion by finding the most similar motion from the training set

Slide credit: Malik

Gathering action data

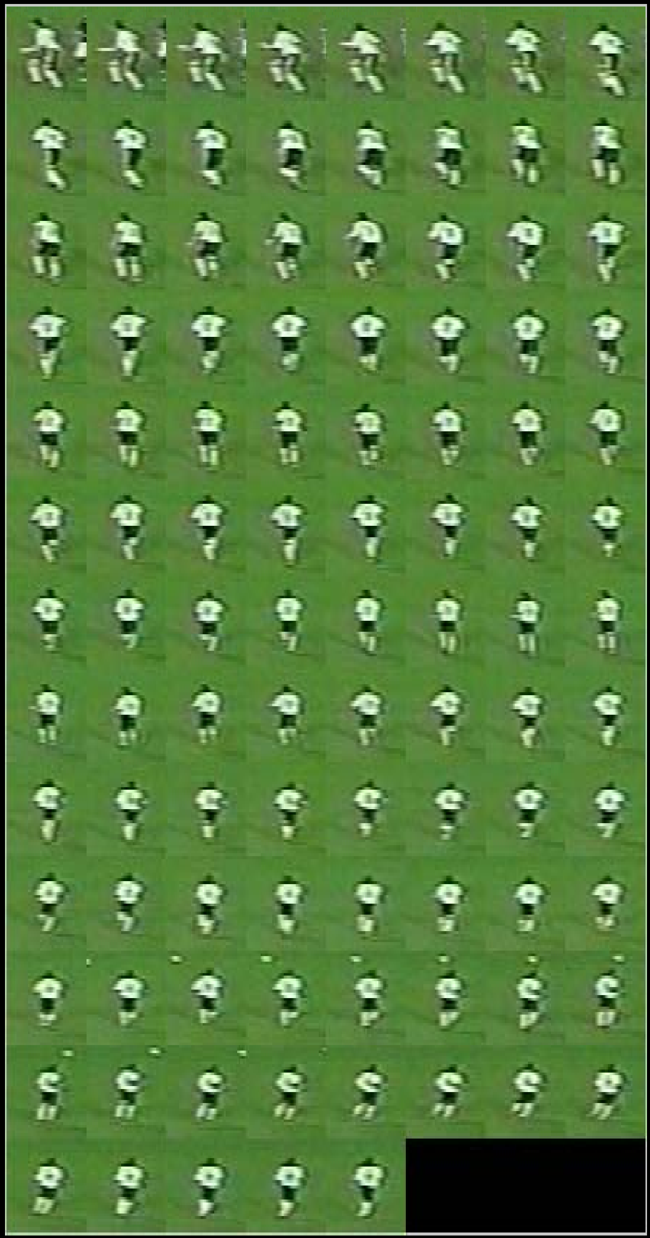


- Tracking
  - Simple correlation-based tracker
  - User-initialized

Slide credit: Malik

Figure-centric Representation

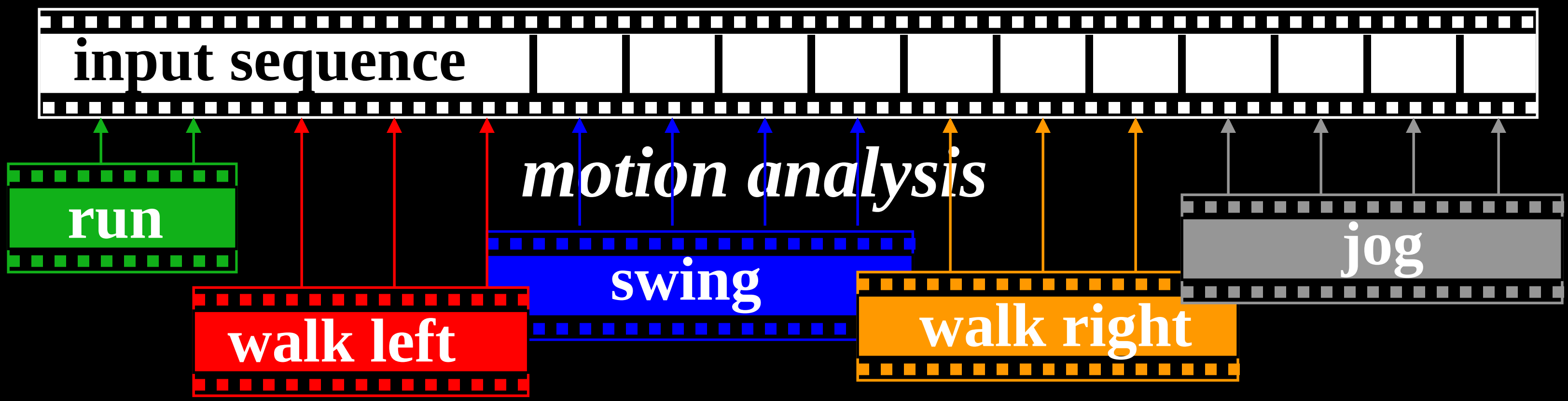
- Stabilized spatio-temporal volume
  - No translation information
  - All motion caused by person's limbs
    - Good news: indifferent to camera motion
    - Bad news: hard!
- Good test to see if actions, not just translation, are being captured



Slide credit: Malik

Remembrance of Things Past

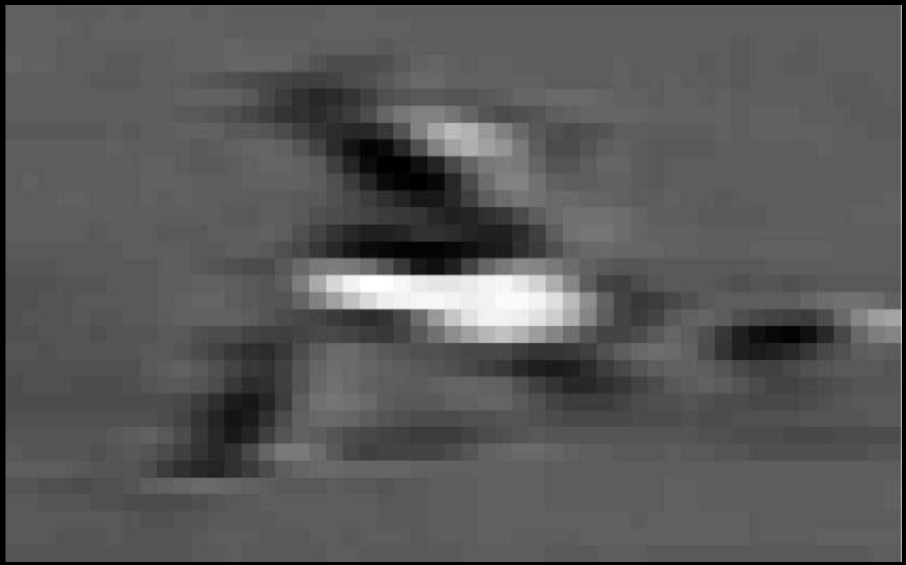
- "Explain" novel motion sequence by matching to previously seen video clips
  - For each frame, match based on some temporal extent



<u>Challenge:</u> how to compare motions?

Slide credit: Malik

Motion Descriptor



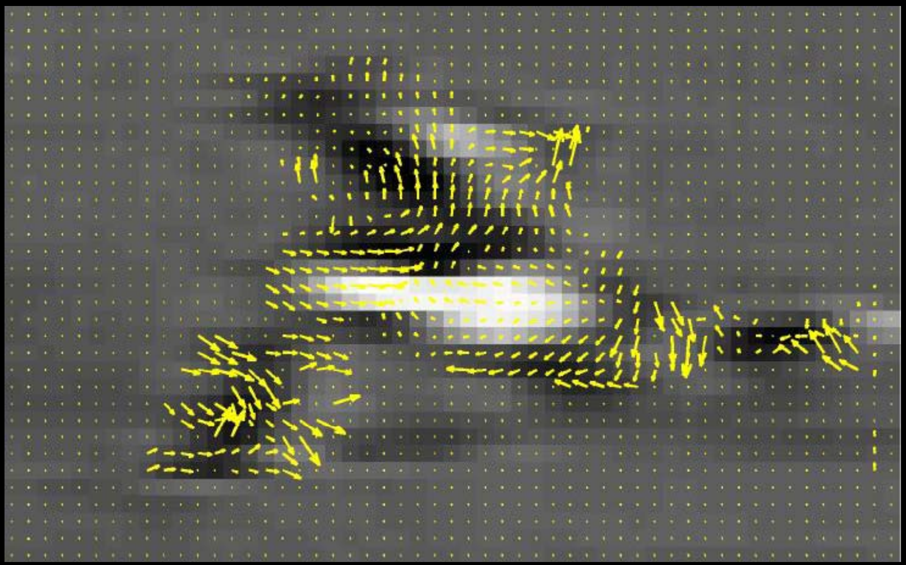
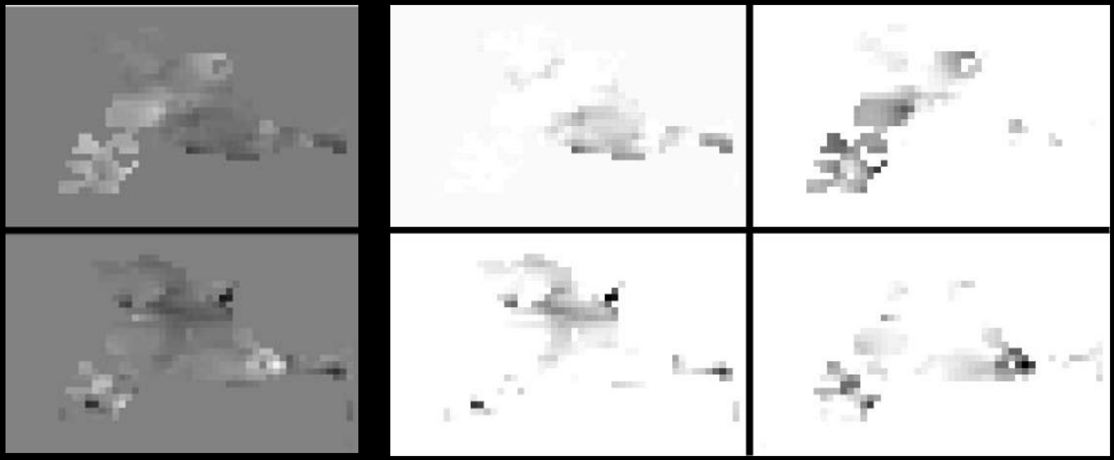
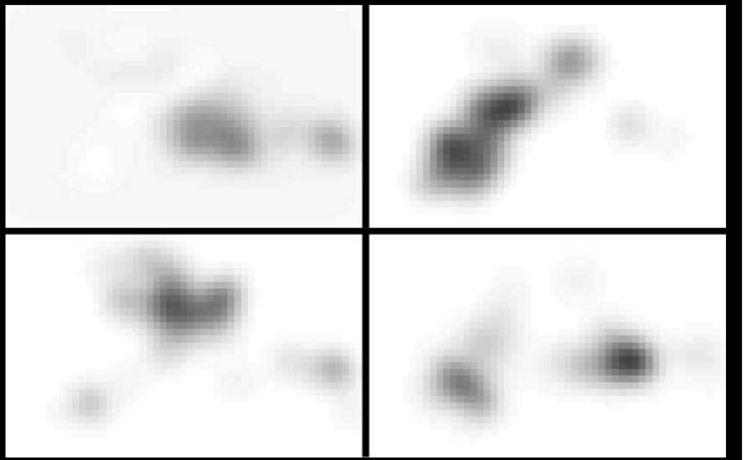


Image frame

Optical flow

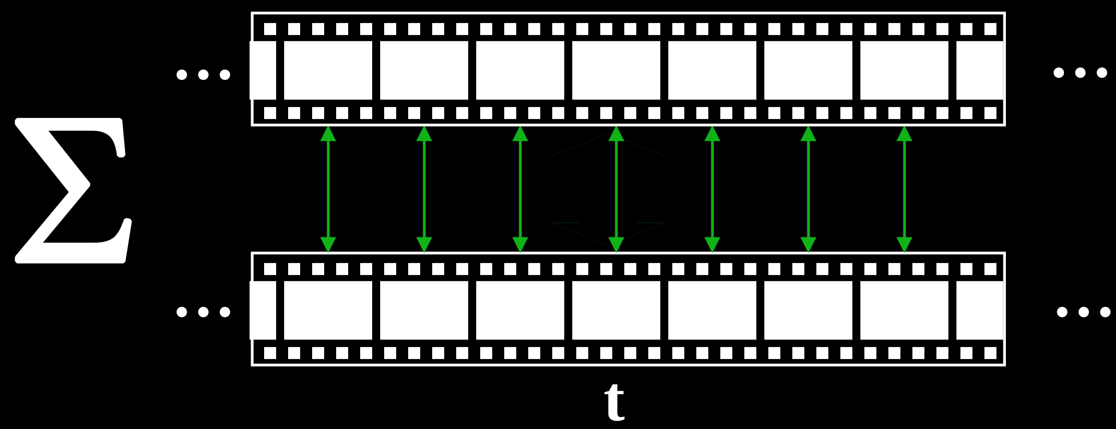


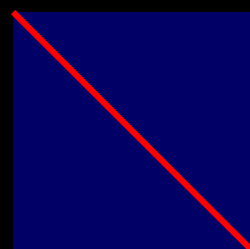


blurred

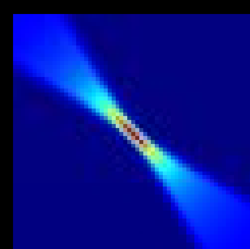
Slide credit: Malik

Comparing motion descriptors

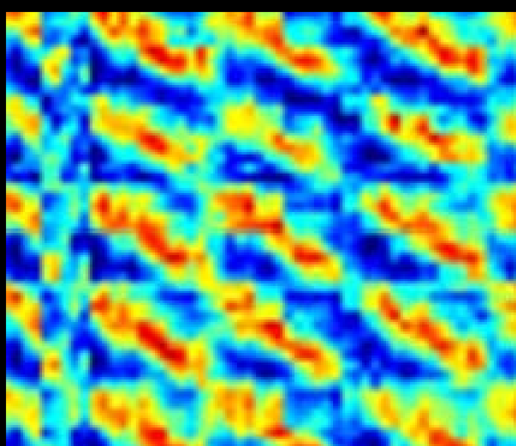




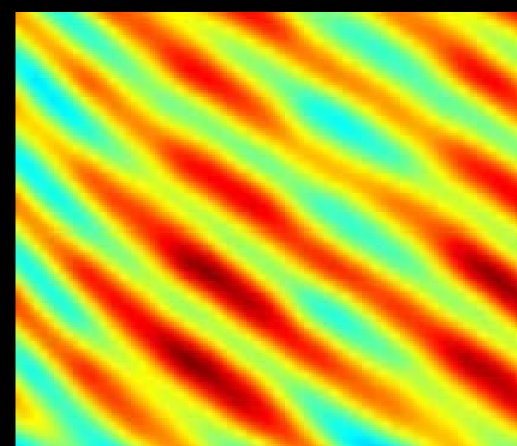
I matrix



blurry I



frame-to-frame
similarity matrix

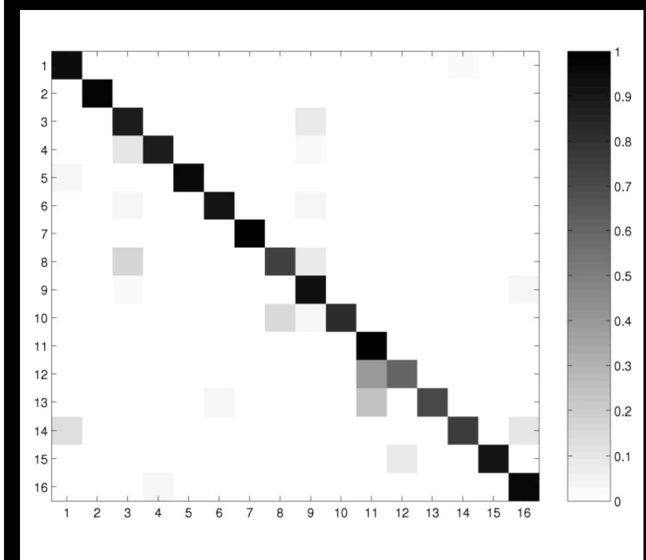


motion-to-motion
similarity matrix

Slide credit: Malik

Classifying Ballet Actions

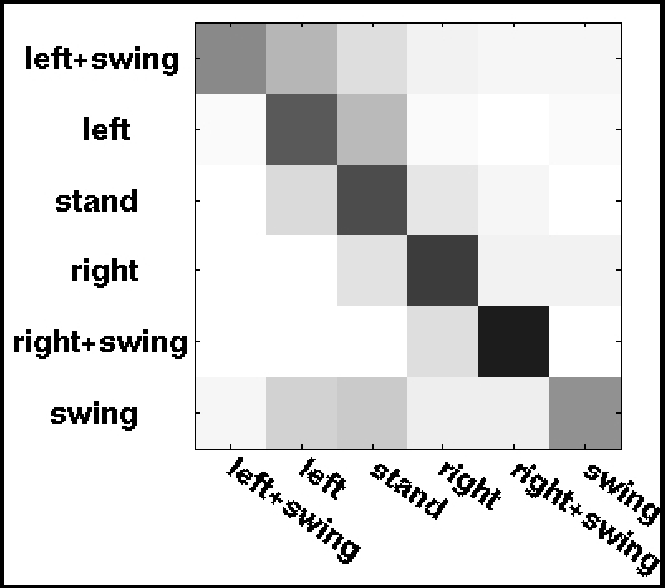
16 Actions. Men used to classify women and vice versa.



Slide credit: Malik

Classifying Tennis Actions

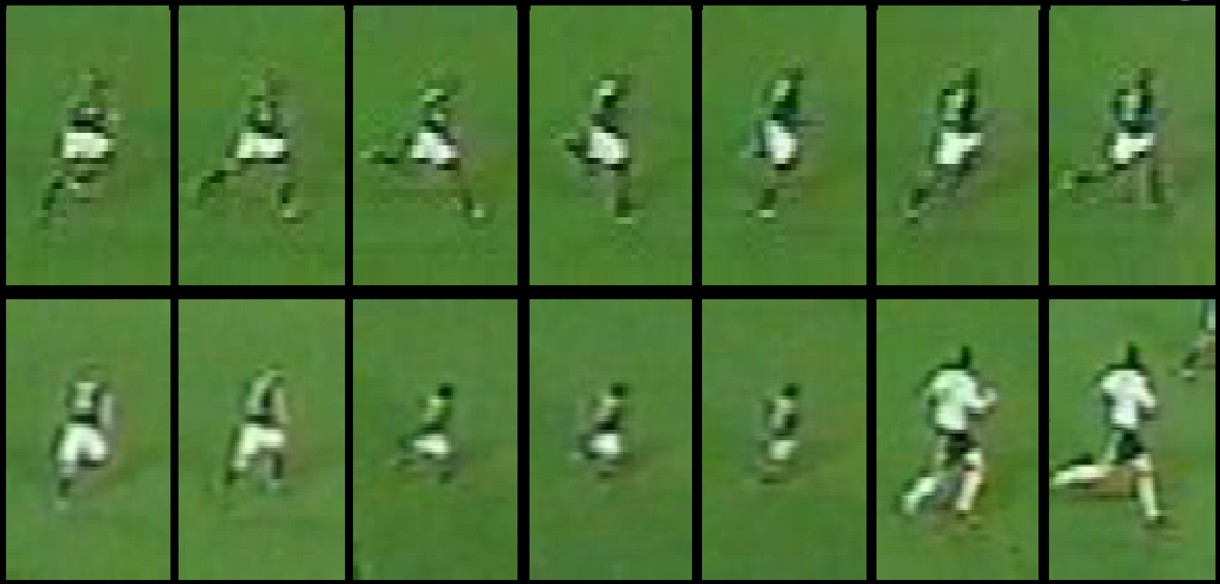
6 actions. Woman player used as training, man as testing.

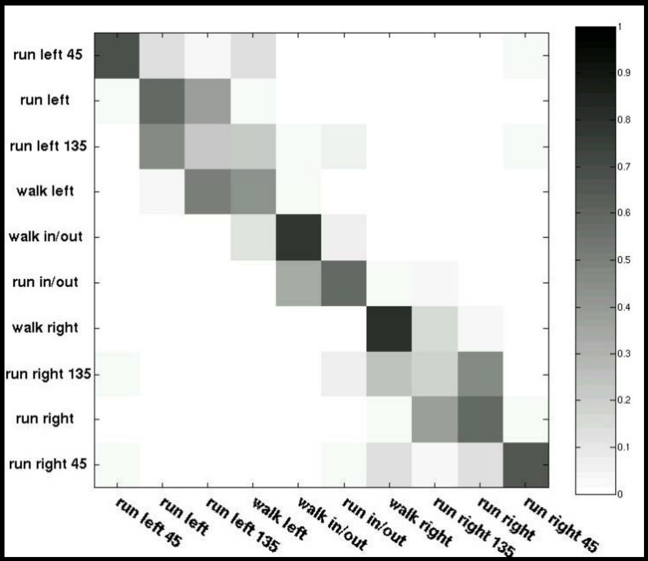


Slide credit: Malik

Classifying Soccer Actions

10 Actions. Leave one sequence out testing.





Slide credit: Malik

Skeleton Transfer

- Annotate database with joint positions
- After matching, transfer data to novel sequence
  - Ajust the match for best fit
- 3D MoCap data as synthetic annotated database

Slide credit: Malik

Remarks

- Purely motion-based descriptor for actions
- Treat optical flow
  - Not as measurement of pixel displacement
  - But as a set of noisy *features* that are carefully smoothed and aggregated

Slide credit: Malik

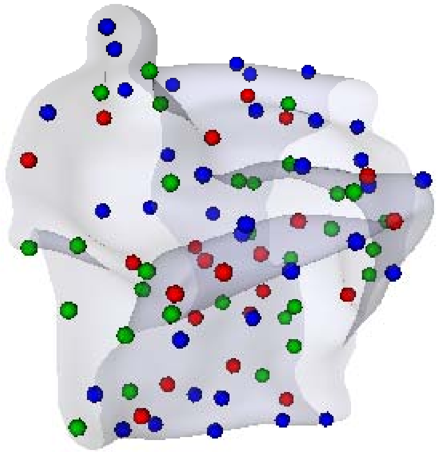
Today

- Background / Overview

- Histograms of edges (Schiele)
- Windowed spectral analysis (GIST)
- Tiled histograms of edges (HOG)

- Motion History Images (Bobick)
- Rectified Flow Descriptors (Efros)
- **Differential Geometry Signatures (Shah)**

Action As Objects



Alper Yilmaz and Mubarak Shah

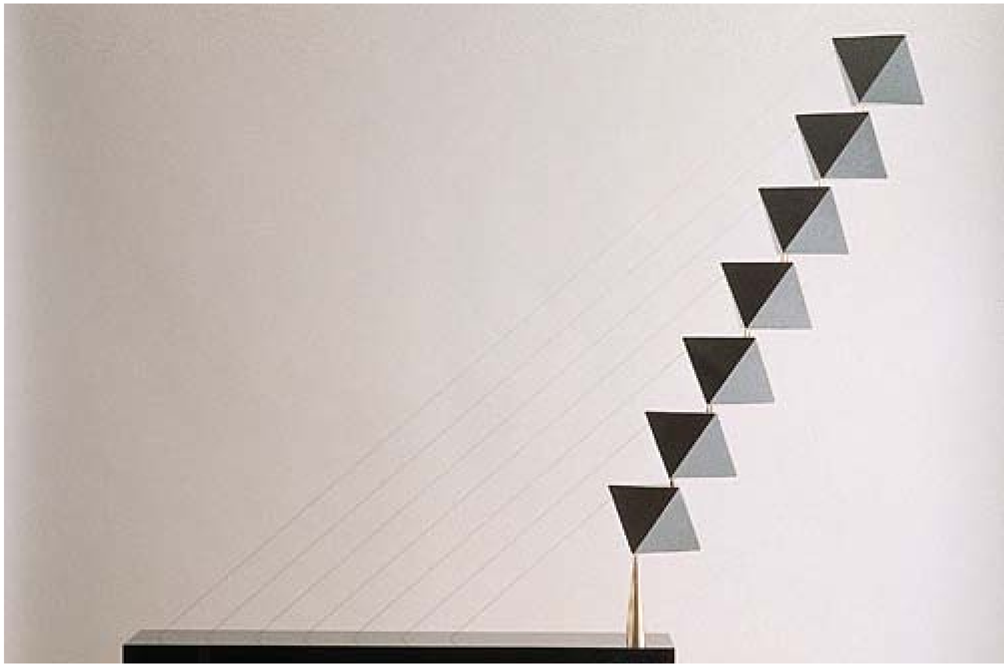
- A. Yilmaz and M. Shah "Actions Sketch: A Novel Action Representation," IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2005.
- A. Yilmaz and M. Shah “Representing Actions Using Differential Geometry," Computer Vision and Image Understanding (CVIU), submitted 2006.

Actions As Objects

When something moves it develops a shape.
Santiago Calatrava
(Sculpture into architecture)

Milwaukee Museum of Art

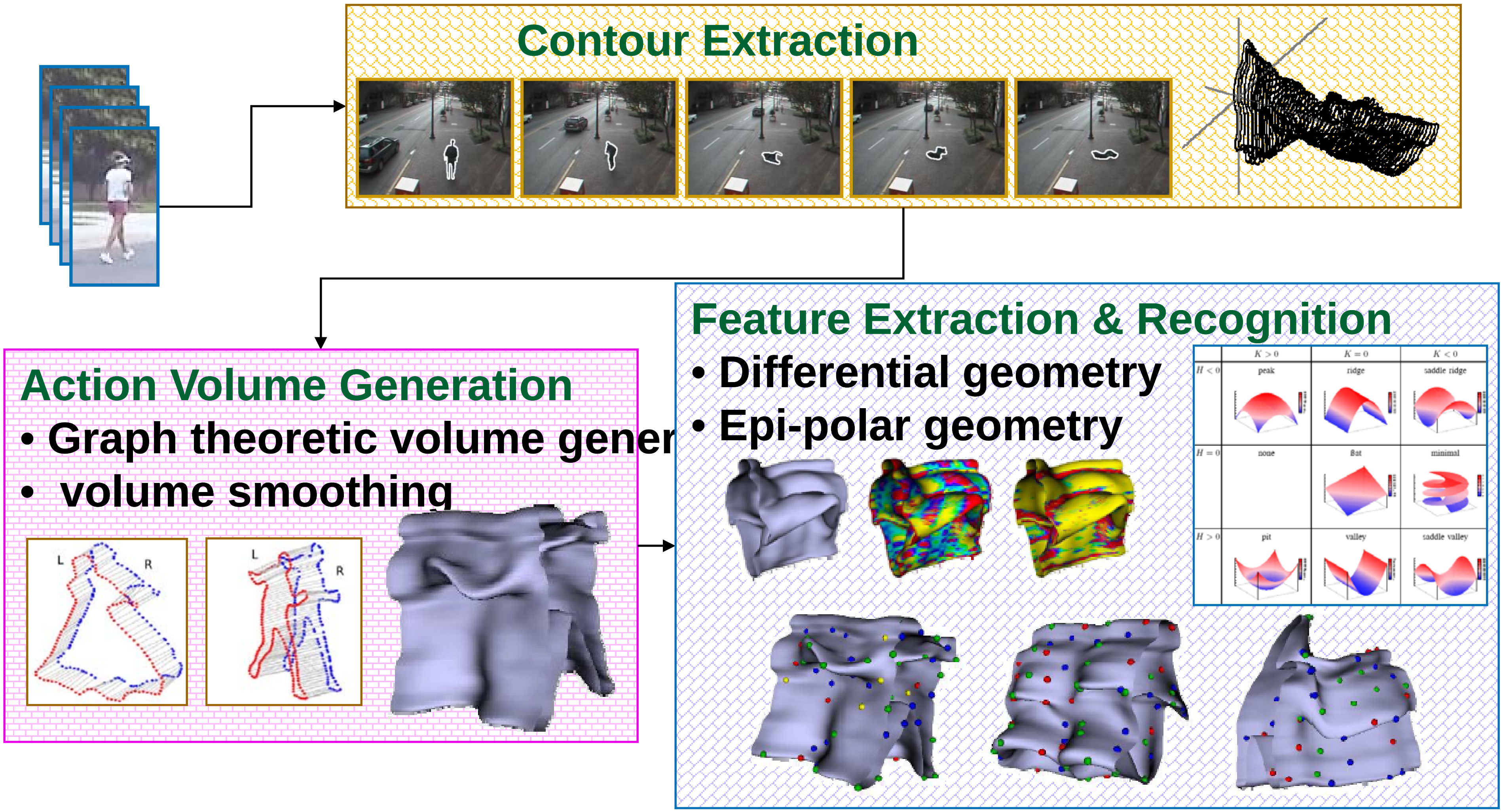
Actions As Objects



Musical Star

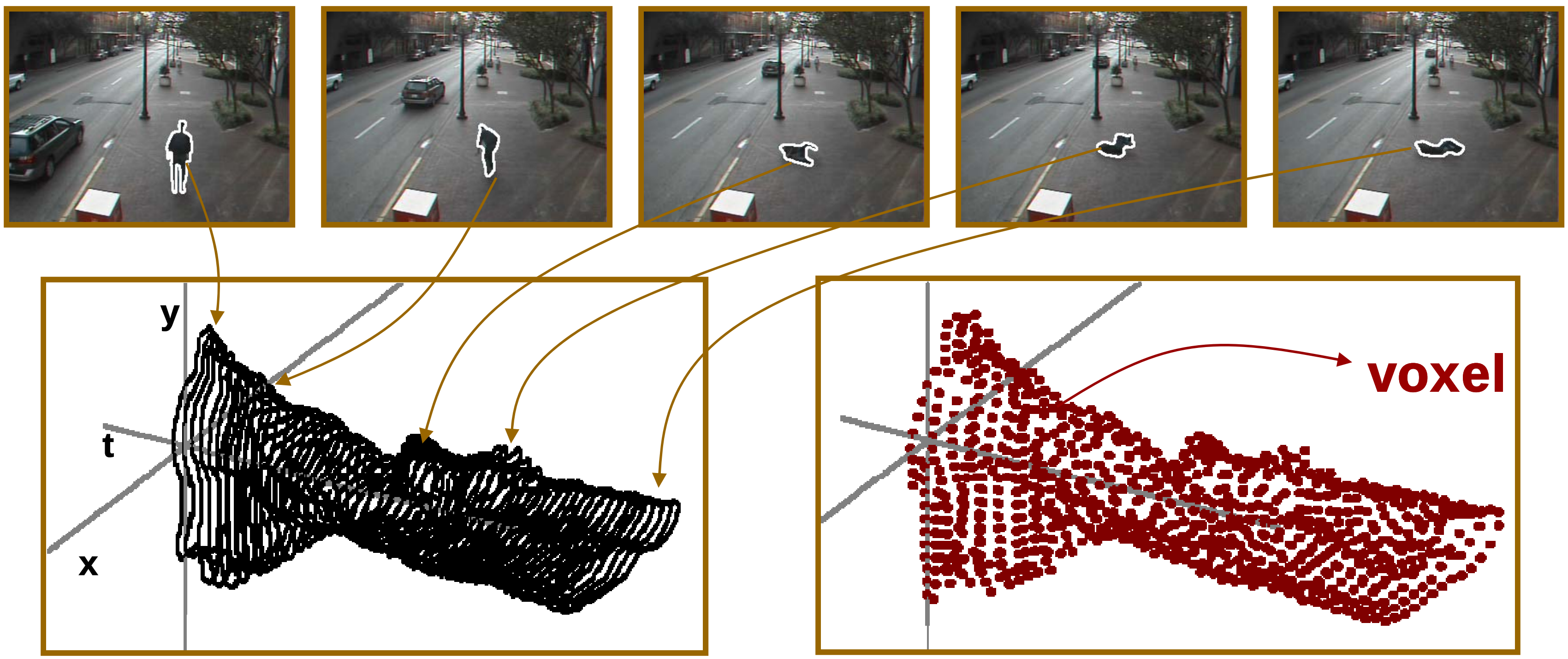
Turning Torso

Flow diagram



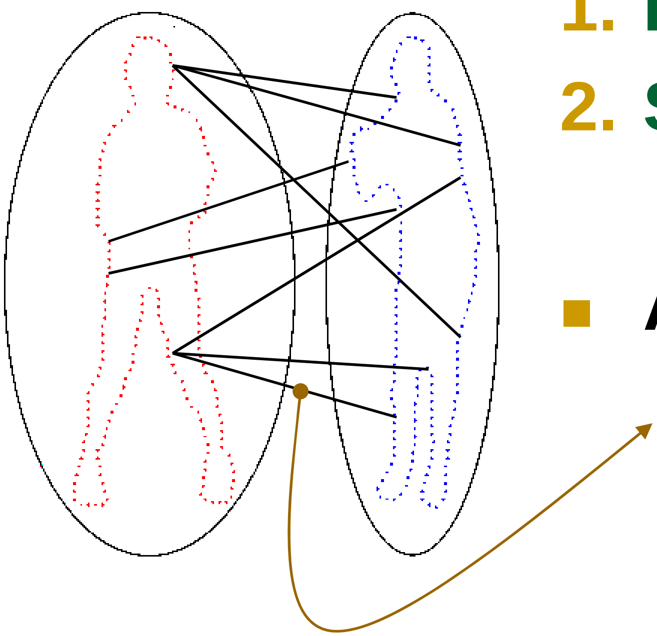
Volume Generation

- Contours from a contour tracker



Volume Generation

- Two pass correspondence approach



1. First pass: **Greedy approach**
2. Second pass: **Spatial coherence**

- **Association likelihood**
  - **Shape similarity**
  - **Proximity**
  - **Orientation similarity**

Volume Generation

- Proximity

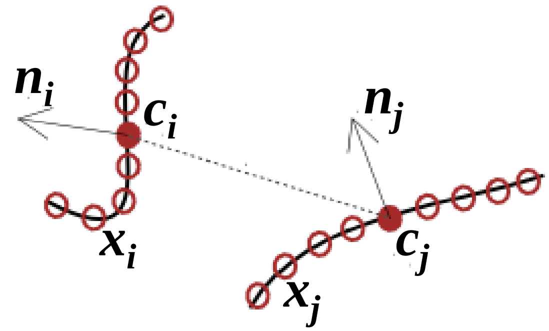
$$d_{i,j} = \left\| c_i - c_j \right\|_2$$

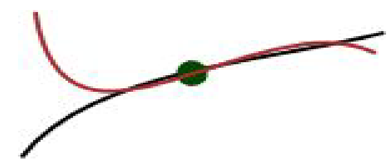
- Alignment similarity

$$\alpha_{i,j} = \arccos \frac{\vec{n}_i \cdot \vec{n}_j}{|\vec{n}_i||\vec{n}_j|}$$

- Shape similarity

$$\varepsilon_{i,j} = \sum_{x_j \in N_j} \left\| \hat{x}_i - x_j \right\|_2$$

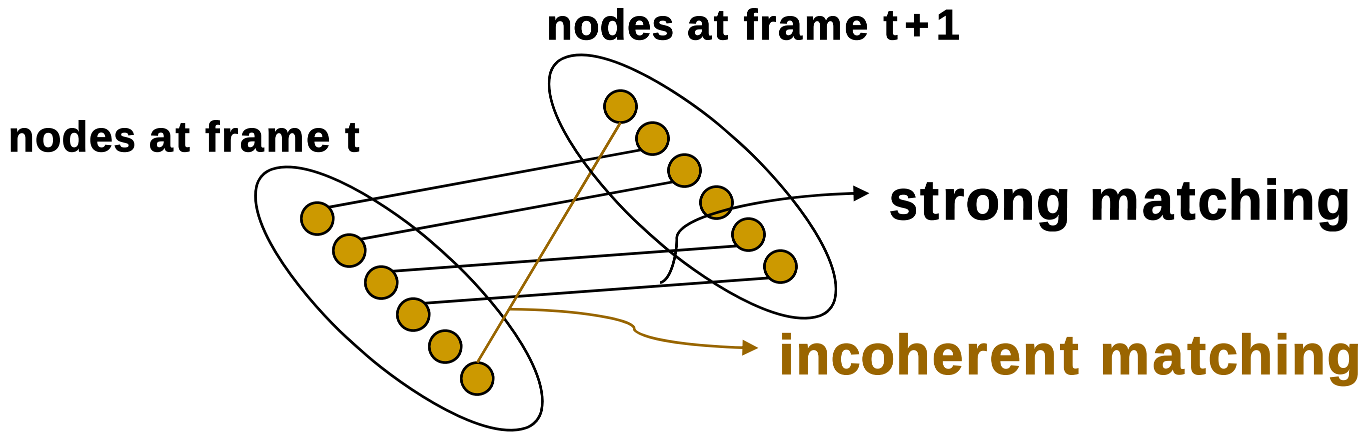




$$\boxed{l_{i,j} = \exp\left(-\frac{d_{i,j}^2}{\sigma_d^2}\right)\exp\left(-\frac{\alpha_{i,j}^2}{\sigma_\alpha^2}\right)\exp\left(-\frac{\varepsilon_{i,j}^2}{\sigma_\varepsilon^2}\right)}$$

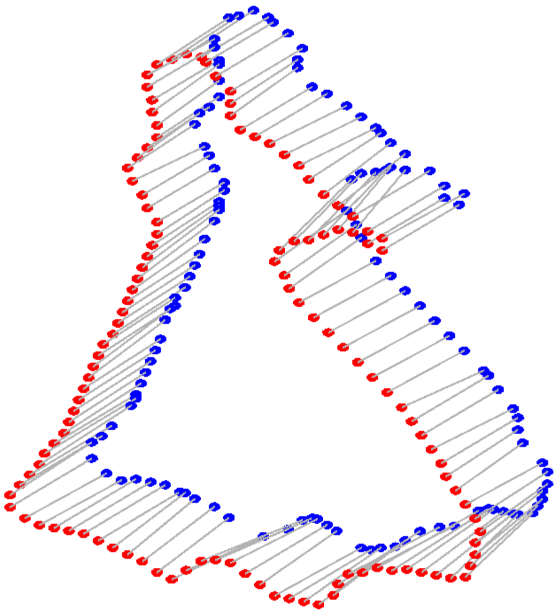
Volume Generation

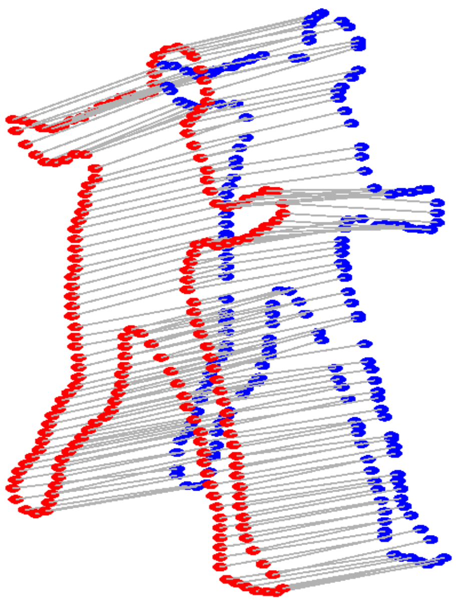
- Associate voxels with high likelihood
- Remove spatially incoherent associations



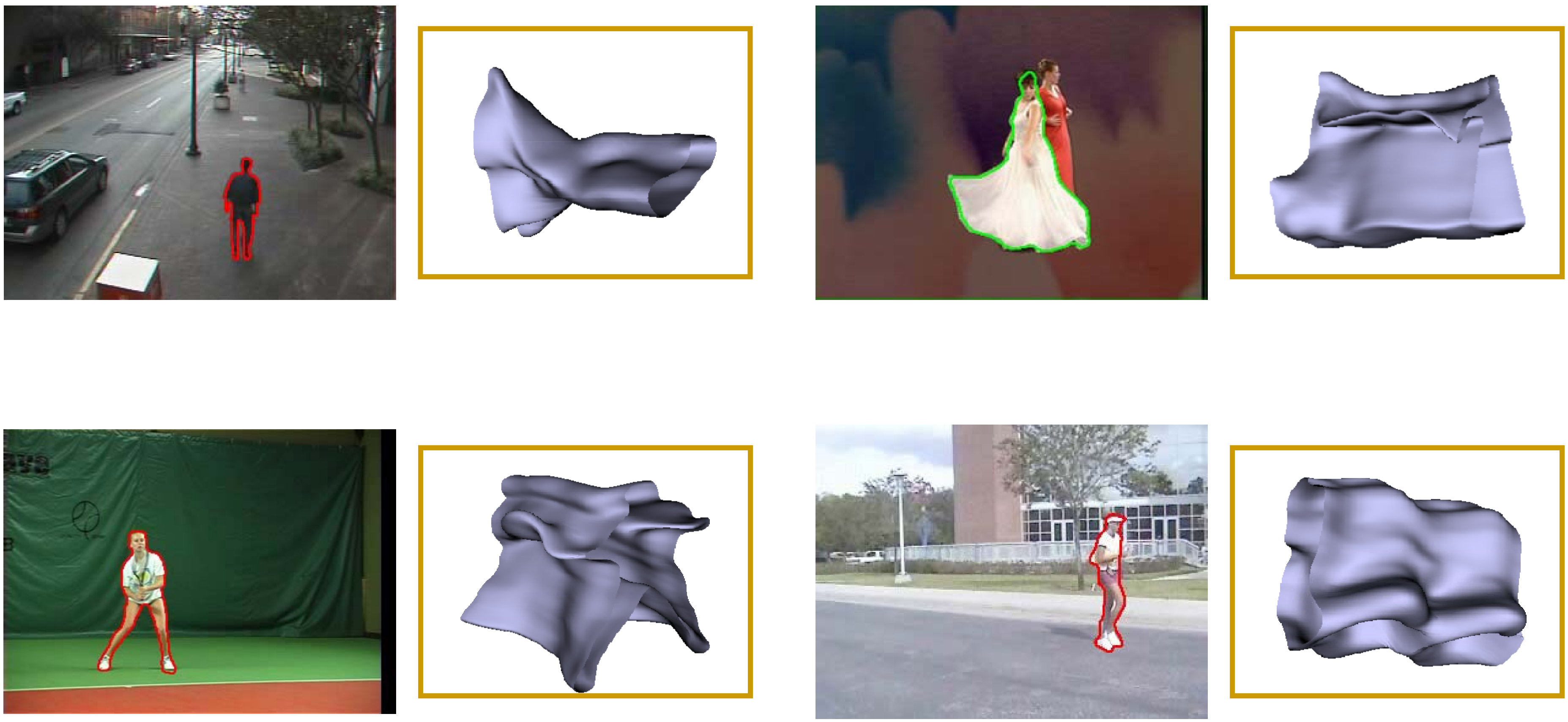
- **Reassign unassigned voxel based on neighboring associations**

Resulting Associations



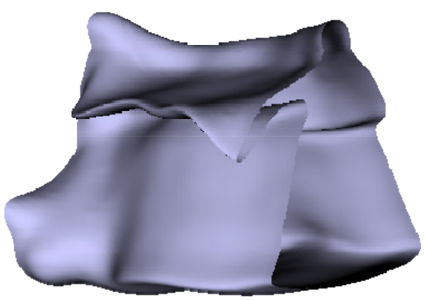


Resulting Volume

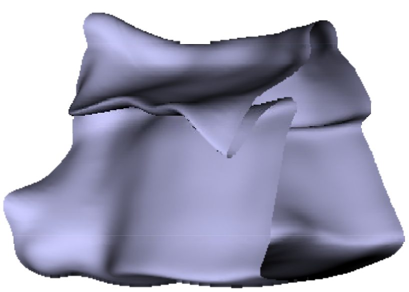


Properties of the Action Volume

- Space-time (3D) object
- Encodes shape and motion
- Uses complete object contours instead of a single point on the object.
- Suitable for fine action analysis
- Continuous representation
  - Same volume for same action of different lengths



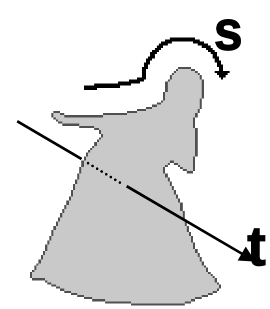
40 frames

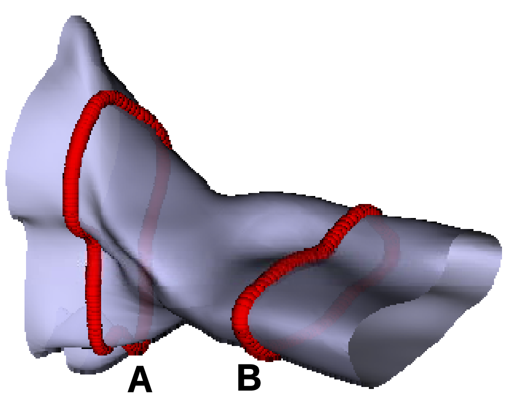


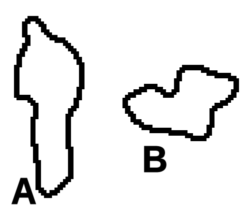
20 random selected frames

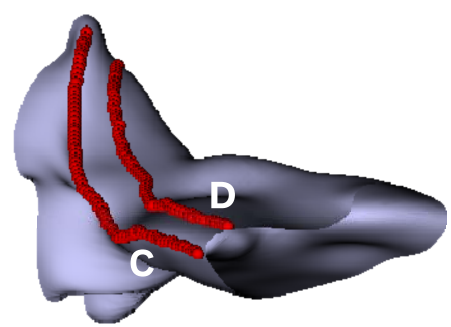
Properties of the Action Volume

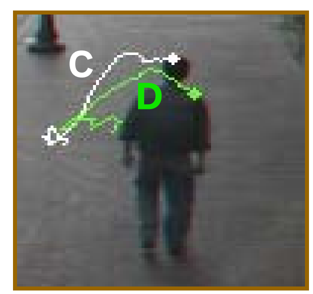
- Can be represented in 2D
  - Arc length and time
- Can regenerate contour at time t
- Can provide spatial trajectory of contour points











What is the Action Sketch?

- Important action descriptors
  - Unique shape and motion characteristics
- Related to differential geometric properties of action volume
  - 1st and 2nd fundamental forms
    - Gaussian and mean curvatures
    - Fundamental surface types

Computing Gaussian (*K*) and Mean (*H*) Curvatures

- K and H are two algebraic invariants of Weingarten mapping S.

$$K = \det(\mathbf{S})$$

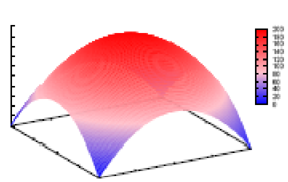
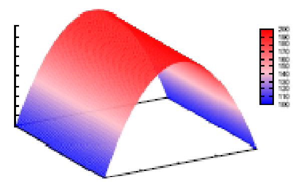
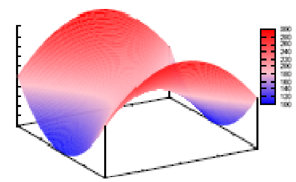
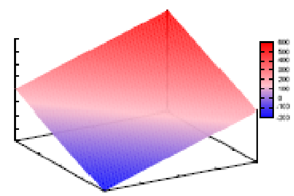
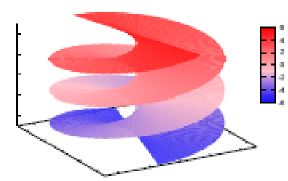
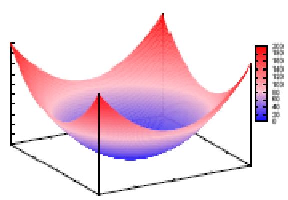
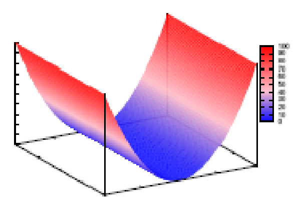
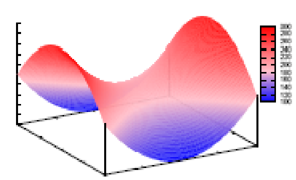
$$H = \frac{1}{2}\,\mathrm{trace}(\mathbf{S})$$

$$\mathbf{S} = \underbrace{\mathbf{g}^{-1}}_{\text{from 1st fundamental f}}\;\underbrace{\mathbf{b}}_{\text{from 2nd fundamental}}$$

$$\mathbf{g} = \begin{bmatrix} f_s f_s & f_s f_t \\ f_s f_t & f_t f_t \end{bmatrix} \qquad \mathbf{b} = \begin{bmatrix} f_{ss}\vec{n} & f_{st}\vec{n} \\ f_{st}\vec{n} & f_{tt}\vec{n} \end{bmatrix}$$

where $f(s,t)$ is a point on the volume, n is normal at f

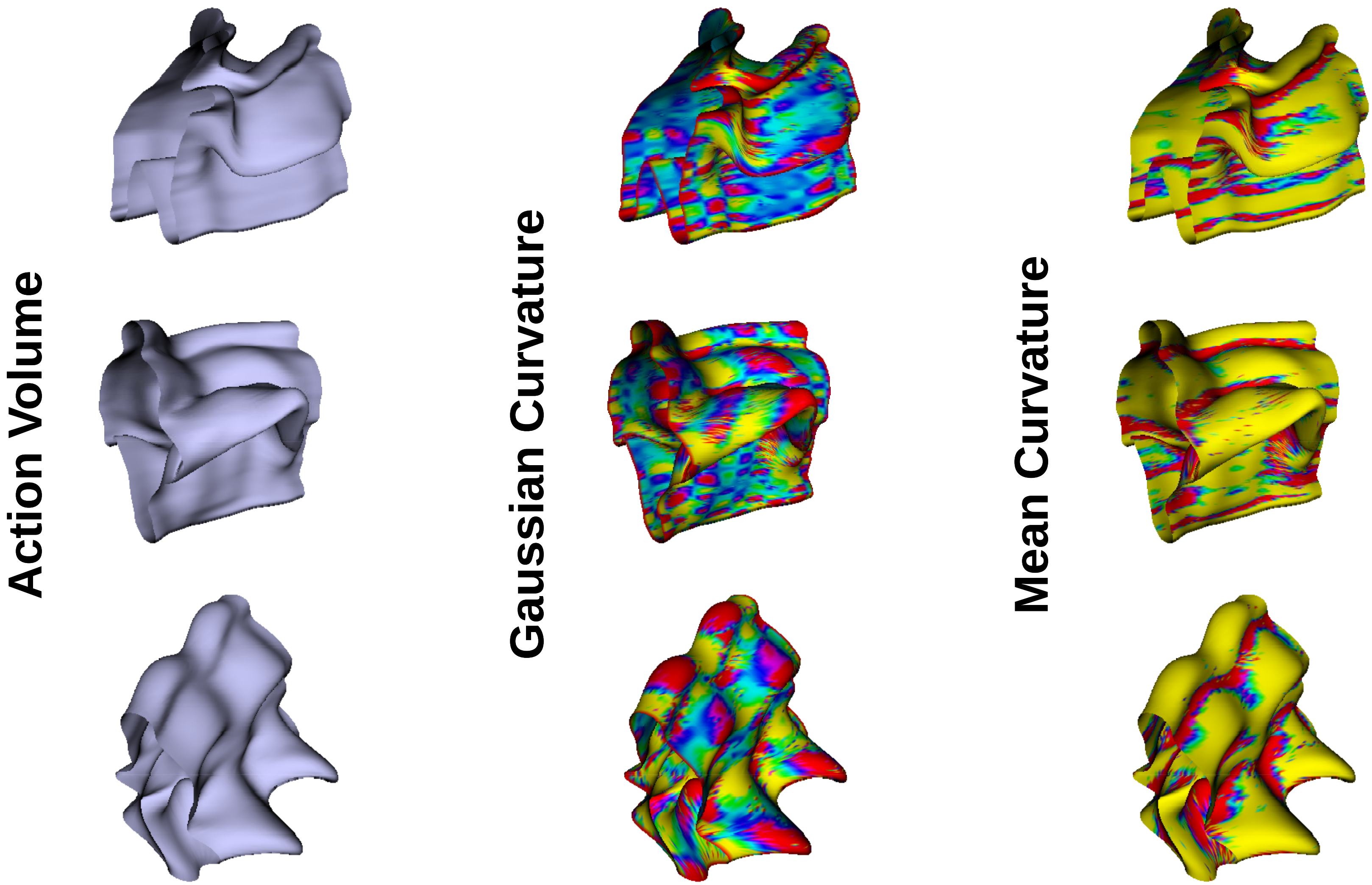
Fundamental Surface Types

	$K > 0$	$K = 0$	$K < 0$
$H < 0$	peak 	ridge 	saddle ridge 
$H = 0$	none	flat 	minimal 
$H > 0$	pit 	valley 	saddle valley 

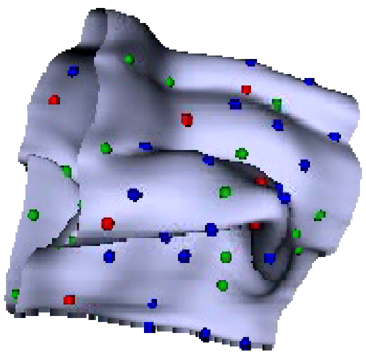
Properties of Surface Types

- Rotation and translation invariant in spatio-temporal space.
- Encodes intrinsic properties of surface.
  - Defines the convexity or concavity of surface.
- Related to speed and acceleration.

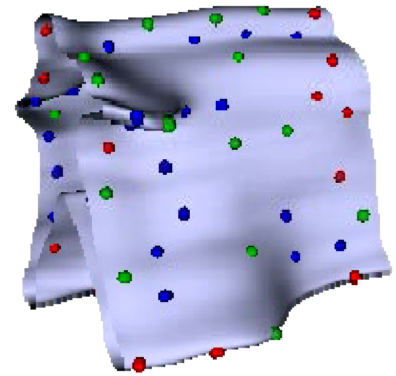
Differential Geometric Surface



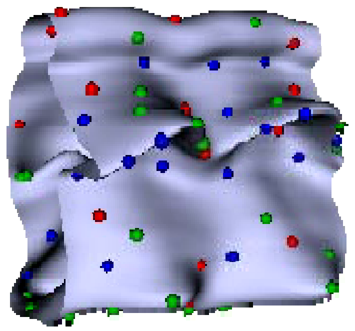
Examples



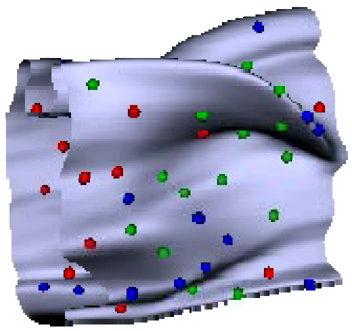
kicking



dance

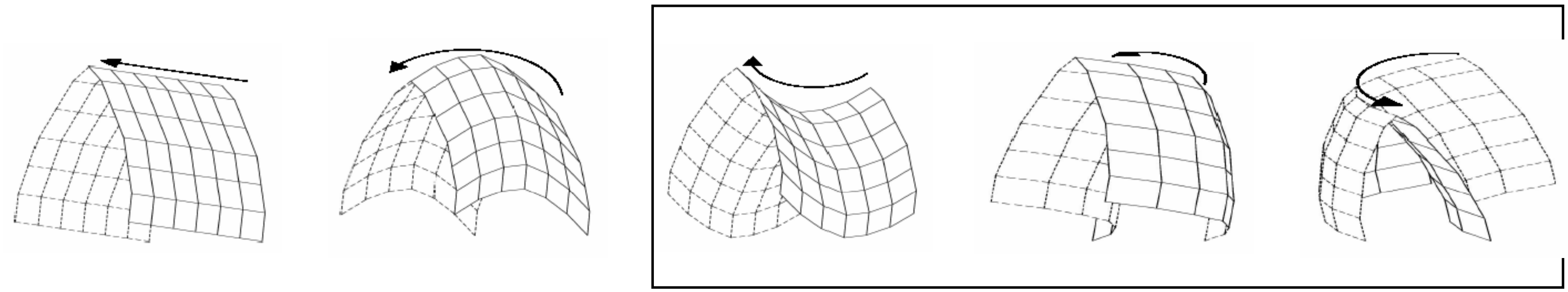


walking



surrender

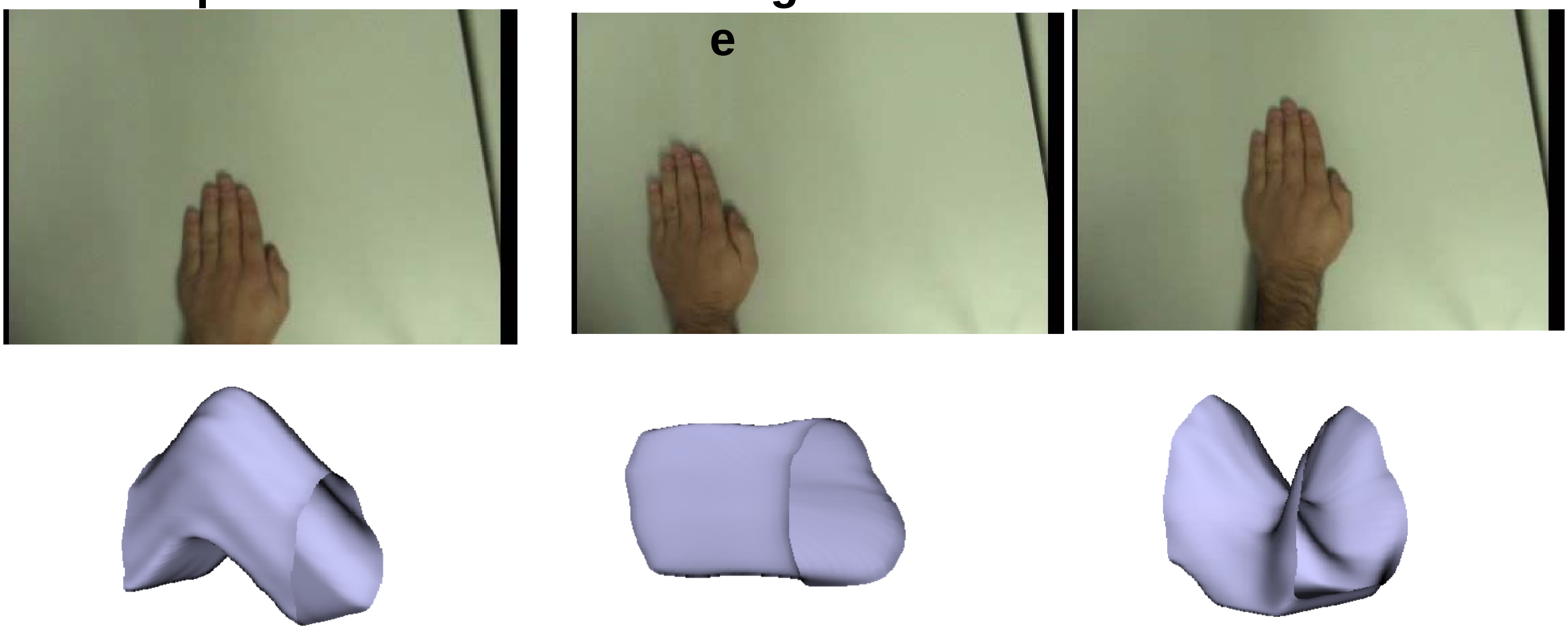
Surface patches & their relation to the object motion



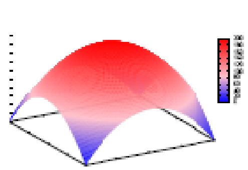
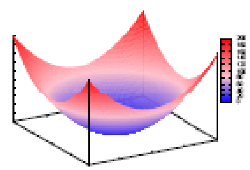
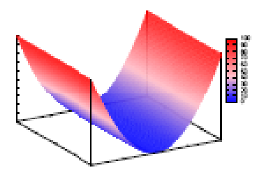
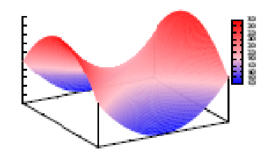
ridge **peak** **saddle ridge**

Action Descriptors Relation to Object Motion

peak | ridg e | saddle

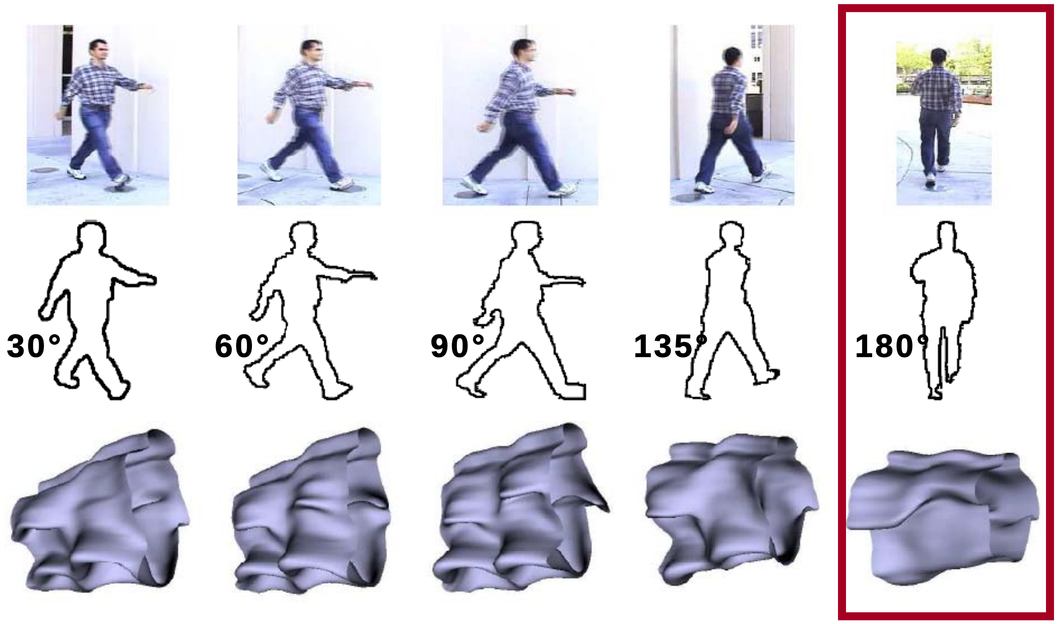


Action Descriptors Relation to Spatial and Trajectory Curvature

	 PEAK	 PIT	 VALLEY	 SADDLE VALLEY
Contour Curvature	maximum	minimum	maximum	maximum
Trajectory Curvature	maximum	minimum	zero	minimum

Changes in viewpoint

- Elements occur on concavities and convexities of contours which are robust to viewpoint changes



Matching Action Volumes

- Epi-polar geometric approach
- Volume registration
- Establishing correspondence
  - Match peaks with peaks, valleys with valleys, etc.

Registration

Level Sets

- Affine transformation

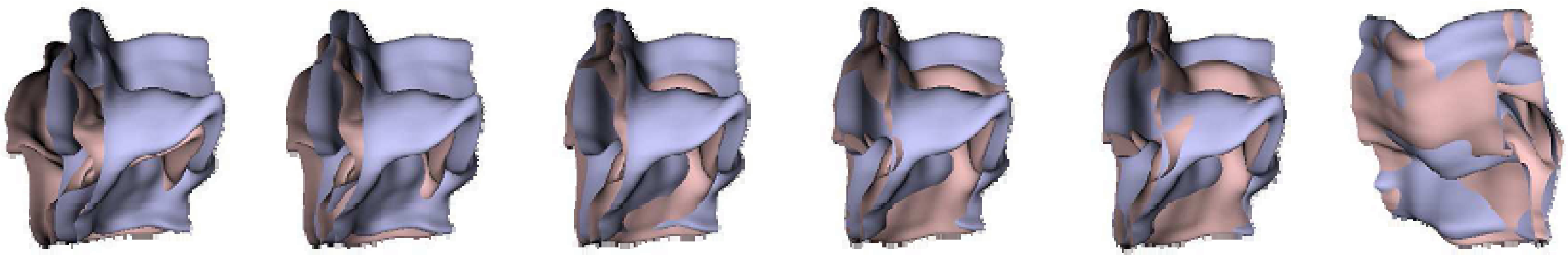
$$\underbrace{\begin{pmatrix} x_{\mathbf{B}} \\ y_{\mathbf{B}} \\ t_{\mathbf{B}} \end{pmatrix}}_{X_{\mathbf{B}}} = \underbrace{\begin{pmatrix} a & b & c \\ d & e & f \\ 0 & 0 & 1 \end{pmatrix}}_{A} \underbrace{\begin{pmatrix} x_{\mathbf{D}_i} \\ y_{\mathbf{D}_i} \\ t_{\mathbf{D}_i} \end{pmatrix}}_{X_{\mathbf{D}_i}} + \underbrace{\begin{pmatrix} t_x \\ t_y \\ 0 \end{pmatrix}}_{T}.$$

- Registration cost

$$E(\phi_D - \phi_S) = \int\int\int_{\Sigma} \left(\phi_D(x,y,t) - \phi_S(x,y,t)\right)^2 ds dt$$

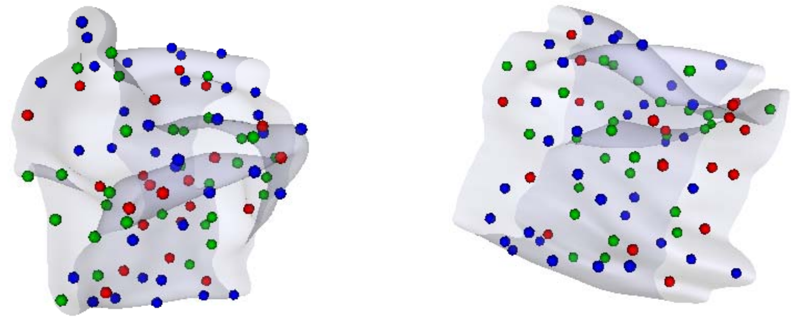
- Speed up by using only zero level set and a random subset of 3D voxels

$$E(\phi_D - \phi_S) = \int\int\int_{\Sigma} \min \parallel \phi_{D0}(x,y,t) - \phi_{S0}(x_i,y_i,t_i) \parallel^2 ds dt$$

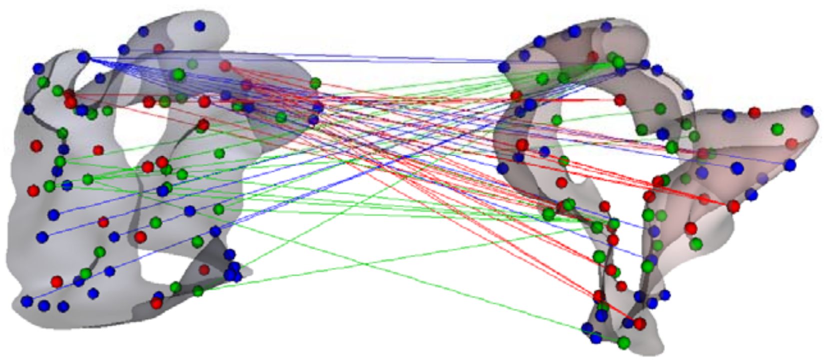


Matching Volumes: Establishing Correspondence

- Generate bipartite action graphs



- Define weights by
  - Space-time proximity
  - Shape similarity



- Find Maximum Matching

Recognition
Epipolar Geometry

- Corresponding points satisfy epipolar geometry

$$\mathbf{x}_B \; f \; \mathbf{x}_{D_i} = 0$$

- Form system of equations

$$\mathbf{A}f = 0$$

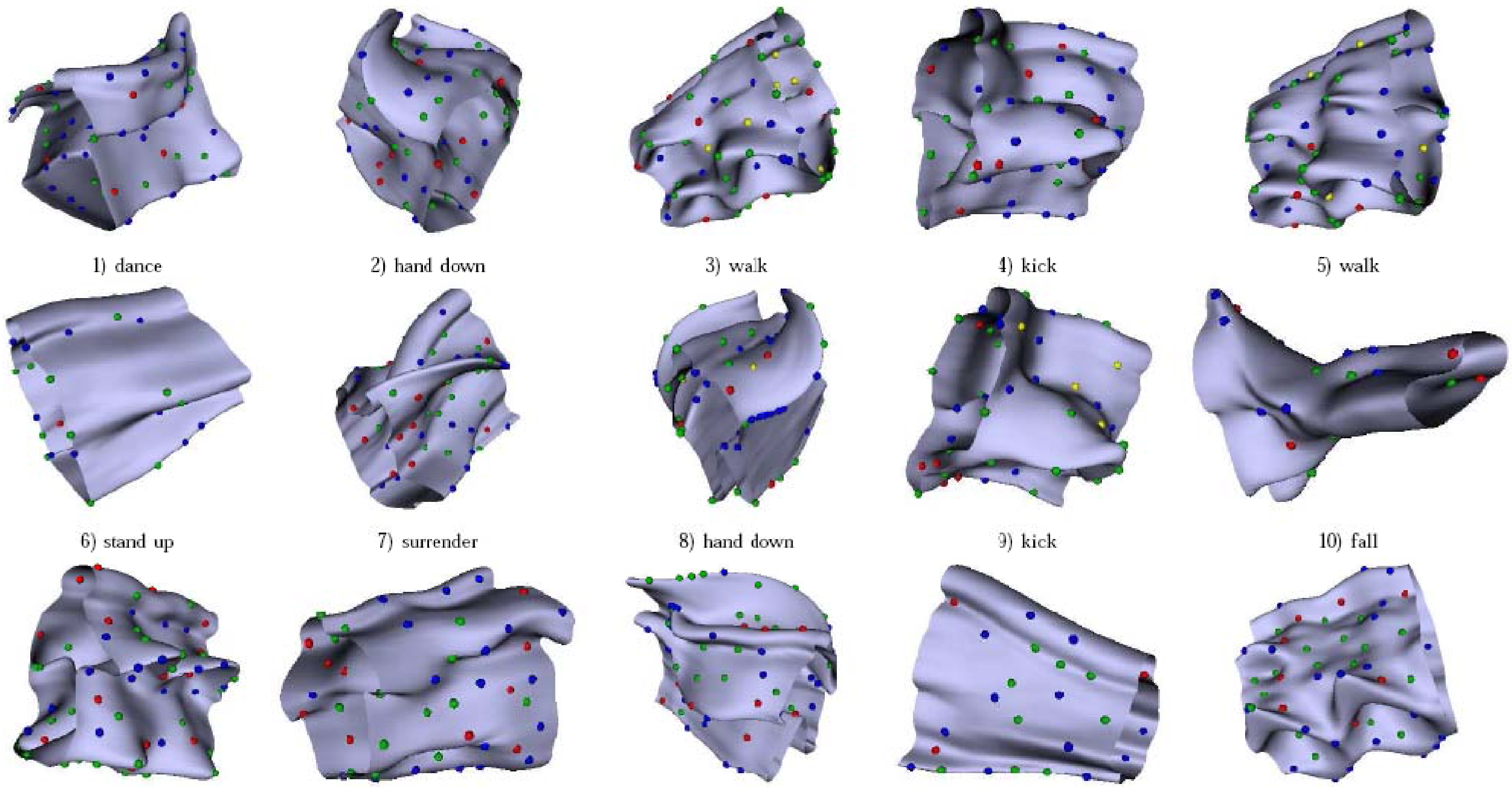
$$\mathbf{A} = [x_{\mathbf{D}_i} x_{\mathbf{B}}, y_{\mathbf{D}_i} x_{\mathbf{B}}, x_{\mathbf{B}}, x_{\mathbf{D}_i} y_{\mathbf{B}}, y_{\mathbf{D}_i} y_{\mathbf{B}}, y_{\mathbf{B}}, x_{\mathbf{D}_i}, y_{\mathbf{D}_i}, 1]$$

$$\mathbf{f} = [\mathcal{F}_{1,1}, \mathcal{F}_{1,2}, \mathcal{F}_{1,3}, \mathcal{F}_{2,1}, \mathcal{F}_{2,2}, \mathcal{F}_{2,3}, \mathcal{F}_{3,1}, \mathcal{F}_{3,2}, \mathcal{F}_{3,3}]$$

- Compute quality from cumulative symmetric epipolar distance

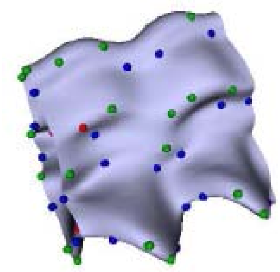
$$d(X_{\mathbf{D}_i}, X_{\mathbf{B}}) = \sqrt{\left(\frac{X_{\mathbf{D}_i}^{\top} U_{\mathbf{B}}}{|U_{\mathbf{B}}|}\right)^2 + \left(\frac{X_{\mathbf{B}}^{\top} U_{\mathbf{D}_i}}{|U_{\mathbf{D}_i}|}\right)^2}$$

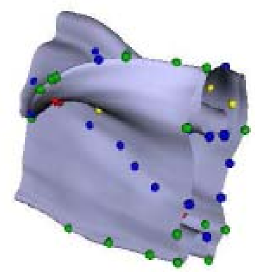
Action Volumes

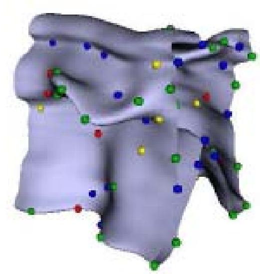


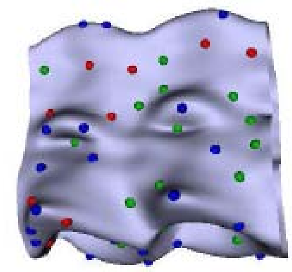
1) dance 2) hand down 3) walk 4) kick 5) walk
6) stand up 7) surrender 8) hand down 9) kick 10) fall
11) walk 12) walk 13) aerobic 1 14) sit down 15) walk

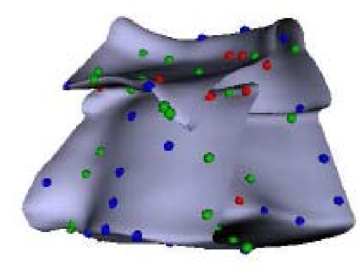
Action Volumes

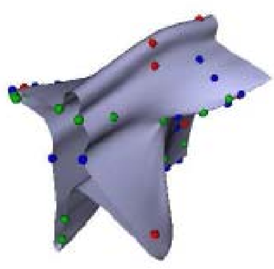

16) running

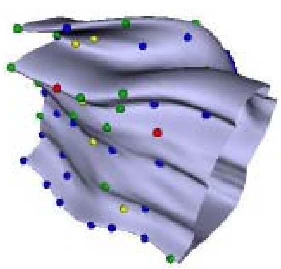

17) surrender

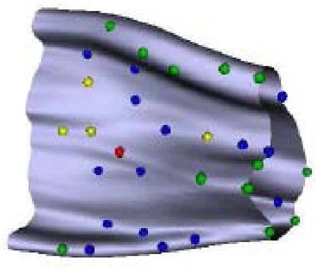

18) stroke

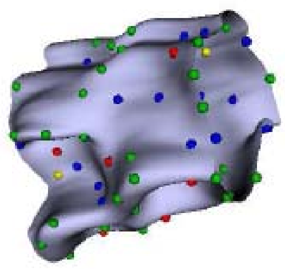

19) walk

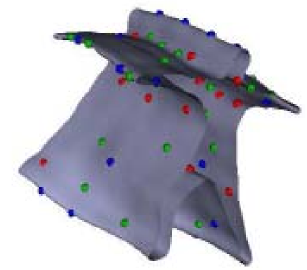

20) dance

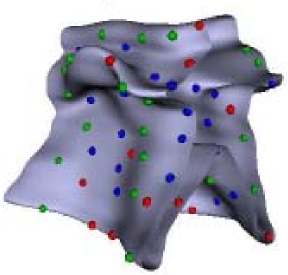

21) aerobic 2

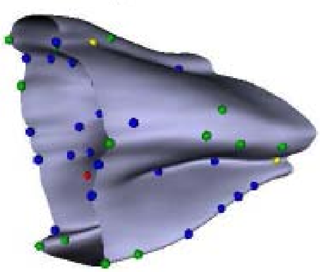

22) aerobic 3

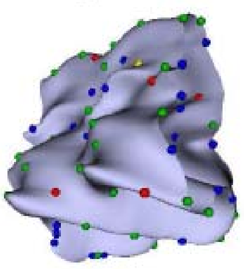

23) sit down

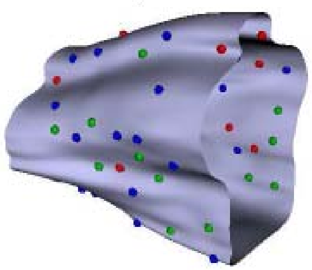

24) walk

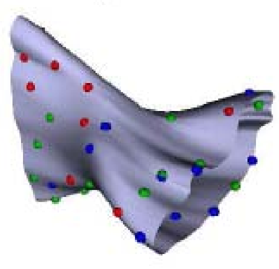

25) aerobic 4


26) stroke


27) stand up


28) running


29) stand up


30) falling

Recognition Results

Input Action	#	Matching action	#
Dance	1	*Dance*	20
Hand down	2	**Stand up**	29
Walking	3	*Walking*	11
Kicking	4	*Kicking*	9
Walking	5	*Walking*	11
Stand up	6	*Stand up*	29
Surrender	7	*Surrender*	17
Hands down	8	*Hands down*	82
Kicking	9	*Kicking*	4
Falling	10	*Falling*	30
Walking	11	*Walking*	11
Walking	12	**Sit down**	23
Sit down	14	*Sit down*	23

Video	#	Matching action	#
Walking	15	*Walking*	11
Running	16	*Running*	28
Surrender	17	*Surrender*	17
Tennis stroke	18	*Tennis stroke*	26
Walking	19	*Walking*	11
Dance	20	*Dance*	1
Sit down	23	*Sit down*	23
Walking	24	*Walking*	11
Tennis stroke	26	*Tennis stroke*	18
Stand up	27	*Stand up*	29
Running	28	*Running*	16
Stand up	29	**Hands down**	8
Falling	30	*Falling*	10

Today

- Histograms of edges (Schiele)
- Windowed spectral analysis (GIST)
- Tiled histograms of edges (HOG)

- Motion History Images (Bobick)
- Rectified Flow Descriptors (Efros)
- Differential Geometry Signatures (Shah)

Feb 10th – Local features (SIFT, Surf, MSER, Shape Context, Self Similarity, etc.)

- T. Lindeberg, "Feature detection with automatic scale selection," International Journal of Computer Vision, vol. 30, no. 2, pp. 79-116, November 1998. Available: http://dx.doi.org/10.1023/A:1008045108935
- J. Matas, O. Chum, U. Martin, and T. Pajdla, "Robust wide baseline stereo from maximally stable extremal regions," in Proceedings of British Machine Vision Conference, vol. 1, London, 2002, pp. 384-393. Available: http://citeseer.ist.psu.edu/608213.html
- K. Mikolajczyk and C. Schmid, "Scale & affine invariant interest point detectors," Int. J. Comput. Vision, vol. 60, no. 1, pp. 63-86, October 2004. Available: http://dx.doi.org/10.1023/B:VISI.0000027790.02288.f2
- I. Laptev, "On space-time interest points," International Journal of Computer Vision, vol. 64, no. 2-3, pp. 107-123, September 2005. Available: http://dx.doi.org/10.1007/s11263-005-1838-7

Optional Readings:

- E. Shechtman and M. Irani, "Matching local self-similarities across images and videos," in Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on, 2007, pp. 1-8. Available: http://dx.doi.org/10.1109/CVPR.2007.383198
- **H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded-up robust features," in *9th European Conference on Computer Vision*, Graz, Austria. Available: http://www.vision.ee.ethz.ch/~surf/eccv06.pdf**

Reminder

Please sign up via email for a paper that you would like to present or show a demonstration of.

- can show demos next week from this week's papers (e.g.,GIST / spatial envelope on some images collected around campus)

- but otherwise should show demo on day of paper (could show Laptev or self-similarity features on Berkeleyish action examples next week…)

I'll expect two demos or one presentation per person taking the course for credit…

N.B., a demo is more than showing author's videos or canned matlab example…must try on something new or extend…