CS294-43: Visual Object and Activity Recognition

Prof. Trevor Darrell

Feb 3\textsuperscript{rd}: 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

Slide credit: K. Grauman, B. Leibe
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

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

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

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
Slide credit: K. Grauman, B. Leibe
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)
• 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)
•...
Histograms of oriented gradients

- SIFT, D. Lowe, ICCV 1999
- Shape context

- Dalal & Trigs, 2006

Slide credit: A. Torralba
Histograms of Gradients, ca. 1996

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.

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


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 classification 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 [21, 8, 10]. 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-line learning, several researchers have adapted such...
recognition system

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 vector 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

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

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

Slide Credit: Torralba, Olivia, J. Huang
Spatially Localized Statistics

- Windowed FFT

<table>
<thead>
<tr>
<th></th>
<th>1 m</th>
<th>10 m</th>
<th>100 m</th>
<th>1000 m</th>
<th>10000 m</th>
</tr>
</thead>
<tbody>
<tr>
<td>Man-made environments</td>
<td>![Symbols]</td>
<td>![Symbols]</td>
<td>![Symbols]</td>
<td>![Symbols]</td>
<td>![Symbols]</td>
</tr>
<tr>
<td>Natural environments</td>
<td>![Symbols]</td>
<td>![Symbols]</td>
<td>![Symbols]</td>
<td>![Symbols]</td>
<td>![Symbols]</td>
</tr>
</tbody>
</table>

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 Spatial Envelope

Aude Oliva

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

**We use a classification task:** 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

Task: 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

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

Perspective - manmade vs. natural structure

Far vs. less far
Task: In the third step, participants split the 4 groups in two more groups.

- Open vs. closed
- Flat vs. oblique structure
- Perspective
- Far vs. near
- Far vs. less far
- manmade vs. natural structure
- Fine vs. coarse texture
# Perceptual Dimensions

<table>
<thead>
<tr>
<th>Dimensions</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naturalness</td>
<td>77</td>
</tr>
<tr>
<td>Openness</td>
<td>83</td>
</tr>
<tr>
<td>Perspective</td>
<td>53</td>
</tr>
<tr>
<td>Size (roughness)</td>
<td>47</td>
</tr>
<tr>
<td>Diagonal planes</td>
<td>41</td>
</tr>
<tr>
<td>Depth</td>
<td>59</td>
</tr>
<tr>
<td>Symmetry</td>
<td>29</td>
</tr>
<tr>
<td>Verticalness</td>
<td>18</td>
</tr>
</tbody>
</table>

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.

**Naturalness:** principal structure of building blocks

**Openness:** the sense of enclosure of the space

**Expansion:** the perspective

**Mean depth:** the scale of the space

**Ruggedness:** deviation of the ground plane

**Roughness:** size of the building blocks

---

Slide Credit: Olivia
Global Properties: Structure of space

Mean depth

Small volume → large volume

Openness

Expansion

Oliva & Torralba, 2001

Slide Credit: Olivia
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  

\[ V_G \quad V_G \quad V_G \]  

\rightarrow \quad \text{closed scenes}  

\[ V_G \quad V_G \]

• 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**

**Method:** **Learning stage:** 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 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).

\[
A(f_x, f_y) = \sum_{f_x, f_y} \text{FT} \sum_{f_x, f_y} \text{DST}(f_x, f_y) = \sum_{f_x, f_y}
\]

Slide Credit: Olivia
Estimation of Space descriptor

- Spatial envelope properties are *continuous* perceptual dimensions

From open scenes to ....

<table>
<thead>
<tr>
<th>Original images</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Image" /></td>
</tr>
<tr>
<td><img src="image2" alt="Image" /></td>
</tr>
<tr>
<td><img src="image3" alt="Image" /></td>
</tr>
<tr>
<td><img src="image4" alt="Image" /></td>
</tr>
<tr>
<td><img src="image5" alt="Image" /></td>
</tr>
<tr>
<td><img src="image6" alt="Image" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Power spectra $A(f_x, f_y)^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image7" alt="Image" /></td>
</tr>
<tr>
<td><img src="image8" alt="Image" /></td>
</tr>
<tr>
<td><img src="image9" alt="Image" /></td>
</tr>
<tr>
<td><img src="image10" alt="Image" /></td>
</tr>
<tr>
<td><img src="image11" alt="Image" /></td>
</tr>
<tr>
<td><img src="image12" alt="Image" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Product $\text{DST}(f_x, f_y) A(f_x, f_y)^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image13" alt="Image" /></td>
</tr>
<tr>
<td><img src="image14" alt="Image" /></td>
</tr>
<tr>
<td><img src="image15" alt="Image" /></td>
</tr>
<tr>
<td><img src="image16" alt="Image" /></td>
</tr>
<tr>
<td><img src="image17" alt="Image" /></td>
</tr>
<tr>
<td><img src="image18" alt="Image" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Opponent energy image</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image19" alt="Image" /></td>
</tr>
<tr>
<td><img src="image20" alt="Image" /></td>
</tr>
<tr>
<td><img src="image21" alt="Image" /></td>
</tr>
<tr>
<td><img src="image22" alt="Image" /></td>
</tr>
<tr>
<td><img src="image23" alt="Image" /></td>
</tr>
<tr>
<td><img src="image24" alt="Image" /></td>
</tr>
</tbody>
</table>

Dark: *Openness* features

Light: *closeness* feature

Slide Credit: Olivia
Windowed Discriminant Template

Spectrogram $\sum_{f_x,f_y} \mathcal{F} \star$ WDST $= \sum_{f_x,f_y}$

< 0 Expanded

> 0 Flat

Expansion

Flat

Slide Credit: Olivia
Space Properties of the *Content* of the scene

**DST**

**WDST**

**Naturalness**

**Roughness** (Natural)

**Roughness** (Manmade)

Stationary distribution of features

Slide Credit: Olivia
Naturalness descriptor

Manmade environments

Natural environments

93%

Center of the axis

Errors

Slide Credit: Olivia
Space Properties of the *Shape* of the scene

- **Openness** (Natural)
  - DST
  - WDST

- **Openness** (Manmade)
  - DST
  - WDST

- **Ruggedness**
  - DST
  - WDST

- **Expansion**
  - DST
  - WDST

*Non-stationary distribution of features*

Slide Credit: Olivia
Spatial Envelope Theory of Scene Recognition

Global Scene Structure

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

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

Slide Credit: Olivia
Modeling Scene Gist

Scenes from the same category share similar global properties.
Categorization of Manmade Scenes

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

<table>
<thead>
<tr>
<th></th>
<th>Highway</th>
<th>Street</th>
<th>City centre</th>
<th>tall building</th>
</tr>
</thead>
<tbody>
<tr>
<td>Highway</td>
<td>91.6</td>
<td>4.8</td>
<td>2.7</td>
<td>0.9</td>
</tr>
<tr>
<td>Street</td>
<td>4.7</td>
<td>89.6</td>
<td>1.8</td>
<td>3.4</td>
</tr>
<tr>
<td>Centre</td>
<td>2.5</td>
<td>2.3</td>
<td>87.8</td>
<td>7.4</td>
</tr>
<tr>
<td>Tall Building</td>
<td>0.1</td>
<td>3.4</td>
<td>8.5</td>
<td>88</td>
</tr>
</tbody>
</table>

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)

<table>
<thead>
<tr>
<th></th>
<th>Coast</th>
<th>Countryside</th>
<th>Forest</th>
<th>Mountain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coast</td>
<td>88.6</td>
<td>8.9</td>
<td>1.2</td>
<td>1.3</td>
</tr>
<tr>
<td>Countryside</td>
<td>9.8</td>
<td>85.2</td>
<td>3.7</td>
<td>1.3</td>
</tr>
<tr>
<td>Forest</td>
<td>0.4</td>
<td>3.6</td>
<td>91.5</td>
<td>4.5</td>
</tr>
<tr>
<td>Mountain</td>
<td>0.4</td>
<td>4.6</td>
<td>3.8</td>
<td>91.2</td>
</tr>
</tbody>
</table>

Local organization:
correct for 92 % images
(4 similar images on 7 K-NN)

Slide Credit: Olivia
Representing Image Structure

\[ V = \{\text{energy at each orientation and scale}\} = 6 \times 4 \text{ dimensions} \]

80 features

Vector of Global features

\[ V_t \xrightarrow{} \text{PCA} \xrightarrow{} V_G \]

Scene Recognition via texture surface
Scene Classification from “Texture”

Slide Credit: Olivia
“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?

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)
Examples (Natural)

Slide Credit: Torralba, Olivia, J. Huang
Some Results

Slide Credit: Torralba, Olivia, J. Huang
$f(D | \text{category})$

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**

- Scenes with animals
- Scenes with cars
- Scenes with people
- Scenes with far people
- Scenes with near people
- Scenes with close-up people

**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


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
• Tested with
  – RGB
  – LAB
  – Grayscale

• Gamma Normalization and Compression
  – Square root
  – Log

Slide credit: Dalal, Triggs, P. Barnum
- Histogram of gradient orientations
  - Orientation
  - Position

  – Weighted by magnitude

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

R-HOG

Cell

Block

C-HOG

Center Bin

Radial Bins, Angular Bins

Slide credit: Dalal, Triggs, P. Barnum
R-HOG

Cell

Block

C-HOG

Center Bin

Block

Radial Bins, Angular Bins

\[
\begin{align*}
L1 - norm : v & \rightarrow v/(\|v\|_1 + \epsilon) \\
L1 - sqrt : v & \rightarrow \sqrt{v/(\|v\|_1 + \epsilon)} \\
L2 - norm : v & \rightarrow v/\sqrt{\|v\|_2^2 + \epsilon^2} \\
L2 - hys : L2-norm, plus clipping at .2 and renomalizing
\end{align*}
\]

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")
Strict Appearance: human movements

- Is recognizing movement a 3D or 2D problem? Simple human psychophysics 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°
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:

\[
\begin{align*}
&\text{if moving } I_\tau(x,y,t) = \tau \\
&\text{otherwise } \\
&\quad I_\tau(x,y,t) = \max(I_\tau(x,y,t-1) - 1, 0)
\end{align*}
\]

- 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

- $MEI + MHI = \text{Temporal template}$
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 $\tau$ 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.
- Sometimes, possible because the machine knows the context.
- A kinder, gentler $C^3$I 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
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

Challenge: how to compare motions?

Slide credit: Malik
Motion Descriptor

Image frame

Optical flow

blurred

Slide credit: Malik
Comparing motion descriptors

\[ \sum \]

frame-to-frame similarity matrix

blurry I

motion-to-motion similarity matrix

Slide credit: Malik
Classifying Ballet Actions

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

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

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

- Contour Extraction
  - Graph theoretic volume generation
  - Volume smoothing

- Action Volume Generation
  - Graph theoretic volume generation
  - Volume smoothing

- Feature Extraction & Recognition
  - Differential geometry
  - Epi-polar geometry

2006 Alper Yilmaz, PhD
Volume Generation

- Contours from a contour tracker

2006

Alper Yilmaz, PhD
Volume Generation

- Two pass correspondence approach

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

- Association likelihood
  - Shape similarity
  - Proximity
  - Orientation similarity

2006 Alper Yilmaz, PhD
Volume Generation

- Proximity
  \[ d_{i,j} = \|c_i - c_j\|_2 \]

- Alignment similarity
  \[ \alpha_{i,j} = \arccos \frac{\vec{n}_i \cdot \vec{n}_j}{|\vec{n}_i||\vec{n}_j|} \]

- Shape similarity
  \[ \epsilon_{i,j} = \sum_{x_j \in N_j} \|\hat{x}_i - x_j\|_2 \]

\[ l_{i,j} = \exp\left(-\frac{d_{i,j}^2}{\sigma_d^2}\right) \exp\left(-\frac{\alpha_{i,j}^2}{\sigma_\alpha^2}\right) \exp\left(-\frac{\epsilon_{i,j}^2}{\sigma_\epsilon^2}\right) \]
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

2006 Alper Yilmaz, PhD
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

2006 Alper Yilmaz, PhD
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

2006 Alper Yilmaz, PhD
Computing Gaussian \((K)\) and Mean \((H)\) Curvatures

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

\[
K = \det(S) \\
H = \frac{1}{2} \text{trace}(S)
\]

\[
g = \begin{bmatrix} f_s f_s & f_s f_t \\ f_s f_t & f_t f_t \end{bmatrix} \quad b = \begin{bmatrix} f_{ss} \vec{n} & f_{st} \vec{n} \\ f_{st} \vec{n} & f_{tt} \vec{n} \end{bmatrix}
\]

where \(f(s,t)\) is a point on the volume, \(n\) is normal at \(f\)
# Fundamental Surface Types

<table>
<thead>
<tr>
<th>$K &gt; 0$</th>
<th>$K = 0$</th>
<th>$K &lt; 0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H &lt; 0$</td>
<td>peak</td>
<td>ridge</td>
</tr>
<tr>
<td>$H = 0$</td>
<td>none</td>
<td>flat</td>
</tr>
<tr>
<td>$H &gt; 0$</td>
<td>pit</td>
<td>valley</td>
</tr>
</tbody>
</table>
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

Action Volume
Gaussian Curvature
Mean Curvature

2006
Alper Yilmaz, PhD
Examples

- kicking
- dance
- walking
- surrender
Surface patches & their relation to the object motion

ridge  peak  saddle ridge
Action Descriptors Relation to Object Motion

peak

ridge

saddle
### Action Descriptors Relation to Spatial and Trajectory Curvature

<table>
<thead>
<tr>
<th></th>
<th>PEAK</th>
<th>PIT</th>
<th>VALLEY</th>
<th>SADDLE VALLEY</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contour Curvature</td>
<td>maximum</td>
<td>minimum</td>
<td>maximum</td>
<td>maximum</td>
</tr>
<tr>
<td>Trajectory Curvature</td>
<td>maximum</td>
<td>minimum</td>
<td>zero</td>
<td>minimum</td>
</tr>
</tbody>
</table>
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

\[
\begin{pmatrix}
x_B \\
y_B \\
t_B
\end{pmatrix} =
\begin{pmatrix}
a & b & c \\
d & e & f \\
0 & 0 & 1
\end{pmatrix}
\begin{pmatrix}
x_{D_i} \\
y_{D_i} \\
t_{D_i}
\end{pmatrix} +
\begin{pmatrix}
t_x \\
t_y \\
0
\end{pmatrix}
\]

- Registration cost

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

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

\[
E(\phi_D, \phi_S) = \int \int \int_{\Sigma} \min \| \phi_{D_0}(x, y, t) - \phi_{S_0}(x_i, y_i, t_i) \|^2 \, ds \, dt
\]
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
  \[ x_B f x_{D_i} = 0 \]

- Form system of equations
  \[ A f = 0 \]
  \[ A = \begin{bmatrix} x_{D_i} & x_B & y_{D_i} & y_B & x_{D_i} & y_{D_i} & y_B & x_{D_i} & y_{D_i} & 1 \end{bmatrix} \]
  \[ f = [F_{1,1}, F_{1,2}, F_{1,3}, F_{2,1}, F_{2,2}, F_{2,3}, F_{3,1}, F_{3,2}, F_{3,3}] \]

- Compute quality from cumulative symmetric epipolar distance
  \[ d(X_{D_i}, X_B) = \sqrt{\left( \frac{X_{D_i}^T U_B}{|U_B|} \right)^2 + \left( \frac{X_B^T U_{D_i}}{|U_{D_i}|} \right)^2} \]
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

2006
Alper Yilmaz, PhD
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

<table>
<thead>
<tr>
<th>Input Action</th>
<th>#</th>
<th>Matching action</th>
<th>#</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dance</td>
<td>1</td>
<td>Dance</td>
<td>20</td>
</tr>
<tr>
<td>Hand down</td>
<td>2</td>
<td>Stand up</td>
<td>29</td>
</tr>
<tr>
<td>Walking</td>
<td>3</td>
<td>Walking</td>
<td>11</td>
</tr>
<tr>
<td>Kicking</td>
<td>4</td>
<td>Kicking</td>
<td>9</td>
</tr>
<tr>
<td>Walking</td>
<td>5</td>
<td>Walking</td>
<td>11</td>
</tr>
<tr>
<td>Stand up</td>
<td>6</td>
<td>Stand up</td>
<td>29</td>
</tr>
<tr>
<td>Surrender</td>
<td>7</td>
<td>Surrender</td>
<td>17</td>
</tr>
<tr>
<td>Hands down</td>
<td>8</td>
<td>Hands down</td>
<td>82</td>
</tr>
<tr>
<td>Kicking</td>
<td>9</td>
<td>Kicking</td>
<td>4</td>
</tr>
<tr>
<td>Falling</td>
<td>10</td>
<td>Falling</td>
<td>30</td>
</tr>
<tr>
<td>Walking</td>
<td>11</td>
<td>Walking</td>
<td>11</td>
</tr>
<tr>
<td>Walking</td>
<td>12</td>
<td>Sit down</td>
<td>23</td>
</tr>
<tr>
<td>Sit down</td>
<td>14</td>
<td>Sit down</td>
<td>23</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Video</th>
<th>#</th>
<th>Matching action</th>
<th>#</th>
</tr>
</thead>
<tbody>
<tr>
<td>Walking</td>
<td>15</td>
<td>Walking</td>
<td>11</td>
</tr>
<tr>
<td>Running</td>
<td>16</td>
<td>Running</td>
<td>28</td>
</tr>
<tr>
<td>Surrender</td>
<td>17</td>
<td>Surrender</td>
<td>17</td>
</tr>
<tr>
<td>Tennis stroke</td>
<td>18</td>
<td>Tennis stroke</td>
<td>26</td>
</tr>
<tr>
<td>Walking</td>
<td>19</td>
<td>Walking</td>
<td>11</td>
</tr>
<tr>
<td>Dance</td>
<td>20</td>
<td>Dance</td>
<td>1</td>
</tr>
<tr>
<td>Sit down</td>
<td>23</td>
<td>Sit down</td>
<td>23</td>
</tr>
<tr>
<td>Walking</td>
<td>24</td>
<td>Walking</td>
<td>11</td>
</tr>
<tr>
<td>Tennis stroke</td>
<td>26</td>
<td>Tennis stroke</td>
<td>18</td>
</tr>
<tr>
<td>Stand up</td>
<td>27</td>
<td>Stand up</td>
<td>29</td>
</tr>
<tr>
<td>Running</td>
<td>28</td>
<td>Running</td>
<td>16</td>
</tr>
<tr>
<td>Stand up</td>
<td>29</td>
<td>Hands down</td>
<td>8</td>
</tr>
<tr>
<td>Falling</td>
<td>30</td>
<td>Falling</td>
<td>10</td>
</tr>
</tbody>
</table>
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.)


Optional Readings:


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…