6.866 projects

Proposals to us by today. We will ok them by Oct. 31.

3 possible project types:
- Original implementation of an existing algorithm.
- Rigorous evaluation of existing algorithm.
- Synthesis or comparison of several research papers.
<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>10/17</td>
<td>Bayesian Analysis</td>
<td></td>
<td></td>
<td>Freeman Slides</td>
</tr>
<tr>
<td>13</td>
<td>10/22</td>
<td>Optic Flow and Direct SFM</td>
<td>Req: H 12, 17</td>
<td>Exam #1 Due</td>
<td>Darrell Slides</td>
</tr>
<tr>
<td>14</td>
<td>10/24</td>
<td>Affine Reconstruction</td>
<td>Req: FP 12</td>
<td>Pset #3 Assigned</td>
<td>Darrell Slides</td>
</tr>
<tr>
<td>15</td>
<td>10/29</td>
<td>TBD</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>10/31</td>
<td>Statistical Classifiers I</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>11/5</td>
<td>Statistical Classifiers II</td>
<td></td>
<td>Pset #3 Due</td>
<td></td>
</tr>
<tr>
<td>Date</td>
<td>Date</td>
<td>Topic</td>
<td>Notes</td>
<td>Slides</td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>-------</td>
<td>--------------------------------------------</td>
<td>----------------------------</td>
<td>-----------------</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>10/17</td>
<td>Bayesian Analysis</td>
<td></td>
<td>Freeman Slides</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>10/22</td>
<td>Optic Flow and Direct SFM</td>
<td>Req: H 12, 17</td>
<td>Darrell Slides</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>10/24</td>
<td>Affine Reconstruction</td>
<td>Req: FP 12</td>
<td>Darrell Slides</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>10/29</td>
<td><strong>Cameras looking at people:</strong> cheap tricks using the tools we've seen so far.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>10/31</td>
<td>Statistical Classifiers I</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>11/5</td>
<td>Statistical Classifiers II</td>
<td></td>
<td>Pset #3 Due</td>
<td></td>
</tr>
</tbody>
</table>

Pset #3 Assigned
<table>
<thead>
<tr>
<th>Date</th>
<th>Date</th>
<th>Topic</th>
<th>Requirements</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>10/17</td>
<td>Bayesian Analysis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>10/22</td>
<td>Optic Flow and Direct SFM</td>
<td>Req: H 12, 17</td>
<td>Exam #1 Due</td>
</tr>
<tr>
<td>14</td>
<td>10/24</td>
<td>Affine Reconstruction</td>
<td>Req: FP 12</td>
<td>Pset #3 Assigned</td>
</tr>
<tr>
<td>15</td>
<td>10/29</td>
<td>Cameras looking at people: cheap tricks using the tools we've seen so far.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>10/31</td>
<td>Statistical Classifiers I</td>
<td>(for faces)</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>11/5</td>
<td>Statistical Classifiers II</td>
<td></td>
<td>Pset #3 Due</td>
</tr>
</tbody>
</table>

Freeman Slides

Darrell Slides
Does anyone mind...

*If I use your photographed face for a simple face-detection demo program that we’ll run in class next time?*

*If you do mind, please let me know (before Thursday).*
Today: Cameras looking at people

A mini-application lecture: under controlled conditions (not general conditions), what human interaction applications can you build with the tools we’ve developed so far?
To be compared with: more sophisticated detection, classification, and tracking tools that we’ll study over the rest of the course.
Yesterday’s tomorrow

New York Worlds Fair, 1939
(Westinghouse Historical Collection)
Computer vision still needs to become more robust

Pavlovic, Rehg, Cham, and Murphy, Intl. Conf. Computer Vision, 1999

Figure 4: (a) Tracker (in white) using constant velocity predictor drifts off track by frame 7. (b) SLDS-based tracker is on track at frame 7. Model (switching state) 3 has the highest likelihood. Black lines show prior mean and observation. (c) SLDS tracker at frame 20.
But we can fake it with clever system design

Research at MERL on fast, low-cost vision systems

From MERL and Mitsubishi Electric:

David Anderson, Paul Beardsley, Chris Dodge, William Freeman, Hiroshi Kage, Kazuo Kyuma, Darren Leigh, Neal McKenzie, Yasunari Miyake, Michal Roth, Ken-ichi Tanaka, Craig Weissman, William Yerazunis
Computer vision based interface

The hope: video input will give a more expressive, natural or engaging interface.
Existing interfaces devices are fast & low-cost.
Applications make the vision easier.

Constraints simplify recognition--if you know where the tracks are, it’s easy to guess where the train is.
There is a human in the loop.

- Rich, immediate visual, audio feedback.
- The player can correct for algorithm imperfections.
Computer vision algorithms as ocean-going vessels
Computer vision algorithms as ocean-going vessels
1. Selected appliance: television
television market

~1 billion television sets
Survey

“What high technology gadget has improved the quality of your life the most?”

What two things were mentioned most?
Survey results

“What high technology gadget has improved the quality of your life the most?”

Microwave ovens and TV remote controls
--Porter/Novelli survey, 1995

message:
People value the ability to control a television from a distance.
Control of television set from a distance

*Wired remote control.*

*Infra-red remote control.*

*Voice control.*

*Gesture control.*
Design constraints

- *From the user’s point of view*

- *From the computer’s point of view*
From the user’s point of view:

Complex commands require complicated gestures?

“mute”
From the computer’s point of view:

**Living room scene is difficult**

How can the computer find the hand, and recognize its gesture, in this complicated, unpredictable visual scene?
Our solution: exploit the visual feedback from the television

user

|televisio

Volume
hand recognition method: template matching

Examine the squared difference between (a) pixel values in the hand template, and (b) pixel values in a square centered at each possible position in the image.
hand recognition method: normalized correlation

template

image

normalized correlation
Normalized correlation

\[
\frac{\mathbf{a} \cdot \mathbf{b}}{\sqrt{(\mathbf{a} \cdot \mathbf{a})(\mathbf{b} \cdot \mathbf{b})}}
\]

Where \( \mathbf{a} \) and \( \mathbf{b} \) are vectors from rasterized patches of the image and template.
Background removal

\[ \text{running average} \times (1 - \alpha) + \alpha \times \text{current image} \]

\[ \text{background removed} \]
Prototype of television controlled by hand signals.
TV screen overlay
Video
Prototype limitations

- **Distance from camera:**
  6 - 10 feet.

- **Field of view:**
  trigger gesture: 15° tracking: 25°

- **Coupling to television is loose.**

- **Two screens instead of one.**

- **Robustness during operation:**
  no template adaptation to different users.  
  background removal may need variable contrast control.
Product hardware requirements

**Short term**
- camera
- video digitizer
- computer

**Long term**
- TV’s / computers / browsers will have cameras and powerful computers.
- a software product.
2. Simple gesture recognition method
Real-time hand gesture recognition by orientation histograms
Orientation measurements (bottom) are more robust to lighting changes than are pixel intensities (top)
Orientation measurements (bottom) are more robust to lighting changes than are pixel intensities (top)
A Simple illustration of an orientation histogram. (1) An image of a horizontal edge has only one orientation at a sufficiently high contrast. (2) Thus the raw orientation histogram has counts at only one orientation value. (3) To allow neighboring orientations to sense each other, we blurred the raw histogram. (4) The same information, plotted in polar coordinates. We define the orientation to be the direction of the intensity gradient, plus 90 degrees.
Images, orientation images, and orientation histograms for training set
Test image, and distances from each of the training set orientation histograms (categorized correctly).
Crane movements controlled by hand gestures
Janken game
7 Problem images for the orientation histogram-based gesture classifier.
Games add fun and purpose: “Get the sprite through the golden rings.”

Field test results from Disney’s VR Aladdin.

“Guests cared about the experience, not the technology.”
Games selected for vision interface
Image moments give a very coarse image summary.

\[
\begin{align*}
M_{00} &= \sum_{x} \sum_{y} I(x, y) \\
M_{10} &= \sum_{x} \sum_{y} x \cdot I(x, y) \\
M_{01} &= \sum_{x} \sum_{y} y \cdot I(x, y) \\
M_{20} &= \sum_{x} \sum_{y} x^2 \cdot I(x, y) \\
M_{11} &= \sum_{x} \sum_{y} x \cdot y \cdot I(x, y) \\
M_{02} &= \sum_{x} \sum_{y} y^2 \cdot I(x, y)
\end{align*}
\]
Hand images and equivalent rectangles having the same image moments
Artificial Retina chip for detection and low-level image processing.
Artificial Retina chip

VSPC: Variable Sensitivity Photodetection Cell

Control Vector $S$

Input Image $W$

Scanner $V_m$

$V_p$

$V_r$

Multiplexer

$V_x$

$\text{out} = W \cdot S$

Processed Image
Artificial Retina functions

- Image Detection
- Edge Extraction
- Smoothing

- Random Access
- Pattern Matching
- Projection (2D->1D compression)
Fast image moment calculation with artificial retina chip

Processing time for image projections:
- w/o AR chip: 10 msec
- with AR chip: 0.3 msec
Hand gesture-controlled robot
Game: Nights
Moment-based pointing control

Center-of-mass of absolute value of difference-image
Moment-based pointing control

Line to difference-image center-of-mass determines flight direction.
Game: Magic Carpet
Magic carpet game--figure analysis by hierarchical image moments
Game: Decathlete
Optical-flow-based Decathlete figure motion analysis
Decathlete 100m hurdles
Decathlete javelin throw
Decathlete javelin throw
video
Nintendo Game Boy Camera

Several million sold (most of any digital camera). Imaging chip is Mitsubishi Electric’s “Artificial Retina” CMOS detector.
Summary

- Fast, simple algorithms and low-cost hardware are well-suited to interactive graphics applications.
- We followed this approach to make a television controlled by hand gestures, simple hand gesture recognition, and vision-based computer game interfaces.
To Trevor’s slides...
Perceptive Context for Pervasive Computing

Trevor Darrell
Vision Interface Group
MIT AI Lab
Perceptually Aware Displays

Camera associated with display
Display should respond to user
  - font size
  - attentional load
  - passive acknowledgement

e.g., “Magic Mirror”, Interval
Compaq’s Smart Kiosk
ALIVE, MIT Media Lab
Example: A Face Responsive Display

• Faces are natural interfaces!
  - Ubiquitous, fast, expressive, general.
  - Want machines to generate and perceive faces.

• A Face Responsive Display...
  - Knows when it’s being observed
  - Recognizes returning observers
  - Tracks head pose
  - Robust to changing lighting, moving backgrounds…
A Face Responsive Display

Tasks
- Detection
- Identification
- Tracking

How? Exploit multiple visual modalities:
- Shape
- Color
- Pattern
### Tasks and Visual Modalities

<table>
<thead>
<tr>
<th></th>
<th>shape</th>
<th>color</th>
<th>pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>detection</strong></td>
<td>silhouette classifier</td>
<td>skin classifier</td>
<td>face detection</td>
</tr>
<tr>
<td><strong>identification</strong></td>
<td>biometrics</td>
<td>flesh hue</td>
<td>face recognition</td>
</tr>
<tr>
<td><strong>tracking</strong></td>
<td>coarse motion estimation</td>
<td>clothing histogram</td>
<td>fine motion estimation / pose tracking</td>
</tr>
</tbody>
</table>
## Mode and Task Matrix

<table>
<thead>
<tr>
<th></th>
<th>shape</th>
<th>color</th>
<th>pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>detection</td>
<td>silhouette classifier</td>
<td>skin classifier</td>
<td>face detection</td>
</tr>
<tr>
<td>identification</td>
<td>biometrics</td>
<td>flesh hue</td>
<td>face recognition</td>
</tr>
<tr>
<td>tracking</td>
<td>Shape change</td>
<td>clothing histogram</td>
<td>Appearance change</td>
</tr>
</tbody>
</table>
Finding Features

2D Head / hands localization
- contour analysis: mark extremal points (highest curvature or distance from center of body) as hand features
- use skin color model when region of hand or face is found (color model is independent of flesh tone intensity)
Flesh color tracking

- Often the simplest, fastest face detector!
- Initialize region of hue space

[ Crowley, Coutaz, Berard, INRIA ]
Color Processing

- Train two-class classifier with examples of skin and not skin
- Typical approaches: Gaussian, Neural Net, Nearest Neighbor
- Use features invariant to intensity
  
  Log color-opponent [Fleck et al.]
  
  \((\log(r) - \log(g), \log(b) - \log((r+g)/2))\)

  Hue & Saturation
Flesh color tracking

Can use Intel OpenCV lib’s CAMSHIFT algorithm for robust real-time tracking. (open source impl. avail.!)
**Detection with multiple visual modes**

<table>
<thead>
<tr>
<th>Shape</th>
<th>Find head sized peaks in 2-D or 3-D.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flesh Color Detection</td>
<td>Detect skin pigment in hue-based color space</td>
</tr>
<tr>
<td>Face Pattern Detection</td>
<td>Classify intensity vector corresponding to face class</td>
</tr>
</tbody>
</table>
**Common Detection Failure Modes**

<table>
<thead>
<tr>
<th>Shape</th>
<th>Fooled by head shaped peaks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flesh Color Detection</td>
<td>Fooled by flesh colored objects</td>
</tr>
<tr>
<td>Face Pattern Detection</td>
<td>Misses out of plane rotation or expression</td>
</tr>
</tbody>
</table>
Robust real-time performance

Shape

Flesh Color Detection

Face Pattern Detection

Integrated Face Detection Algorithm (temporally asynch. voting scheme)
## Mode and Task Matrix

<table>
<thead>
<tr>
<th></th>
<th>shape</th>
<th>color</th>
<th>pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>detection</td>
<td>silhouette classifier</td>
<td>skin classifier</td>
<td>face detection</td>
</tr>
<tr>
<td>identification</td>
<td>biometrics</td>
<td>flesh hue</td>
<td>face recognition</td>
</tr>
<tr>
<td>tracking</td>
<td>Shape change</td>
<td>clothing histogram</td>
<td>Appearance change</td>
</tr>
</tbody>
</table>
A Key Technology: Video-Rate Stereo

- Two cameras $\rightarrow$ stereo range estimation; disparity proportional to depth
- Depth makes tracking people easy
  - segmentation
  - shape characterization
  - pose tracking
- Real-time implementations becoming commercially available.
Video-rate stereo

Left and right images

Computed disparity

Foreground pixels; grouped by local connectivity
RGBZ input
RGBZ input
RGBZ input
Video-Rate Stereo

• Multiple cameras $\rightarrow$ stereo range estimation; disparity proportional to depth
• Real-time implementations becoming commercially available.
• Depth makes tracking people easier
  - segmentation
  - shape characterization
  - pose tracking
Range feature for ID!

- Body shape characteristics -- e.g., height measure.
- Normalize for motion/pose: median filter over time

- Near future: full vision-based kinematic estimation and tracking--active research topic in many labs.
Color feature for ID!

For long-term tracking / identification, measure color hue and saturation values of hair and skin.

For same-day ID, use histogram of entire body / clothing
## Mode and Task Matrix

<table>
<thead>
<tr>
<th>Detection</th>
<th>Shape</th>
<th>Color</th>
<th>Pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>silhouette classifier</td>
<td>skin classifier</td>
<td>face detection</td>
</tr>
<tr>
<td>Identification</td>
<td>biometrics</td>
<td>flesh hue</td>
<td>face recognition</td>
</tr>
<tr>
<td>Tracking</td>
<td>Shape change</td>
<td>clothing histogram</td>
<td>Appearance change</td>
</tr>
</tbody>
</table>

See lectures by Trevor later in the course
Robust, Multi-modal Algorithm

Combine modules for detection:

- Silhouette finds body
- Color tracks extremities
- Pattern discriminates head from hands.

Use each also to recognize returning people:

- Face recognition
- Biometrics (skeletal structure)
- Hair and Skin hue
- Clothing (intra-day.)

[ CVPR '98; T. Darrell, G. Gordon, M. Harville, J. Woodfill ]
System Overview

Diagram showing the system overview with various processes and steps:
- Face pattern detection
- Color classification and segmentation
- Mask
- Range computation and foreground segmentation
- Head detection from body silhouette
- Normalization for scale and alignment
- Skin and Hair color description
- Height computation
- Person identification for extended tracking
- Classification of identity
- Selection of face region and tracking
- Application computing
Classic Background Subtraction model

- Background is assumed to be mostly static
- Each pixel is modeled as by a gaussian distribution in YUV space
- Model mean is usually updated using a recursive low-pass filter

Given new image, generate silhouette by marking those pixels that are significantly different from the “background” value.
Static Background Modeling Examples

[MIT Media Lab Pfinder / ALIVE System]
Static Background Modeling Examples

[MIT Media Lab Pfinder / ALIVE System]
Static Background Modeling Examples

[MIT Media Lab Pfinder / ALIVE System]
The ALIVE System

Video Screen

Camera

User

Autonomous Agents
ALIVE

- Real sensing for virtual world
- Tightly coupled sensing-behavior-action
- Vision routines: body/head/hand tracking

[ Blumberg, Darrell, Maes, Pentland, Wren, ... 1995 ]
ALIVE system, MIT

The ALIVE System:

Wireless, Full-body Interaction with Autonomous Agents

Pattie Maes, Trevor Darrell, Bruce Blumberg, Alex Pentland
MIT Media Laboratory

http://vismod.www.media.mit.edu/cgi-bin/tr_pagemaker (TR 257)
A Face Responsive Display
Vision-only Application: Interactive Video Effects
end