Looking at people (again!)

D.A. Forsyth, UIUC (was U.C. Berkeley; was U.Iowa)

contributions from:
Derek Hoiem (UIUC), Leslie Ikemoto, Okan Arikan, of Animeeple
Deva Ramanan of TTI/UC Irvine
Ali Farhadi of UIUC Nazli Ikizler of Bilkent U (now Boston U; soon Hacettepe U)
Alex Sorokin, UIUC Du Tran, UIUC Duan Tran, UIUC, Wei Yan, Texas A+M

Thanks to: Electronic Arts, Sony SCEA, ONR MURI, NSF, DHS
Why are humans important?

• **Surveillance**
  • prosecution; intelligence gathering; crime prevention
  • HCI; architecture;

• **Synthesis**
  • games; movies;

• **Safety applications**
  • pedestrian detection

• **People are interesting**
  • movies; news
Where you are can suggest you are doing something you shouldn’t be
Boult 2001
Bill Freeman flies a magic carpet.

Orientation histograms detect body configuration to control bank, raised arm to fire magic spell.

Freeman et al, 98.
An example of a user playing a Decathlon event, the javelin throw. The computer’s timing of the set and release for the javelin is based on when the integrated downward and upward motion exceeds predetermined thresholds.

Motion fields set javelin timing
Freeman et al 98
Sony’s eyetoy estimates motion fields, links these to game inputs. Huge hit in EU, well received in US.
Computational Behavioural Science

- Observe people
  - Using vision, physiological markers
    - Interacting, behaving naturally
  - In the wild
- drive feedback for therapy
  - Eg reward speech
- Applications
  - Model: screen for ASD
- Other:
  - Any where large scale observations help
    - Support in home care
    - Support care for demented patients
    - Support stroke recovery
    - Support design of efficient buildings
- 10M$, 5yr NSF award under Expeditions program
  - GaTech, UIUC(DAF, Karahalios), MIT, CMU, Pittsburgh, USC, Boston U
Rapid ABC

- Easily administered screening test
- Challenge:
  - Automatic evaluation
  - To use unskilled screeners
Why are humans important?

- **Surveillance**
  - prosecution; intelligence gathering; crime prevention
  - HCI; architecture;
- **Synthesis**
  - games; movies;
- **Safety applications**
  - pedestrian detection
- **People are interesting**
  - movies; news
Why are humans important?

- **Surveillance**
  - prosecution; intelligence gathering; crime prevention
  - HCI; architecture;
- **Synthesis**
  - games; movies;
- **Safety applications**
  - pedestrian detection
- **People are interesting**
  - movies; news
From Dalal+Triggs, 05

DET – different descriptors on MIT database

- Lin. R–HOG
- Lin. C–HOG
- Lin. EC–HOG
- Wavelet
- PCA–SIFT
- Lin. G–ShapeC
- Lin. E–ShapeC
- MIT best (part)
- MIT baseline

DET – different descriptors on INRIA database

- Ker. R–HOG
- Lin. R2–HOG
- Lin. R–HOG
- Lin. C–HOG
- Lin. EC–HOG
- Wavelet
- PCA–SIFT
- Lin. G–ShapeC
- Lin. E–ShapeC
Why are humans important?

- **Surveillance**
  - prosecution; intelligence gathering; crime prevention
  - HCI; architecture;
- **Synthesis**
  - games; movies;
- **Safety applications**
  - pedestrian detection
- **People are interesting**
  - movies; news
News Faces

- 5e5 captioned news images
- Mainly people “in the wild”
- Correspondence problem
  - some images have many (resp. few) faces, few (resp. many) names (cf. Srihari 95)

• Process
  - Extract proper names
  - Detect faces (Vogelhuber Schmid 00) 44773 big face responses
  - Rectify faces 34623 properly rectified
  - Kernel PCA rectified faces
  - Estimate linear discriminants
  - Now have (face vector; name_1,.,., name_k) 27742 for k<=4

• Apply a form of modified k-means
Structure

• **What can we do?**
  • mainly, tag some known activities with classifiers

• **What should we be doing?**
  • building representations to describe the unfamiliar

• **How do we get information from the image signal?**
  • tracking/parsing the body to get arms and legs

• **What’s the form of the representation?**
  • contact, timing, style attributes
What we can do

- Primary machine is the classifier
  - features in, decision out
  - train with examples
- Decision is typically motion label
  - “run”, “walk”, “fight”, etc.
  - drawn from vocabularies of 5-50 (or so, depending on paper)
Datasets

- **IXMAS**
  - Actors: 12, 13, 36, 5

- **Weizman**
  - Actors: 9, 10, 93, 1
  - Actions: 9, 10, 93, 1
  - Sequences: 9, 10, 93, 1
  - Views: 9, 10, 93, 1

- **Our dataset**
  - Images of various poses

- **UMD**
  - Images of various poses
  - Actors: 8, 14, 532, 1
  - Actions: 10, 100, 2
Discriminative results

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Algorithm</th>
<th>Chance</th>
<th>Protocols</th>
<th>Discriminative task</th>
<th>Reject</th>
<th>Few examples</th>
<th>L1SO</th>
<th>L1AAO</th>
<th>L1AO</th>
<th>L1VO</th>
<th>UNa</th>
<th>FE-1</th>
<th>FE-2</th>
<th>FE-4</th>
<th>FE-8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weizman</td>
<td>NE(k=300)</td>
<td>10.00</td>
<td>91.40</td>
<td>93.50</td>
<td>95.70</td>
<td>N/A</td>
<td>0.00</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1NN</td>
<td>10.00</td>
<td>95.70</td>
<td>95.70</td>
<td>96.77</td>
<td>N/A</td>
<td>0.00</td>
<td>53.00</td>
<td>73.00</td>
<td>89.00</td>
<td>96.00</td>
<td>N/A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1NN-M</td>
<td>10.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>N/A</td>
<td>0.00</td>
<td>72.31</td>
<td>81.77</td>
<td>92.97</td>
<td>100.00</td>
<td>N/A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1NN-R</td>
<td>9.09</td>
<td>83.87</td>
<td>84.95</td>
<td>84.95</td>
<td>N/A</td>
<td>0.00</td>
<td>84.95</td>
<td>17.96</td>
<td>42.04</td>
<td>68.92</td>
<td>84.95</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1NN-MR</td>
<td>9.09</td>
<td>89.66</td>
<td>89.66</td>
<td>89.66</td>
<td>N/A</td>
<td>0.00</td>
<td>90.78</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Our</td>
<td>NE(k=600)</td>
<td>7.14</td>
<td>98.70</td>
<td>98.70</td>
<td>98.70</td>
<td>N/A</td>
<td>0.00</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1NN</td>
<td>7.14</td>
<td>98.87</td>
<td>97.74</td>
<td>98.12</td>
<td>N/A</td>
<td>0.00</td>
<td>58.70</td>
<td>76.20</td>
<td>90.10</td>
<td>95.00</td>
<td>N/A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1NN-M</td>
<td>7.14</td>
<td><strong>99.06</strong></td>
<td><strong>97.74</strong></td>
<td><strong>98.31</strong></td>
<td>N/A</td>
<td>0.00</td>
<td>88.80</td>
<td>94.64</td>
<td>95.63</td>
<td>98.86</td>
<td>N/A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1NN-R</td>
<td>6.67</td>
<td>95.86</td>
<td>81.40</td>
<td>82.10</td>
<td>N/A</td>
<td>0.00</td>
<td>81.20</td>
<td>27.40</td>
<td>37.90</td>
<td>51.00</td>
<td>65.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1NN-MR</td>
<td>6.67</td>
<td><strong>98.68</strong></td>
<td><strong>91.73</strong></td>
<td><strong>91.92</strong></td>
<td>N/A</td>
<td>0.00</td>
<td><strong>91.11</strong></td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IXMAS</td>
<td>NE(k=600)</td>
<td>7.89</td>
<td>80.00</td>
<td>78.00</td>
<td>79.90</td>
<td>N/A</td>
<td>0.00</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1NN</td>
<td>7.89</td>
<td>81.00</td>
<td>75.80</td>
<td>80.22</td>
<td>N/A</td>
<td>0.00</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1NN-R</td>
<td>7.14</td>
<td><strong>65.41</strong></td>
<td><strong>57.44</strong></td>
<td><strong>57.82</strong></td>
<td>N/A</td>
<td>0.00</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td><strong>57.48</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UMD</td>
<td>NE(k=300)</td>
<td>10.00</td>
<td>100.00</td>
<td>N/A</td>
<td>N/A</td>
<td>97.50</td>
<td>0.00</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1NN</td>
<td>10.00</td>
<td>100.00</td>
<td>N/A</td>
<td>N/A</td>
<td>97.00</td>
<td>0.00</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1NN-R</td>
<td>9.09</td>
<td><strong>100.00</strong></td>
<td>N/A</td>
<td>N/A</td>
<td><strong>88.00</strong></td>
<td><strong>88.00</strong></td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Works well, depending on task; not rejecting improves things. Metric learning improves things.

Tran + Sorokin 08
Laptev Perez 2007
see also Laptev et al 08
<table>
<thead>
<tr>
<th>AnswerPhone</th>
<th>GetOutCar</th>
<th>HandShake</th>
<th>HugPerson</th>
<th>Kiss</th>
<th>SitDown</th>
<th>SitUp</th>
<th>StandUp</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="AnswerPhone" /></td>
<td><img src="image2" alt="GetOutCar" /></td>
<td><img src="image3" alt="HandShake" /></td>
<td><img src="image4" alt="HugPerson" /></td>
<td><img src="image5" alt="Kiss" /></td>
<td><img src="image6" alt="SitDown" /></td>
<td><img src="image7" alt="SitUp" /></td>
<td><img src="image8" alt="StandUp" /></td>
</tr>
<tr>
<td>TP</td>
<td>TN</td>
<td>FP</td>
<td>FN</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Movies and captions: Laptev et al 08
Choi Shahid Savarese 09
Predicting stylized narrations

Pitcher pitches the ball before Batter hits. Batter hits and then simultaneously Batter runs to base and Fielder runs towards the ball. Fielder catches the ball after Fielder runs towards the ball. Fielder catches the ball before Fielder throws to the base. Fielder throws to the base and then Fielder at Base catches the ball at base.

Pitcher pitches the ball and then Batter hits. Fielder catches the ball after Batter hits.

Pitcher pitches the ball before Batter hits. Batter hits and then simultaneously Batter runs to base and Fielder runs towards the ball. Fielder runs towards the ball and then Fielder catches the ball. Fielder throws to the base after Fielder catches the ball. Fielder throws to the base and then Fielder at Base catches the ball at base.

Pitcher pitches the ball and then Batter does not swing.

Gupta ea 09
Structure

• What can we do?
  • mainly, tag some known activities with classifiers

• What should we be doing?
  • building representations to describe the unfamiliar

• How do we get information from the image signal?
  • tracking/parsing the body to get arms and legs

• What’s the form of the representation?
  • contact, timing, style attributes
What should activity recognition say?

- **Report names of activity of all actors (?!?)**
  - but we might not have names
  - and some might not be important

- **Make useful reports about what’s going on**
  - what is going to happen?
  - how will it affect me?
  - who’s important?

- **Do activity categories exist?**
  - allow generalization
    - future behavior; non-visual properties of activities
Unfamiliar activities present no real problem
Unfamiliar activities present no real problem
Unfamiliar activities present no real problem
How is it going to affect me?
What outcome do we expect?

How are other people feeling?

What will they do?
What outcome do we expect?

How are other people feeling?

What will they do?
How many adults were on the platform and what were they doing?
What’s going to happen to the baby?
What outcome do we expect?

How are other people feeling?

What will they do?
Choosing what to report

Two girls take a break to sit and talk.

Two women are sitting, and one of them is holding something.

Two women chatting while sitting outside.

Two women sitting on a bench talking.

Two women wearing jeans, one with a blue scarf around her head, sit and talk.

Sentences from Julia Hockenmaier’s work

Rashtchian ea 10
Figure 2: Images with annotating sentences, produced by workers on Amazon’s Mechanical Turk. Notice that generally the sentences are quite consistent, being simple descriptions of the image. There are some issues with annotators whose first language is not English (e.g. “goats” in the picture of sheep). Occasionally, annotators produce mysterious sentences (e.g. “Blue and red in the air”), and sentences are not always just lists of visible entities (e.g. “A car on a dirt and gravel road approaches a group of three sheep grazing” — the car is hardly visible). When the object is not known, we can still make useful statements about its properties. A paper describing this method has been accepted to CVPR 2008. We have shown that image annotations that mention appearance properties (for example, “a green hat”) allow more accurate visual learning than pure object names; this is because the appearance attribute can be used to focus the search for the object example. A paper describing this procedure has been accepted to ICCV 2009.

Machine learning methods: Our proposed work requires large-scale machine learning using supervised and partially supervised data. We have developed methods to train kernel SVM classifiers that are simple and fast, and produce a classifier that can be evaluated quickly. We have shown how the classifier can be used to categorize images very effectively in a paper accepted to ICCV 2009. We have demonstrated that a method for partially supervised learning that we developed for other purposes can be used to improve the accuracy of image labelling. A paper describing these results has been accepted to ICCV 2009.

Data sets: We have produced a large attribute data set, which we will release to the community. We are in the process of collecting a second attribute data set, which will include more objects, more attributes, and exact spatial maps of the occurrence of the attributes. We have collected two sets of images annotated with sentences using Amazon’s Mechanical Turk (examples appear in figure 2). The first data set contains 1000 images from the PASCAL object recognition challenge, whereas the second data set consists of 8000 action images harvested from Flickr.com. Each image is annotated with five independently produced captions. We find that different annotators produce surprisingly consistent sentences for an image, and that these sentences tend to list the main entities present in the image, with the probability an entity is mentioned rather roughly proportional to its importance. We find that it is important to prequalify annotators so that only first-language English speakers annotate; careful instructions to the annotators are helpful, too. Our findings are described in a paper presented at the Mechanical Turk workshop associated with NAACL 2010.

The goats on the way
A car on a rural dirt and gravel road approaches a group of three sheep grazing. A small group of sheep in a dirt road.
Three sheep on a rural road, about to block traffic.
Three sheeps on the road out of nowhere.
Structure

• What can we do?
  • mainly, tag some known activities with classifiers

• What should we be doing?
  • building representations to describe the unfamiliar

• How do we get information from the image signal?
  • tracking/parsing the body to get arms and legs

• What’s the form of the representation?
  • contact, timing, style attributes
Point tracks reveal curious phenomena in public spaces

Yan+Forsyth, 04
Goals, intentions, outcomes

- Probably need to know some of body configuration
  - to reason about current contacts
    - man is on bicycle
    - woman is on platform
  - to reason about future contacts, eg
    - man is flying off bicycle and will hit water
    - woman is reaching for baby carriage
  - to reason about unfamiliar movements
    - what is he doing with his arm?
Why is kinematic tracking hard?

- It’s hard to detect people
  - until recently, human trackers were manually started
- People move fast, and can move unpredictably
  - dynamics gives limited constraint on future configuration
  - appearance changes over time (shading, aspect, etc)
- Some body parts are small and tend to have poor contrast
  - particularly difficult to track
    - lower arms (small, fast, look like other things);
    - upper arms (poor contrast)

variation in pose & aspect
self-occlusion & clutter
variation in appearance
Build and detect models

small scale

learn limb classifiers

label pixels

torso

arm

leg

head

unusual pose

"Lola" likelihood

Ramanan, Forsyth and Zisserman CVPR05
Coming to tracking

- Advances in human parsing
  - Appearance/layout interaction (Ramanan 06)
  - Improved appearance models (Ferrari et al 08; Eichner Ferrari 10)
  - Branch+bound (Tian Sclaroff 10)
  - Interactions with objects (Yao Fei-Fei 10; Desai et al 10)
  - Coverage and background (Buehler ea 08; Jiang 09)
  - Complex spatial models (Sapp ea 10a)
  - Cascade models (Sapp ea 10b)
  - Full relational models (Tran Forsyth 10)
Naming activities

• Build a set of basic labels
  • guess them: walk, run, stand, reach, crouch, etc.

• Composite Activity model:
  • Product of finite state automata for arms, legs built from MoCap
  • Arms, legs each have local short timescale activity models for basic labels
  • Link these models into a large model, using animation-legal transitions
Searching for complex human activities with no visual examples N İkizler, DA Forsyth - IJCV, 2008
Rapid ABC

- Easily administered screening test
- Challenge:
  - Automatic evaluation
  - To use unskilled screeners
Structure

• What can we do?
  • mainly, tag some known activities with classifiers

• What should we be doing?
  • building representations to describe the unfamiliar

• How do we get information from the image signal?
  • tracking/parsing the body to get arms and legs

• What’s the form of the representation?
  • contact, timing, style attributes
What is an object like?

Viz comic, issue 101
Possible architecture
Attribute phenomena

- Some are easily predicted from pictures
  - eg “red”, “wooden”
- Some are properly inherited from category
  - eg “mammal”
- They are heavily correlated
  - easy binary variable argument
- Some are “stuff”-like
  - eg “red”, “wooden”
- Others “thing”-like
  - eg “wheel”, “leg”
- Within class variation
  - Different instances of the same category could have different attributes

“Stuff” -- shape doesn’t matter (sky, grass, bush)  cf mass noun

“Thing” -- shape matters (cow, cat, car)  cf count noun
Farhadi ea 09
Missing attributes

- Aeroplane: No "wing"
- Car: No "window"
- Boat: No "sail"
- Aeroplane: No "jet engine"
- Motorbike: No "side mirror"
- Car: No "door"
- Bicycle: No "wheel"
- Sheep: No "wool"
- Train: No "window"
- Sofa: No "wood"
- Bird: No "tail"
- Bird: No "leg"
- Bus: No "door"

Farhadi ea 09
Extra attributes

Bird
"Leaf"

Bus
"face"

Motorbike
"cloth"

DiningTable
"skin"

People
"Furn. back"

Aeroplane
"beak"

People
"label"

Sofa
"wheel"

Bike
"Horn"

Monitor
"window"

Farhadi ea 09
Activity attributes

- Gaze and focus
- Style
  - Fast/Gentle
- Timing
  - arms in phase with legs
- Contact
  - Having hand contact
- Kinematic
  - Arms sticking out

Nearby objects and free space
Gaze and focus: Rapid ABC

- Easily administered screening test
  - Challenge:
    - Automatic evaluation
    - To use unskilled screeners
Contact and kinematics: Picking things up
Animation tells us about attributes

- Relative timing of movements across the body matters
  - however, no real models here

- Contact matters
  - people are highly sensitive to incoherent contacts

- Style matters
  - people are good at consistency between motion style and body shape
Relative timing matters
Relative timing matters
Different bodies have different styles
Different bodies have different styles
Open question: similarity

• This activity is like that one
  • therefore, the outcome might be similar

• In what way like? how do we score this?

• Advantages
  • strong improvements in training with few examples
    • (Wang, 10; some cases)
    • perhaps allows recognition/prediction with no examples
Summary

• Extend attribute based representations to describe activity
  • starting at least with
    • Gaze/focus
    • Style
    • Timing
    • Contact
    • Kinematics
    • Nearby objects or free space
• Select what is important from sequences
  • perhaps for predictive purposes
• Build procedures to use similarity of motion/outcome
  • to train models with little data
Thanks

UIUC Vision & Graphics groups
UC Berkeley Vision & Graphics Groups
Oxford Visual Geometry Group, particularly Andrew Zisserman
Dept. Homeland Security
ONR MURI
NSF
Electronic Arts
Sony Computer Entertainment