Intro to People & Actions

Tamara Berg

CSE595 Words & Pictures
Reminders

• In class project proposals March 22 – come see me in office hours (today) to clear project topics.
• HW3 due March 27.
Face Detection

Sliding window approach: for each possible rectangular region in the image asks the question, is there a face here?

Now even integrated into many consumer digital cameras.

What is a face?
Face Detection in OpenCV

OpenCV - (Open Source Computer Vision) is a library of programming functions for real time computer vision. http://opencv.willowgarage.com/wiki/
Face Recognition

Face Recognition

Face recognition by fitting a 3d model - Volker Blanz and Thomas Vetter 2003.
Recognition by Attributes
... but people are more than just faces
Capturing Humans in Motion

Loss of *depth* and *motion* in projection to 2D images.

Eadweard Muybridge, 1884-5. Multiple cameras.

Articulated Body Model

Kinematic tree:
Marr & Nishihara '78

\[ \begin{bmatrix} \theta_{1,0}^{1,0}, \theta_{1,0}^{1,0}, \theta_{1,0}^{1,0} \end{bmatrix} \]

> 3D space

\[ \begin{bmatrix} \tau_{0,g}^{0,0}, \tau_{0,g}^{0,0}, \tau_{0,g}^{0,0} \end{bmatrix} \]

\[ \begin{bmatrix} \theta_{x}^{0,0}, \theta_{y}^{0,0}, \theta_{z}^{0,0} \end{bmatrix} \]

Represent a “pose” at time \( t \) by a vector of these parameters: \( X_t \)
Motion Capture

From images to “models” that support reasoning.

Recover 3D pose and motion.
“Mocap” Today

CMU Mocap lab. 2003
“Mocap” Today

Walking motion learned from mocap.
“Mocap” Today
Mocap in the Wild

Humans in captivity

Humans in their natural habitat
Mocap Now?

Xbox Kinect in our lab with OpenNI extracting skeletons.
Problems

The appearance/size/shape of people can vary dramatically (high-D space).

Underlying structure (bones and joints) is unobservable (obscured by muscle, skin, clothing).

Occlusion and partial views.
Problems

- Loss of 3D in 2D projection
- Unusual poses
- Self occlusion
- Low contrast
Problems

Multiple people and occlusion leads to ambiguity. Moving cameras & complex changing backgrounds.
Problems

Accidental alignment

Motion blur.
(nothing to match)
Detecting and Tracking People

* Where are the people?
* What are their poses?
* How are they moving?
* What are they doing?
Why?

surveillance

interfaces

video data mining

motion capture

Slide credit: D. Ramanan
Tasks

• Pedestrian detection
• Pose Recognition
• Tracking
• Activity Recognition
Where are the people?
Pedestrian detection

Bastian Leibe, Edgar Seemann, and Bernt Schiele
Detection: The Pure ML Approach

Single image

\[ \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_N \end{bmatrix} \]

Classifier

Person/Not-person
Support Vector Machines

Support Vector Machines

Product of wavelet templates and filtered image regions gives a vector of responses for each region.

Bootstrapped SVM learns the classify pedestrian/background.

AdaBoost

45,396 possible features

Viola, Jones and Snow, ICCV’03
Boosting

- Defines a classifier using an additive model:

\[ F(x) = \alpha_1 f_1(x) + \alpha_2 f_2(x) + \alpha_3 f_3(x) + \ldots \]
Boosting

• Defines a classifier using an additive model:

\[ F(x) = \alpha_1 f_1(x) + \alpha_2 f_2(x) + \alpha_3 f_3(x) + \ldots \]

• We need to define a family of weak classifiers

\[ f_{k'}(x) \] from a family of weak classifiers

---

Slide credit: Antonio Torralba
Adaboost

Given: \((x_1, y_1), \ldots, (x_m, y_m)\) where \(x_i \in X, y_i \in Y = \{-1, +1\}\)
Initialize \(D_1(i) = 1/m\).
For \(t = 1, \ldots, T\):

- Train weak learner using distribution \(D_t\).
- Get weak hypothesis \(h_t : X \to \{-1, +1\}\) with error
  \[
  \epsilon_t = \Pr_{i \sim D_t} [h_t(x_i) \neq y_i].
  \]
- Choose \(\alpha_t = \frac{1}{2} \ln \left( \frac{1 - \epsilon_t}{\epsilon_t} \right)\).
- Update:
  \[
  D_{t+1}(i) = \frac{D_t(i)}{Z_t} \times \begin{cases} 
    e^{-\alpha_t} & \text{if } h_t(x_i) = y_i \\
    e^{\alpha_t} & \text{if } h_t(x_i) \neq y_i
  \end{cases} = \frac{D_t(i) \exp(-\alpha_t y_i h_t(x_i))}{Z_t}
  \]
  where \(Z_t\) is a normalization factor (chosen so that \(D_{t+1}\) will be a distribution).

Output the final hypothesis:
\[
H(x) = \text{sign} \left( \sum_{t=1}^{T} \alpha_t h_t(x) \right).
\]

Slide credit: Antonio Torralba
Each data point has a class label:

\[ y_t = \begin{cases} +1 & \text{(\textcolor{red}{$\bullet$})} \\ -1 & \text{(\textcolor{blue}{$\bullet$})} \end{cases} \]

and a weight:

\[ w_t = 1 \]

• It is a sequential procedure:
Toy example

Weak learners from the family of lines

Each data point has a class label:

\[ y_t = \begin{cases} 
+1 & \text{(red)} \\
-1 & \text{(blue)} 
\end{cases} \]

and a weight:

\[ w_t = 1 \]

\[ h \Rightarrow p(\text{error}) = 0.5 \] it is at chance

Slide credit: Antonio Torralba
This is a ‘weak classifier’: It performs slightly better than chance.
We set a new problem for which the previous weak classifier performs at chance again.

Each data point has a class label:

\[ y_t = \begin{cases} 
+1 & (\bullet) \\
-1 & (\circ) 
\end{cases} \]

We update the weights:

\[ w_t \leftarrow w_t \exp\{-y_t H_t\} \]
We set a new problem for which the previous weak classifier performs at chance again.

Each data point has a class label:

\[ y_t = \begin{cases} +1 & (\bullet) \\ -1 & (\circ) \end{cases} \]

We update the weights:

\[ w_t \leftarrow w_t \exp\{-y_t H_t\} \]
We set a new problem for which the previous weak classifier performs at chance again.

Each data point has a class label:

\[ y_t = \begin{cases} +1 & (\bullet) \\ -1 & (\circ) \end{cases} \]

We update the weights:

\[ w_t \leftarrow w_t \exp\{-y_t H_t\} \]
We set a new problem for which the previous weak classifier performs at chance again.

Each data point has a class label:

\[ y_t = \begin{cases} 
+1 & (\bullet) \\
-1 & (\bigcirc) 
\end{cases} \]

We update the weights:

\[ w_t \leftarrow w_t \exp\{-y_t H_t\} \]
The strong (non-linear) classifier is built as the combination of all the weak (linear) classifiers.
AdaBoost

45,396 possible features

Viola, Jones and Snow, ICCV’03
Pedestrian Detection

Viola, Jones and Snow, ICCV’03
Where are the people?

Pedestrian detection

Bastian Leibe, Edgar Seemann, and Bernt Schiele
Tasks

- Pedestrian detection
- Pose Recognition
- Tracking
- Activity Recognition
What pose are they in?
Pose Recognition

Mori, Ren, Efros and Malik
What about 3D Pose?

Contour Points / Shape Model

30+ dimensions

Learned mapping

K. Grauman, G. Shakhnarovich, T. Darrell, ICCV’03
Pose matching

“Fast Pose Estimation with Parameter Sensitive Hashing”
Gregory Shakhnarovich, Paul Viola, Trevor Darrell
Tasks

- Pedestrian detection
- Pose Recognition
- Tracking
- Activity Recognition
Why is tracking hard?

- Fast
- Hard to track
- Unpredictable
- Variety of poses
- Hard to detect
- Interactions & occlusions

Slide credit: Deva Ramanan
Basic Approaches

• Model-Free
  – annotate video directly using local features
    (Efros et al., Cutler & Davis, Bobick, Nelmin, Manor & Irani)

• Skeleton Model
  – track skeleton and then recognize pose
    (Yacoob & Black)
Tracking - more formal view

- Very general model:
  - We assume there are moving objects, which have an underlying state $X$
  - There are observations $Y$, some of which are functions of this state
  - There is a clock
    - at each tick, the state changes
    - at each tick, we get a new observation

- Examples
  - object is ball, state is 3D position+velocity, observations are stereo pairs
  - object is person, state is body configuration, observations are frames, clock is in camera (30 fps)
Tracking - Probabilistic formulation

- **Given**
  - \( P(X_{i-1}|Y_0, ..., Y_{i-1}) \)
  - “Prior”

- **We should like to know**
  - \( P(X_i|Y_0, ..., Y_{i-1}) \)
    - “Predictive distribution”
  - \( P(X_i|Y_0, ..., Y_i) \)
    - “Posterior”

Slide credit: D.A. Forsyth
The three main issues in tracking

- **Prediction:** we have seen $y_0, \ldots, y_{i-1}$ — what state does this set of measurements predict for the $i$’th frame? To solve this problem, we need to obtain a representation of $P(X_i|Y_0 = y_0, \ldots, Y_{i-1} = y_{i-1})$.

- **Data association:** Some of the measurements obtained from the $i$-th frame may tell us about the object’s state. Typically, we use $P(X_i|Y_0 = y_0, \ldots, Y_{i-1} = y_{i-1})$ to identify these measurements.

- **Correction:** now that we have $y_i$ — the relevant measurements — we need to compute a representation of $P(X_i|Y_0 = y_0, \ldots, Y_i = y_i)$. 
Key assumptions:

- **Only the immediate past matters:** formally, we require

  \[ P(X_i|X_1, \ldots, X_{i-1}) = P(X_i|X_{i-1}) \]

  This assumption hugely simplifies the design of algorithms, as we shall see; furthermore, it isn’t terribly restrictive if we’re clever about interpreting \( X_i \) as we shall show in the next section.

- **Measurements depend only on the current state:** we assume that \( Y_i \) is conditionally independent of all other measurements given \( X_i \). This means that

  \[ P(Y_i, Y_j, \ldots Y_k|X_i) = P(Y_i|X_i)P(Y_j, \ldots, Y_k|X_i) \]

  Again, this isn’t a particularly restrictive or controversial assumption, but it yields important simplifications.
Model based
Key Idea: Represent Ambiguity

* Represent a multi-modal posterior probability distribution over model parameters
  - sampled representation
  - each sample is a pose and its probability (likelihood weighting)
  - predict over time using a particle filter.

Samples from a distribution over 3D poses.
Appearance based
Tracking Algorithm

Model Build → Detect

original video → 2D track

Slide credit: D. Ramanan
Look for candidate torsos

detected torsos

bag of detected torso patches

Slide credit: D. Ramanan
Cluster

Slide credit: D. Ramanan
Find new torsos using appearance

Slide credit: D. Ramanan
Find arms & legs near torsos
Final Tracks

Slide credit: D. Ramanan
What to detect?

People take on a variety of poses, aspects, scales

self-occlusion  rare pose  motion blur

non-distinctive pose  too small  just right

Slide credit: D. Ramanan
Stylized Pose Person Detector

Slide credit: D. Ramanan
How well do classifiers generalize?
How likely is a ‘typical’ pose?
How likely is a ‘typical’ pose?
Results

Slide credit: D. Ramanan
Results

Slide credit: D. Ramanan
Tasks

• Pedestrian detection
• Pose Recognition
• Tracking
• Activity Recognition
What are they doing?  
Recognizing Actions

Figure from Laptev et al. 2008
Common Issue: No vocab

Bobick & Davis. PAMI01

What’s right annotation?

Slide credit: D. Ramanan
Temporal scale and activity

- **Very short timescales**
  - not much happens
    - low dimensional models seem to work in animation
    - motion compresses well
  - but body configuration is diagnostic

- **Medium timescales**
  - Motions can be (at least):
    - sustained (running, walking, jogging, etc. --- typically periodic)
    - punctate (jump, punch, kick)
    - parametric (reach, etc.)

- **Long timescales**
  - Motions are complex composites
    - visiting an ATM
    - reading a book
    - cooking a meal
Recognizing Actions

Figure from Laptev et al. 2008

Slide credit: Derek Hoiem
Appearance

- Activities lead to characteristic patterns of image appearance
  - in grey level
  - in optic flow
Where you are is often a very powerful guide to what you are doing

Intille et al 95, 97

Slide credit: D.A. Forsyth
And can suggest you are doing what you should not be.

Boult et al 2001

Surveillance by omni-directional cameras, detection of anomalous pixel groups.
Numerous curious phenomena related to location

Yan + Forsyth 04
Particular activities often have characteristic appearance patterns. Braids appear at the legs of a walker.
The appearance of a silhouette can show whether a person is carrying something

Haritaoglu, Cutler, Harwood, Davis

Slide credit: D.A. Forsyth
Motion is a powerful cue at low resolution

Efros et al 03
Spatio-temporal volume is important

Blank et al 05
Working in a motion query framework relieves the need for a motion taxonomy. Features computed as before, we now seek sequences with small distances.

Blank et al 05