Tracking People by Learning Their Appearance

Developed by
Deva Ramanan, David Forsyth & Andrew Zisserman

Presenter: Jinyan Guan
09/10/2010
Road Map

• Motivation

• Approach Overview
  • Model representation
  • Model learning
  • Model detection

• Advantages

• Demo
Road Map

• Motivation

• Approach Overview
  • Model representation
  • Model learning
  • Model detection

• Advantages

• Demo
Why do We Want to Track People?

- Action recognition
- 3D pose estimation & reconstruction
Tracking People is Hard

- People move fast and unpredictably
- One can appear in variety of poses & clothes, and surrounded by limb-like clutter
Common Approach of Tracking

• Hidden Markov Model

\[ P(X_{1:T}, I_{1:T}) = \prod_{t} P(X_t | X_{t-1}) P(I_t | X_t) \]

\[ X_{1:T} = \{X_1, \ldots, X_t\} \]

• Tracking corresponds to inference on this HMM: Given a sequence of images, find the MAP sequence of poses.
Why Tracking by Learning the Appearance?

• Tracking by capturing the motion of people
  • What if the background moves rather than the people?

• An uninformative prior on motion (dynamics) models may cause the tracker to drift.

• Once the tracking fails, it has to be manually reinitialized.
Road Map

- Motivation:
- Approach Overview
  - Model representation
  - Model learning
  - Model detection
- Advantages
- Demo
“Tracking by Detecting”

Overview

• Step 1: Build a model of appearance of each person from a sequence of frames - learning the appearance

• Step 2: Track the person by detecting those models in each frame
Road Map

• Motivation:

• Approach Overview
  • Model representation
  • Model learning
  • Model detection

• Advantages

• Demo
Model representation

• How to represent people’s appearance?

• Pictorial Structure:
  • Model the human body as a puppet of rectangles
Temporal Pictorial Structure

from 1:t time

at time t

torso-lua assembly

Friday, September 10, 2010
Road Map

- Motivation:
- Approach Overview
  - Model representation
  - Model learning
  - Model detection
- Advantages
- Demo
Build the Models

- Bottom-up: group together candidate body parts found throughout a sequence of frames.
- Top-down: automatically build people-models by detecting convenient key poses within a single frame.
Bottom-up Approach: Clustering

- Looks for candidate in each frame
- Cluster the candidates to find assemblies of parts that might be people.
Clustering Steps

• **Detect** Candidate parts in each frame with an edge-based part detector

• **Cluster** the resulting image patches to identify body parts that look similar across time

• **Prune** clusters that move too fast in some frames and those do not move.
Learning a Model of Torso Appearance
Learning Multiple Appearance Models

Cluster

Candidate arms

Learned arm template

person model

Bryan model

John model

Deva model
Road Map

- Motivation:
- Approach Overview
  - Model representation
  - Model learning
  - Model detection
- Advantages
- Demo
Graphical Model

\[ P \left( P_{1:T}^{1:N}, I_{1:T} | C^{1:N} \right) = \prod_{t} \prod_{i} P \left( P_{t}^{i} | P_{t-1}^{i} \right) P \left( P_{t}^{i} | P_{t}^{\pi(i)} \right) P \left( I_{t} | P_{t}^{i}, C_{i} \right) \]

(a) Motion Model
(b) Spatial Kinematics
(c) Image Likelihood

---

Friday, September 10, 2010
Model Detection

• Finding an optimal track given a video sequence corresponds to find the MAP estimate of $C_t^i$ and $P_t^i$

• Exact inference is difficult because of loops and large state spaces of variables.

• Approximate inference: Ignore the loops and pass local messages
Approximate inference

- torso -> lower-arm -> upper-arm ...

\[ C^{\text{tor}} \]

\[ P_1^{\text{tor}}, P_2^{\text{tor}}, P_T^{\text{tor}} \]

\[ C^{\text{l1a}} \]

\[ P_1^{\text{l1a}}, P_2^{\text{l1a}}, P_T^{\text{l1a}} \]

\[ P_1^{\text{tor}}, P_2^{\text{tor}}, \ldots, P_T^{\text{tor}} \]

\[ P_1^{\text{l1a}}, P_2^{\text{l1a}}, \ldots, P_T^{\text{l1a}} \]
Building a Model of Arms and Legs
**Bottom-up Detection is Hard**

self occlusion  rare pose  motion blur

non-distinctive pose  too small  just right – detect this

Top-down Approach
Top-down Model: Building Models with Stylized Detectors

- Opportunistic detection
- Convenient poses:
  - 1) Easy to detect.
  - 2) Easy to learn appearance from, such as lateral walking.
Detect a Stylized Person Detector

- Use a single-frame pictorial structure model:
  \[ P \left( P^{1:N}, I \mid C^{1:N} \right) = \prod_{i}^{N} P \left( P^{i} \mid P^{\pi(i)} \right) P \left( I \mid P^{i}, C^{i} \right) \]

- \( P \left( P^{i} \mid P^{\pi(i)} \right) \): manually set the kinematic shape potential.

- \( P \left( I \mid P^{i}, C^{i} \right) \): use a chamfer template edge mask.
Lateral-walking Pose Finder
Discriminative Appearance Models

• Assume each limb detector is (more or less) color constant.

• Then we can train a quadratic logistic regression classifier in RGB space.
Tracking by Model Detection

• Given either model building method (bottom-up or top-down), we can build a representation (either a template patch or a classier) of a specific person.

• Multiple scales: The system searches this representation over an image pyramid.
An Overview
Road Map

• Motivation:

• Approach Overview
  • Model representation
  • Model learning
  • Model detection

• Advantages

• Demo
Advantages

• Track people with automatic initialization in front of complex backgrounds.

• Track people that standing in front of moving backgrounds.

• Two model-building algorithm are complementary.

• Initial detection can be done opportunistically.
Road Map

● Motivation:
  ● Why “tracking by learning appearance”?

● Approach Overview
  ● Model representation
  ● Model learning
  ● Model detection

● Advantages

● Demo
Demo
Demo
Demo
Questions?