CS131
Tracking people

A. Alahi
November 19th 2014
Understanding human behavior
Motivation: Elderly monitoring

Courtesy of Ph.D Guido Pusiol
Motivation:
Path-to-purchase understanding
Motivation: Space analytics

**Interaction**
- Number of social interaction

**Distances**
- Walked distance

**Durations**
- Duration in each room

**Heatmaps**
- Hot spots
Motivation:
Performance analysis
Motivation:
Behavior monitoring
Motivation:
Large-space analytics
Motivation: Large-space analytics

- Number of visitors
- Path of visitors
- Duration
- Rank spots

Large-space analytics
Moreover, how do shoppers behave?

- Number of visitors
- Path of visitors
- Duration
- Rank spots

Motivation:

- Maximize productivity
- Increase revenues
- Optimize staff
- Justify infrastructure change

Large-space analytics
Understand human mobility in a large terminal over a year

>120,000 commuters/day in 2013
>240,000 commuters/day by 2030*

Lausanne, Switzerland

* SBB CFF FFS report leman 2030
Track crowds in large spaces

Analytics
Flow in/out
Heatmap
Most used paths

Destination
Origin
Destination
Origin
A corridor with 14 Origin/Destination (O/D)
Collect long-term trajectories

# People | Av. duration | Av distance | Density (up to) | # Paths (O/D)
---|---|---|---|---
42 million | 1 minute | 100m | 1 pedestrian/m² | 196
Tracking people

- Thousands scientific publications about tracking people
- What is related to tracking people?
Tracking people

Focus of today’s lecture:
- Tracking people with
  - Video streams (vs static images)
    => Prior from the scene
    => Calibration data
    => Mapping from image plane to real-world
    => Model temporal variation/changes
Outline:

From Foreground Extraction To Tracking 42 million Pedestrians

Network of 3 sensors:
4 people walking (top view)

1. Detection intra-sensor
2. Tracklet Generation
3. Tracklet Association
4. Statistic Generation

- >100 GB/day
- 5 GB/day
- 4 GB/day
- 5 GB/day
- 0.05 GB/day
Outline:

From Foreground Extraction To Tracking 42 million Pedestrians

I. Detection
   I. Foreground extraction
   II. Pedestrian localization

II. Tracklet Generation
   I. Data association problem
   II. Matching appearance cues

III. Tracklet Association
   I. Modeling Social Affinities
I. Detection: Foreground extraction

- Severely degraded foreground silhouettes
- Spatially dense distribution
- Strong occlusions
I. Detection: Foreground extraction

\[ F(x,y, t+1) = \begin{cases} 
  1 \text{ if } |I(x,y, t) - B(x,y)| > T \\
  0 \text{ otherwise} 
\end{cases} \]

- Frame differencing
  \[ B(x,y) = I(x,y, t-1) \]

- Mean filter
  \[ B(x,y) = \frac{1}{N} \sum_{i=1}^{N} I(x,y, t-i) \]

- Gaussian averaging
- GMM
- ...

\[ => \text{Library of 32 algorithms (BGS library)} \]
I. Detection: Foreground extraction

- Severely degraded foreground silhouettes
- Spatially dense distribution
- Strong occlusions
I. Detection: Pedestrian localization
I. Detection: Calibrated Camera

- Create a dictionary $D$ of atoms approximating the ideal foreground silhouette for every position in $x$.
I. Detection: Sparsity driven framework

- Inverse problem:

\[ y = Dx + n \]
1. Detection: Greedy approach

\[ D \times x = y \]

I. Detection: Sparsity driven framework

- Inverse problem: 
  \[ y = Dx + n \]

- Sparsity prior: 
  \[ \min \| x \|_0 \quad \text{s. t.} \quad y = Dx + n \]
I. Detection: In praise of sparsity

“Creation is based on small number of primary, indivisible elements that combine with one another according to a few simple patterns.” [1]

I. Detection: Sparsity driven framework

- Inverse problem:
  \[ y = Dx + n \]

- Sparsity prior:
  \[ \min \| x \|_0 \quad \text{s.t.} \quad y = Dx + n \]

- Basis Pursuit [1]:
  \[ \min \| x \|_1 \quad \text{s.t.} \quad y = Dx \]

- BPDN:
  \[ \min \| x \|_1 \quad \text{s.t.} \quad \| y - Dx \| \leq n \]

- Lasso:
  \[ \min \| y - Dx \|_2 \quad \text{s.t.} \quad \| x \|_1 \leq \epsilon \]

I. Detection: Pedestrian localization

- Foreground extraction
- Used Dictionary
- Localization
Outline:

From Foreground Extraction To Tracking 42 million Pedestrians

Network of 3 sensors:
4 people walking (top view)

- *Detection intra-sensor*
  - >100 GB/day

- *Tracklet Generation*
  - 5 GB/day

- *Tracklet Association*
  - 4 GB/day

- *Statistic Generation*
  - 5 GB/day

- 0.05 GB/day
II. Tracklet generation: Data Association Problem

- Create a Directed Acyclic Graph $G = (N, E)$ where
  - $N$ = The detected ground plane points across time
  - $E$ = The connectivity cost between the detections (based on motion/appearance model)
II. Tracklet generation: Data Association Problem

II. Tracklet generation: Create a DAG

II. Tracklet generation: Select longest shortest path with smallest cost

II. Tracklet generation: Iterate

II. Tracklet generation: Iterate

II. Tracklet generation: Till no more paths


II. Tracklet Generation: Edge cost

- Motion model with social interactions
- Appearance model
II. Tracklet generation: Modeling social interactions

\[ \mathbf{F}_i = \mathbf{F}_{i}^{Goal} + \mathbf{F}_{i}^{Avoidance} + \mathbf{F}_{i}^{Attraction} + \mathbf{F}_{i}^{Scene} \]

\[ \mathbf{F}_{i}^{Avoidance} = \sum_{j \in P \setminus i} \mathbf{f}_{j \rightarrow i}^{Avoidance}, \]

where

\[ \mathbf{f}_{j \rightarrow i}^{Avoidance} = \alpha e^{-\frac{d_p - d_{ij}}{\beta}} \mathbf{n}_{j \rightarrow i} \]

\[ \frac{d}{dt} \mathbf{v} = \frac{\mathbf{F}_i}{m}, \]

II. Tracklet Generation: Modeling appearance cues
II. Tracklet Generation: An arm-race of image descriptors

- Vector of pixel intensities
  - Histogram of pixel intensities
  - Haar-wavelet responses
  - Gaussian derivatives
  - Complex filters
  - Steerable filters

- Distribution-Based Descriptor
  - HOG
  - SIFT
  - GLOH
  - Shape context
  - Spin images

- Moment invariants

- Differential Descriptor
  - Gabor-wavelet responses

- Spatial-Frequency-Based Descriptor
  - SURF

- Binary Descriptor
  - BRIEF
  - ORB
  - BRISK
  - FREAK

Low Performance

High Performance

II. Tracklet Generation: An arm-race of image descriptors

Vector of pixel intensities

Distribution-Based Descriptor

Histogram of pixel intensities

HOG

SIFT

GLOH

Shape context

Spin images

Moment invariants

Covariance of set of features

Differential Descriptor

Steerable filters

Gaussian derivatives

Complex filters

Spatial-Frequency-Based Descriptor

Gabor-wavelet responses

Haar-wavelet responses

Binary Descriptor

BRIEF

ORB

BRISK

FREAK

SURF

II. Tracklet Generation: HOG

Image gradient

- The gradient of an image: \( \nabla f = \left[ \frac{\partial f}{\partial x}, \frac{\partial f}{\partial y} \right] \)

\[ \nabla f = \left[ \frac{\partial f}{\partial x}, 0 \right] \]

The gradient points in the direction of most rapid increase in intensity.

The gradient direction is given by: \( \theta = \tan^{-1} \left( \frac{\partial f}{\partial y} / \frac{\partial f}{\partial x} \right) \)

- how does this relate to the direction of the edge?

The edge strength is given by the gradient magnitude:

\[ || \nabla f || = \sqrt{\left( \frac{\partial f}{\partial x} \right)^2 + \left( \frac{\partial f}{\partial y} \right)^2} \]

<table>
<thead>
<tr>
<th>Orientation</th>
<th>Magnitude</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>45</td>
<td>3</td>
</tr>
<tr>
<td>90</td>
<td>3</td>
</tr>
<tr>
<td>135</td>
<td>4</td>
</tr>
<tr>
<td>180</td>
<td>7</td>
</tr>
<tr>
<td>225</td>
<td>3</td>
</tr>
</tbody>
</table>

Source: Steve Seitz

Fei-Fei Li  Lecture 5 - 18  25-Sep-13
II. Tracklet Generation: An arm-race of image descriptors

- Vector of pixel intensities
  - Histogram of pixel intensities
  - Haar-wavelet responses
  - Gabor-wavelet responses
- Distribution-Based Descriptor
  - Moment invariants
  - Covariance of set of features
- Differential Descriptor
  - Steerable filters
  - Gaussian derivatives
  - Complex filters
- Spatial-Frequency-Based Descriptor
  - Gabor-wavelet responses
  - Haar-wavelet responses
- Binary Descriptor
  - BRIEF
  - ORB
  - BRISK
  - FREAK

II. Tracklet Generation: An arm-race of image descriptors

- **Vector of pixel intensities**
  - Histogram of pixel intensities
  - HOG
  - SIFT
  - GLOH
  - Shape context
  - Spin images
  - Moment invariants
- **Distribution-Based Descriptor**
  - Covariance of set of features
- **Differential Descriptor**
  - Steerable filters
  - Gaussian derivatives
  - Complex filters
- **Spatial-Frequency-Based Descriptor**
  - Gabor-wavelet responses
  - Haar-wavelet responses
  - SURF

Low Performance

High Performance

---

II. Tracklet Generation: Binary descriptors


---


A sequence of 1-bit DoG


- Select
  1) Most discriminant
  AND
  2) Less correlated

II. Tracklet Generation: Retina-inspired [1]

II. Tracklet Generation: Saccadic search

Object of interest

Target image to search

Matched objects

512 bits descriptor

Distance map

Distance map

Filtering with first 128 bits
Matching with last 128 bits

Tracking Example
Outline:

From Foreground Extraction To Tracking 42 million Pedestrians

Network of 3 sensors: 4 people walking (top view)

1. Detection intra-sensor
2. Tracklet Generation
3. Tracklet Association
4. Statistic Generation

- >100 GB/day
- 5 GB/day
- 4 GB/day
- 5 GB/day
- 0.05 GB/day
Tracklet association in scattered network

Network of 3 sensors:
4 people walking (top view)

Detection
intra-sensor

Tracklet
Generation

Tracklet
Association

Statistic
Generation

>100 GB/day

5 GB/day

4 GB/day

5 GB/day

0.05 GB/day
III- Tracklet association: Problem formulation

Let:

- $\mathbf{T}$: all long term trajectories
- $\mathbf{t}$: tracklets (tracklets capture within each camera)
- Problem: Maximizing the a posterior probability (MAP) of $\mathbf{T}$:

$$
\mathbf{T}^* = \arg \max_T P(\mathbf{T} | \mathbf{t})
= \arg \max \prod_i P(\mathbf{t}_i | \mathbf{T}) P(\mathbf{T}) \quad (1),
$$

where $P(\mathbf{T}) = \prod_k P(\mathbf{T}_k)$ (since trajectories do not overlap)

$$
P(\mathbf{T}_k) = P(\mathbf{t}^s_k) \ldots P(\mathbf{t}^{i-1}_k | \mathbf{t}^{i-1}_k) P(\mathbf{t}^e_k) \quad (markov \ chain)
$$
III. Tracklet association (Top view)
III. Tracklet association (Top view)

DAG = (V=tracklets,E= cost)

E=cost based:
- spatial,
- velocity
- travel time
- social affinity map
III. Tracklet association (Top view)

DAG = (V=tracklets,E= cost)

E=cost based:
- spatial,
- velocity
- travel time
- social affinity map
III. Tracklet association
(Top view)

\[ \text{DAG} = (V=\text{tracklets}, E= \text{cost}) \]

\[ E= \text{cost based:} \]
\[- \text{ spatial,} \]
\[- \text{ velocity} \]
\[- \text{ travel time} \]
\[- \text{ social affinity map} \]
Network flow optimization

Objective: minimum cost maximum flow

$$\arg \min_c c(f)$$

$$c(f) = \sum \alpha_i f_i + \sum \beta_{ij} f_{ij}$$

Where $\alpha_i, \beta_{ij}, \gamma_{OD}$ are the costs, and $f_i$ the flows
Network flow optimization

Objective: minimum cost maximum flow

$$\arg\min_{f} c(f)$$

$$c(f) = \sum \alpha_i f_i - \sum \beta_{ij} f_{ij}$$

Cost $\alpha_i$ based:
- Detection likelihood
Network flow optimization

Objective: minimum cost maximum flow

$$\arg\min_{f} c(f)$$

$$c(f) = \sum_{i} \alpha f_i + \sum_{ij} \beta_{ij} f_{ij}$$

Cost $\beta_{ij}$ based:
- spatial
- velocity
- Social Affinity Map
Tracklet association
With Social Affinity Map
Conclusion

A new dimension to “Google Analytics”: Analyzing people outside of website