CS131
Tracking millions of people

A. Alahi
November 8th 2013
What is related to tracking people?
Outline:

From foreground extraction To tracking millions of 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 millions of 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

- Severeley 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
- ...

=> Library of 32 algorithms (*BGS library*)
I. Detection: Pedestrian localization
I. Detection: Calibrated Camera

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

- Inverse problem:
  \[ y = Dx + n \]
I. Detection: Greedy approach

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: \( \min \| y - Dx \|_2 \quad \text{s. t.} \quad \| x \|_1 \leq \varepsilon \)

I. Detection: Pedestrian localization

- Foreground extraction
- Used Dictionary
- Localization

<table>
<thead>
<tr>
<th>$H_i$</th>
<th>$H_j$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha_1$</td>
<td>$\ldots$</td>
</tr>
<tr>
<td>$\alpha_2$</td>
<td>$\ldots$</td>
</tr>
<tr>
<td>$\ldots$</td>
<td>$\ldots$</td>
</tr>
<tr>
<td>$\alpha_k$</td>
<td>$\ldots$</td>
</tr>
</tbody>
</table>

Ground position $P_c(x_w, 0, z_w)$

Cuboid $S_x, S_y, S_z$
Outline:

**From foreground extraction To tracking millions of pedestrians**

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

1. Detection intra-sensor
   - >100 GB/day

2. Tracklet Generation
   - 5GB/day

3. Tracklet Association
   - 4 GB/day

4. Statistic Generation
   - 5 GB/day

- 50 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

- Distribution-Based Descriptor
  - Histogram of pixel intensities
  - Spin images
  - Shape context

- Moment invariants

- Differential Descriptor
  - Steerable filters
  - Gaussian derivatives
  - Complex filters

- Spatial-Frequency-Based Descriptor
  - Gabor-wavelet responses
  - Haar-wavelet responses

- Binary Descriptor
  - BRIEF
  - ORB
  - BRISK
  - FREAK

- Covariance of set of features

Low Performance

High Performance

II. Tracklet Generation: HOG

**Image gradient**

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

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

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

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

The gradient direction is given by:

\[
\theta = \tan^{-1}\left( \frac{\frac{\partial f}{\partial y}}{\frac{\partial f}{\partial x}} \right)
\]

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

The **edge strength** is given by the gradient magnitude:

\[
||\nabla f|| = \sqrt{\left( \frac{\partial f}{\partial x} \right)^2 + \left( \frac{\partial f}{\partial y} \right)^2}
\]
II. Tracklet Generation: An arm-race of image descriptors

- **Vector of pixel intensities**
  - Histogram of pixel intensities
  - HOG
  - SIFT
  - Shape context
  - Spin images
  - GLOH
  - Gaussian derivatives
  - Complex filters
  - Steerable filters
  - Moment invariants
  - Covariance of set of features

- **Spatial-Frequency-Based Descriptor**
  - Gabor-wavelet responses
  - Haar-wavelet responses

- **Differential Descriptor**
  - SIFT
  - SURF
  - BRIEF
  - ORB
  - BRISK
  - FREAK

II. Tracklet Generation: Binary descriptors

II. Tracklet Generation: BRIEF\([1]\) / ORB\([2]\) / BRISK\([3]\)  

- 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

Filtering with first 128 bits

Matching with last 128 bits

Outline:

From foreground extraction To tracking millions of pedestrians

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

1. Detection intra-sensor
   - >100 GB/day

2. Tracklet Generation
   - 5GB/day

3. Tracklet Association
   - 4 GB/day

4. Statistic Generation
   - 5 GB/day

- 0.05 GB/day

A. Alahi
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
Let:

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

\[
T^* = \arg \max_T P(T \mid t) = \arg \max \prod_i P(t_i \mid T)P(T)
\]  

(1)

(2),

where \( P(T) = \prod_k P(T_k) \) (since trajectories do not overlap)

\[
P(T_k) = P(t^s_k) \ldots P(t^t_k \mid t^{t-1}_k) P(t^e_k)
\]  

(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
Conclusion

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