Session-based biometrics for tracking people across disjoint camera views

Massimo Piccardi, Eric Dahai Cheng, Chris Madden
Faculty of Information Technology
University of Technology, Sydney (UTS)
{massimo, cheng, cmadden}@it.uts.edu.au
Tracking objects in videos

- Typical approaches to tracking objects in videos are based on some combination of:
  - motion coherency
    - motion history allows motion prediction (possibly within an ellipsoid of uncertainty)
  - appearance (texture) coherency
  - shape coherency
    - not obvious for deformable objects such as humans
Tracking objects in videos
Multi-camera tracking

• Multi-camera tracking requires tracking over more than one camera’s field of view (possibly, the entirety of the camera network)

• Similar to conventional tracking if
  – camera views overlap, or
  – camera views do not overlap but only for a small blind area -> motion prediction still acceptable
Disjoint camera views

- **Disjoint** camera views:
  - large blind areas between cameras
  - possibly long time between two consecutive views of a same object
  - great changes in apperance between two consecutive views of a same object

- A typical case for many real surveillance camera networks
Views: examples
Tracking across disjoint camera views

- Our goal: tracking single people across disjoint camera views
  - under the typical constraints of real surveillance camera networks

- Potentially useful for:
  - replacing intensive labour in determining the path of people who have been involved in an event of interest
  - detecting particular wanderabouts in real time
Session-based biometrics

To make this problem approachable, we introduced the concept of session-based biometrics

Physical features likely invariant within a session (i.e. entry to exit): clothes, hairdo, shoes...

Our set of visual features:
- clothes colour histogram
- hair colour histogram
- gait-filtered height
Matching two disjoint tracks

- Assumption: each object is reliably tracked within a single camera view
- $track = \text{associated state information (position, mask, texture for each frame)}$
- Problem: matching $track_i$ and $track_j$ for all eligible $i$ and $j$
Appearance-based matching

- Use of colour histograms: Major Color Spectrum Histogram Representation (MCSHR)

- Features:
  - in RGB as a whole (not on separate components)
  - color distance: magnitude-normalised L2 distance in the RGB space
  - variable number of bins, all of a same size in the normalised space (regular grid in RGB: way too many bins)
  - only the main bins retained, in frequency order to a fixed volume (typically 90%)
MCSHR: example

- main 15 colours
- colour distance threshold for bin creation (normalised space): 0.01
MCSHR: creation

• Two steps:
  – initial solution:
    • raster scan the object’s pixels
    • every time a pixel is outside any existing bins, create a new bin
  – K-means improvement
    • an online K-means algorithm is applied to refine the solution
    • typically 3 iterations
    • acceptable speed (more sophisticated EM algorithms: too slow)
Comparison of two MCSHRs

- Representations for objects $A$ and $B$:

\[
\text{MCSHR}(A) = \{C_{A_1}, C_{A_2}, \ldots, C_{A_i}, \ldots, C_{A_M}\}, \quad p(A) = \{p(A_1), p(A_2), \ldots, p(A_i), \ldots p(A_M)\}
\]

\[
\text{MCSHR}(B) = \{C_{B_1}, C_{B_2}, \ldots, C_{B_j}, \ldots, C_{B_N}\}, \quad p(B) = \{p(B_1), p(B_2), \ldots, p(B_j), \ldots, p(B_N)\}
\]

- Def.: the colour of $B$ closest to $C_{A_i}$ and within a given distance, $\sigma$:

\[
C_{B_j|A_i} : j = \arg \min_{k=1,\ldots,L} \{d(C_{B_k}, C_{A_i})\}
\]

\[
\text{MCSHR}'(B \mid C_{A_i}, \sigma) = \{C_{B_{i'}}, C_{B_{2'}}, \ldots, C_{B_{l'}}\}
\]
Comparison of two MCSHRs (2)

- Portion of $C_{A_i}$ in object A:
  
  $$p_{\text{norm}}(A_i) = \frac{p(A_i)}{\sum_{i=1,2,\ldots,M} p(A_i)}$$

- Portion of $C_{B_j|A_i}$ in object B:
  
  $$p_{\text{norm}}^{[A_i]}(B_j) = \frac{p_{[A_i]}(B_j)}{\sum_{j=1,2,\ldots,N} p(B_j)}$$

- Similarity between $C_{A_i}$ and $C_{B_j|A_i}$ (an arbitrary definition):

  $$\text{Sim}(C_{A_i}, C_{B_j|A_i}) = \min\{ p_{\text{norm}}(A_i), p_{\text{norm}}^{[A_i]}(B_j) \}$$
Comparison of two MCSHRs (3)

• Similarity between the whole objects $B$ and $A$:

$$Sim(B, A) = \sum_{i=1}^{M} Sim(C_{Bj}, C_{A|Bj})$$

• Due to construction of the representations, the above similarity measurement is obviously asymmetric; thus:

$$\rightarrow \text{Similarity}(A, B) = \min\{ Sim(A, B), Sim(B, A) \}$$

(NB: we do not want partial matches; alternative “symmetric” definitions are possible)
Matching between two tracks

- How to extend this approach to two entire tracks?

- Basically, two alternative approaches:
  - extending the representation to cover a whole track and perform a single, global matching operation
  - leaving the representation as is and perform repeated matching operations along successive frames
Post-matching integration

- Integration of the binary outcome of single-frame matching along the sequence
Incremental MCSHR

• Incremental color histograms (IMCSHR):
  – computing each color histogram over a small window of consecutive frames instead of a single frame
  – it averages/compensates for small pose changes in the person’s appearance

• We replace MCHSR with IMCHSR in the previous approach
Experimental results

- A same object in two separate tracks:

Matching results: 5 out of 5, average similarity = 0.9268
Experimental results (2)

- Two different objects from two separate tracks:

Matching results: 0 out of 6, average similarity = 0.3917

Compensating for varying illumination

- Use of cumulative color histograms (equalization)
  - seems to improve invariance while at the same time retaining discriminative power

- Use of intrinsic images? Perhaps
Shape-based matching

- Can we match two people from disjoint views based on their shape??

- Sophisticated 3D articulated models for human shape exist – fitting them to the required level of resolution on such a few pixels is unrealistic

- What shape features can be made invariant?
  - our answer: gait-filtered, perspective corrected height
Image point locations

Ground-truth salient points
(lateral view, after correction of radial lens distortion)
• Person A: tracks 1, 2, 3
• Person B: tracks 4, 5
• Camera 1: tracks 1, 2, 4
• Camera 2: tracks 3, 5
Height estimation

Calculated Track Heights

- Track 1
- Track 2
- Track 3
- Track 4
- Track 5
### Height estimation and matching

<table>
<thead>
<tr>
<th>Track</th>
<th>Average Height (cm)</th>
<th>Standard Deviation (cm)</th>
<th>Matching Tracks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Track 1</td>
<td>170.64</td>
<td>1.63</td>
<td>2, 3</td>
</tr>
<tr>
<td>Track 2</td>
<td>171.85</td>
<td>1.34</td>
<td>1, 3</td>
</tr>
<tr>
<td>Track 3</td>
<td>171.23</td>
<td>1.73</td>
<td>1, 2</td>
</tr>
<tr>
<td>Track 4</td>
<td>166.15</td>
<td>1.04</td>
<td>5</td>
</tr>
<tr>
<td>Track 5</td>
<td>166.64</td>
<td>0.92</td>
<td>4</td>
</tr>
</tbody>
</table>

Session-based biometrics: use

- Session-based biometrics can be fed together into any classifier
- Stringent real-time requirements -> use computationally-cheaper features for early dismissals of impossible matching
Conclusions

- Correct matching of single objects from disjoint camera views is possible.

- Session based-biometrics: a set of features capable of supporting such a matching.

- Appearance-based features:
  - MCSHRs (and IMCSHR) and their similarity are an effective way to compare appearance.
  - Post-matching integration mitigates the effect of typical segmentation errors.
Conclusions (2)

- **Shape-based features:**
  - Height measurement proves reliable
  - Initial camera calibration and ground-plane homography for correction of distortion (radial lens distortion, perspective)
  - Filtering “out” the gait effects through simple measurement averaging

- **Further testing**
The “Any questions?” slide

- Any questions?