Where’s Waldo: Matching People in Images of Crowds

Rahul Garg, Deva Ramanan, Steve Seitz, Noah Snavely

(Presented by Deepti Ghadiyaram)
Motivation

{ all photos }
Motivation

Question – How to browse such a collection and search for someone?
Problem Definition
Applications

1. Photo browsing
2. Surveillance
3. Content based querying / search
   - Richer search experience.
CHALLENGES
Challenges

Pose Change
Severe occlusion
Low resolution
Challenges (contd..)

Photos from 100s of users; different viewpoints

Different capture devices from different people.

Matching 100s of people where even faces are not clearly visible.

Challenges (contd..)

A particular “Waldo” appears in a small fraction of the entire collection.

Solution: Make Realistic Assumptions

1. People are relatively stationary over large intervals.
   Advantage? Multi View Stereo is applicable.

2. Images contain additional contextual information.
   – GPS tags, time stamps.
   – Social context.
   Advantage? Markov Random Field model is applicable.
MAIN CONTRIBUTIONS
Main contributions

1. Generalizing multi-view stereo to people-matching problem
   – NOT template matching
   – Use of a part-based appearance classifier instead of a window-based classifier

1) Generalizing multi-view stereo to people-matching problem.

<table>
<thead>
<tr>
<th>MVS</th>
<th>Waldo Problem</th>
</tr>
</thead>
<tbody>
<tr>
<td>Photo consistency through NCC etc.</td>
<td>Appearance consistency through a part based classifier</td>
</tr>
<tr>
<td>3D Localization</td>
<td>3D Localization with <strong>custom priors</strong></td>
</tr>
<tr>
<td>Smoothness in space via MRF</td>
<td>“Smoothness” over time and people via MRF</td>
</tr>
</tbody>
</table>

Main contributions (contd..)

2) Exploiting contextual-cues via MRF
   – Co-occurrence of people
   – Timestamps.

3) Making an extensively labeled dataset available.
METHOD OVERVIEW
Step#0

Image Collection of an event

Register the Photo Collection using SFM

Structure From Motion

http://grail.cs.washington.edu/projects/cpc/
Learn Part Based Appearance classifier

Estimate the 3D Location of the person

Search for the person in the entire image collection

Refine search using MRF optimization

User Input

Results
Learn Part Based Appearance classifier

Estimate the 3D Location of the person

Search for the person in the entire image collection

Refine search using MRF optimization

User Input

Results
User Input

• Input – Single instance of each person to be searched \( (p_i) \)

• Effective since the pose variation is implicitly captured.
Part specific Color Model

Challenges:
- View point
- Scale
- Exposure
- Occlusion

\[
w_{part} = \arg\min_w \sum_j \log(1 + \exp(-y_j w^T x_j)).\]
Scoring a candidate match

- Isotropic Transform
- Part-specific classifier
- Binary Classification Mask (per part)
Scoring a candidate match

\[ \sum S_{part_k} \]

Sum of the number of positively classified pixels inside a specific part \( k \)

Summation accounts for occlusion.

*Some parts are more discriminating than others.*

Score =

\[ \begin{cases} 
0 & \text{if } \sum S_{part_1} = 0 \\
\sum_{k=1}^{k'} \sum S_{part_k} & \text{otherwise} 
\end{cases} \]
Discussion

• Very high dependence on the lighting conditions.
  – *Normalize* the RGB values in the appearance model?
  – HSV space or a *different color space*?

• Performance on a similarly dressed crowd images. Eg: Convocation ceremony.
  – Requires additional cues beyond appearance.

• *Face detection* during appearance modeling (when applicable)

• *Soft threshold* on the appearance score rather than a hard threshold as it is now.
Learn Part Based Appearance classifier → Estimate the 3D Location of the person

User Input

Search for the person in the entire image collection

Refine search using MRF optimization

Results
3D Localization

Assumption: Orientation of the person is along the vertical.

- Searching in 1-D for $P_{iground}$
Learn Part Based Appearance classifier

Estimate the 3D Location of the person

Search for the person in the entire image collection

Refine search using MRF optimization

User Input

Results
3D Localization (contd..)

- For each candidate pair in $(P_{i_{head}}, P_{i_{ground}})$
  
  Project it into all the images (timestamp constrained)
  - Get $(p^k_{i_{head}}, p^k_{i_{ground}})$

  Score the projection using appearance model $S_i$

  $$\sum_{I_k \in A} \max(S_i(p^k_{i_{head}}, p^k_{i_{ground}}) - \text{thresh}, 0)$$
Wiggle search

The score is multiplied by height and ground priors.

Search window of size 2hX2h

h: projected height of the candidate location

MAX(S1,S2,S3)
When orientation of the person is not vertical.

\[ P_{i\text{head}} \] should be marked at a sitting height (Sitting prior)
Learn Part Based Appearance classifier → Estimate the 3D Location of the person

Search for the person in the entire image collection

Refine search using MRF optimization
Contextual Cues

1. People appear *together* with the same group of people.

2. Images which are *nearby in time* are likely to contain...
People who are likely to appear together in an image

A person in a set of similar images.

Minimize Objective Function

$$E(\mathcal{I}) = \sum_{ij} U(l_{ij}) + \sum_{ij} \sum_{i'j'} \phi(l_{ij}, l_{i'j'})$$

Unary Potential

Pairwise potentials
Discussion

• For the MRF model to be applicable, is every person, in every image, every time?
  – (OR) Is every person in the training image identified?

• Cues hallucinate the person when not present if other people with high affinities with that person are detected in the image.
  – Wont the appearance score be zero for this missing person?
Learn Part Based Appearance classifier

Estimate the 3D Location of the person

Search for the person in the entire image collection

Refine search using MRF optimization

User Input

Results
Datasets & Ground Truth Data

• Dataset#1
  – 34 photos; single photographer; Trafalgar Square; single day.

• Dataset#2
  – 282 photos; 89 different photographers; Trafalgar Square; single day.

• Dataset#3
  – 45 photos from 19 different users taken; Hackday; over two days. (Indoor)

• Ground truth labeling
  – Manually labeled with assistance from geometry
  – Does not follow the contextual cues.
Results – Dataset#1

- Pose change
- Occlusion

False Negatives

Pose change

Occlusion
Results – Dataset#2
Illustrating failure to identify matches

- Torso (Red) not distinct from the background.
- Blue – too many colors.
Extensions

• Relaxing each of the assumptions made.
  – Allow large motion of people.

• Track people’s movement through the scene.

• More powerful and accurate appearance models.

• Larger image datasets.
Understanding Images of Groups of People