Where's Waldo: Matching People in Images of Crowds

Rahul Garg — University of Washington
Deva Ramanan — University of California, Irvine
Steven M. Seitz — University of Washington, Google Inc.
Noah Snavely — Cornell University

Problem Definition

Given a person in a single image, find all other photos of that person from a collection of community photos of the same event.

Our approach finds 4 of the 5 matches shown above

Challenges

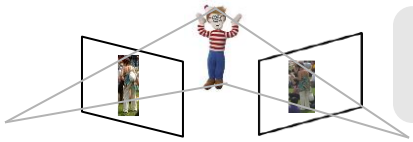

Non-rigid Pose Change


Severe Occlusion


Low Resolution


Photos from hundreds of users, Different viewpoints

Key Idea



Generalization Of Multi View Stereo (MVS)

Assumptions: Known camera pose, small person movement over short time interval

MVS	Waldo Problem
Photoconsistency through NCC, etc.	**Appearance Consistency through a custom classifier**
3D localization	3D localization with **custom priors**
Smoothness in space via MRF	**"Smoothness" over time and people via MRF**


Single labeled instance

Learn Appearance Classifier → 3D Localization → MRF Refinement →


Matches

Details of the Approach

Appearance Classifier :

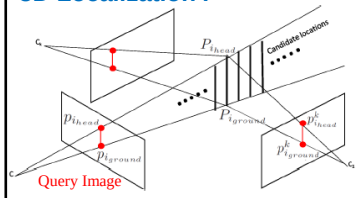

User input on query image

User Input – p_{head}, p_{ground}, masks for head, torso and legs on a single query image. Learn pixel level RGB classifier using logistic regression for the three parts

Scoring a candidate: Align the candidate with the template. Run the part classifiers and sum the pixel classification weighted using part masks


Align candidate match with query image and classify pixels part classifiers

3D Localization :



Camera pose from Structure from Motion

Propose candidate locations by backprojecting rays from query image. Project candidates into other images and score using learnt classfier

Height Prior: Prior on average height of a person

Ground Prior: Encourage backprojection of p_{ground} to be close to the ground plane in 3D

MRF Refinement :

Choose 3D location with highest score for each person. Project into each image and decide which projections are true matches. Use **co-occurrence** and **time** cues.



Co occurrence and Time Cues: People appear with the same group of people. Images nearby in time are likely to contain the same set of people.

MRF Model:

Node for every person-image pair, (p_i, I_j). Solve for a binary labelling where label = 1 if p_i occurs in I_j

Unary term: Appearance classifier score

Add edges between people with weights determined by *people affinity*, edges between images with weights determined by *image affinity*

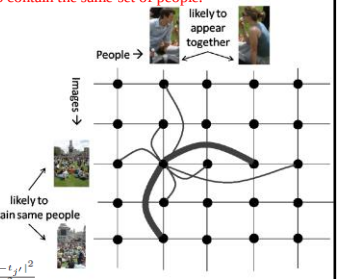


Image Affinity: $\alpha_I(I_j, I_{j'}) = \lambda_1 e^{\frac{-|t_j - t_{j'}|^2}{2\sigma_I^2}}$ where t_j is the corresponding time stamp

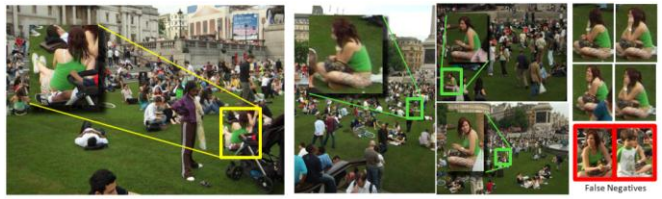
People Affinity: $\alpha_p(p_i, p_{i'}) = \lambda_2 \frac{|D_i \cap D_{i'}|}{|D_i| + |D_{i'}|}$ where D_i is the set of images that contain p_i.

Solve MRF iteratively updating D_i each time.

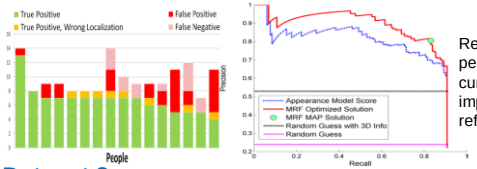
Results

All datasets downloaded from Flickr and manually matched with assistance from geometry

Dataset 1: 34 photos taken by a single photographer at Trafalgar Square on a single day. 16 different people to match, 130 total matches



Sample matching result for one person: 7/9 matches found. The query image was a back pose while the found matches are all side poses. There are two missed matches, one with extreme pose change and the other with severe occlusion.



Results for individual people and precision recall curve showing the improvement from MRF refinement.

Dataset 2: 282 photos taken by 89 different photographers at Trafalgar Square on a single day. 57 people, 244 total matches.

A representative result: 6/7 matches found are correct. One of the missed matches has extreme occlusion and the false positive is due to presence of a similar color. Precision Recall curve on the right.

Dataset 3: 45 photos from 19 different users taken during an indoor event – *Hackday London 2007* over two days. 16 people, 56 matches.

All 5 matches are found. Note that the laptop is not visible in the query image.

Conclusion

- Very hard problem made tractable by simplifying assumptions: Known camera pose, relatively static people
- Relax assumptions in future: "track" people from photos, use stronger appearance cues in photos with unknown camera pose
- Lack of datasets presently – will change with more cameras and more photo sharing