# Where's Waldo: Matching People in Community Photo Collections

CVPR 2011

Rahul Garg, Deva Ramanan,
Steve Seitz, Noah Snavely

Slides from Rahul Garg


I MADE A SIGN
VOTE SANITY 11.02.10
MY SIGN IS BETTER THAN HIS

# Can you find them in ***this*** photo?






**Where's Waldo?**

***Manually*** matched photos from a collection of 282

***Automatically*** matched photos from a collection of 282

**Problems:** Severe occlusion, Varying poses and cameras, Low resolution

Non-rigid
Pose Change

Severe
Occlusion

Low
Resolution

Photos from hundreds of
users, Different viewpoints

**Problems:** Severe occlusion, Varying poses and cameras, Low resolution

**Scene Geometry**
And
**Camera Pose**

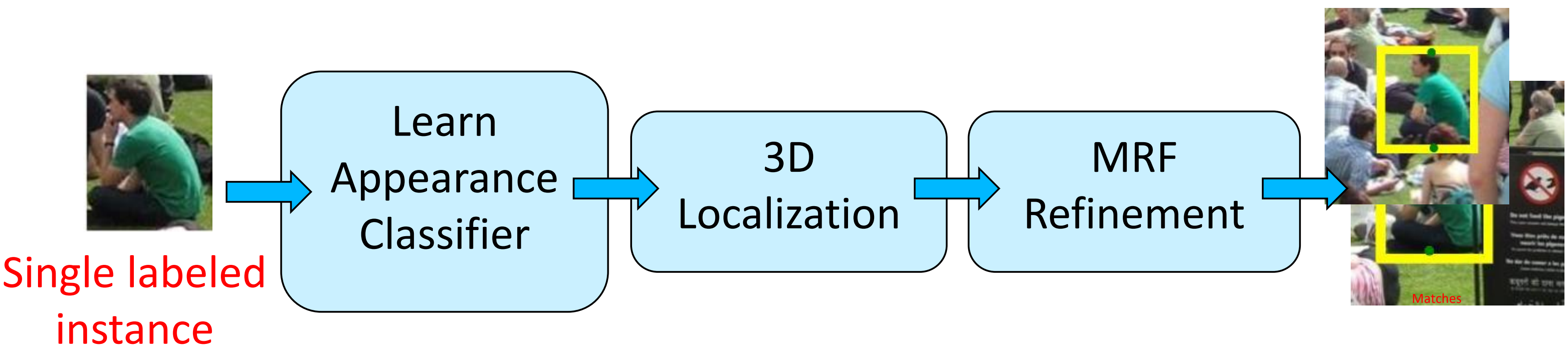
**Part based appearance model**
based on cloth color

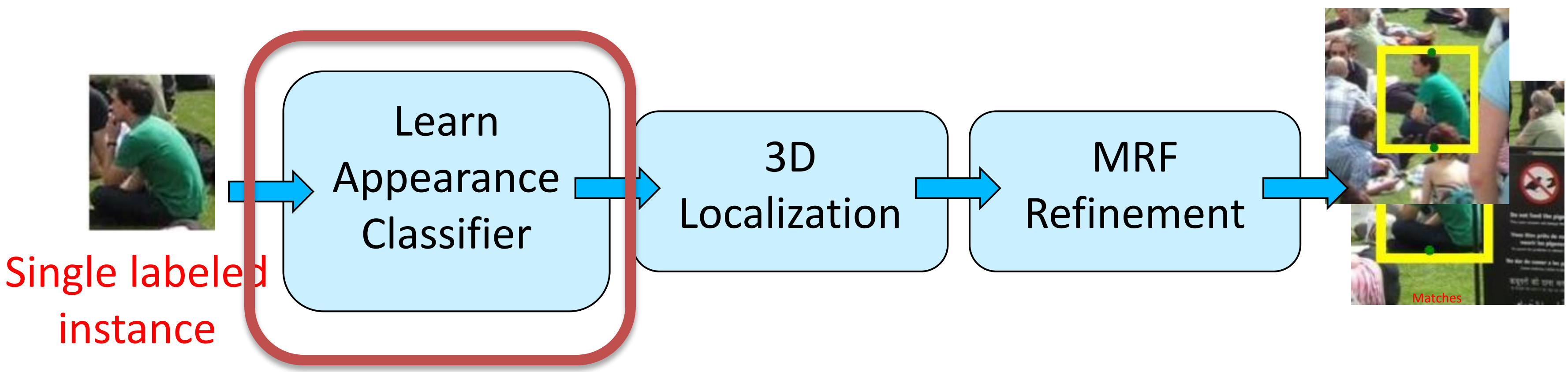
**Detection**
**and**
**Recognition**

**Time stamps**: people are
relatively still
over short intervals

**Co-occurrence cues**: People move
about with
same set of people

# Details of the approach

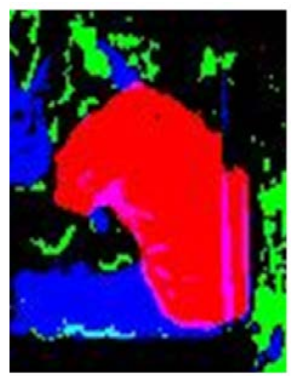

Single labeled
instance
Learn
Appearance
Classifier
3D
Localization
MRF
Refinement
Matches


Single labeled instance
Learn Appearance Classifier
3D Localization
MRF Refinement
Matches

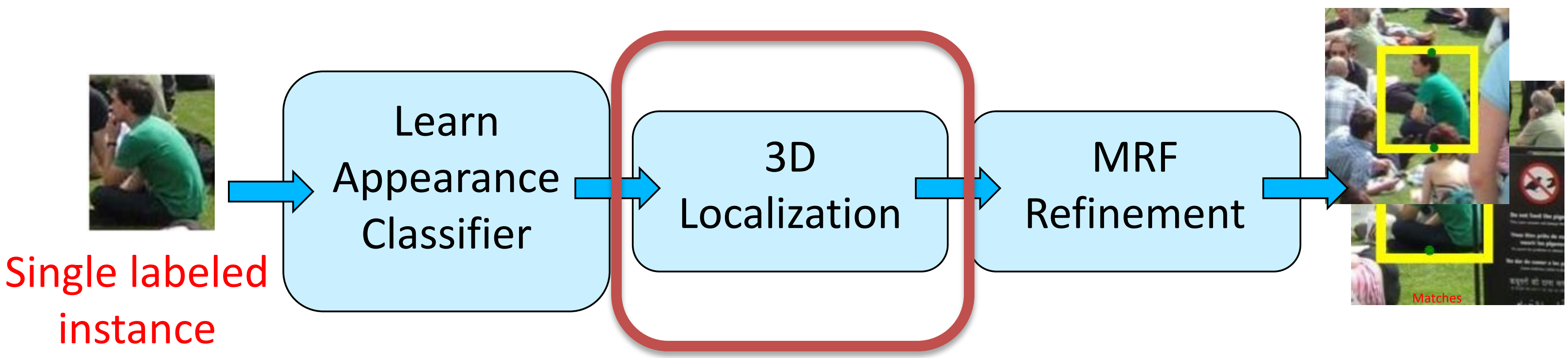
# Appearance Classifier


**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.

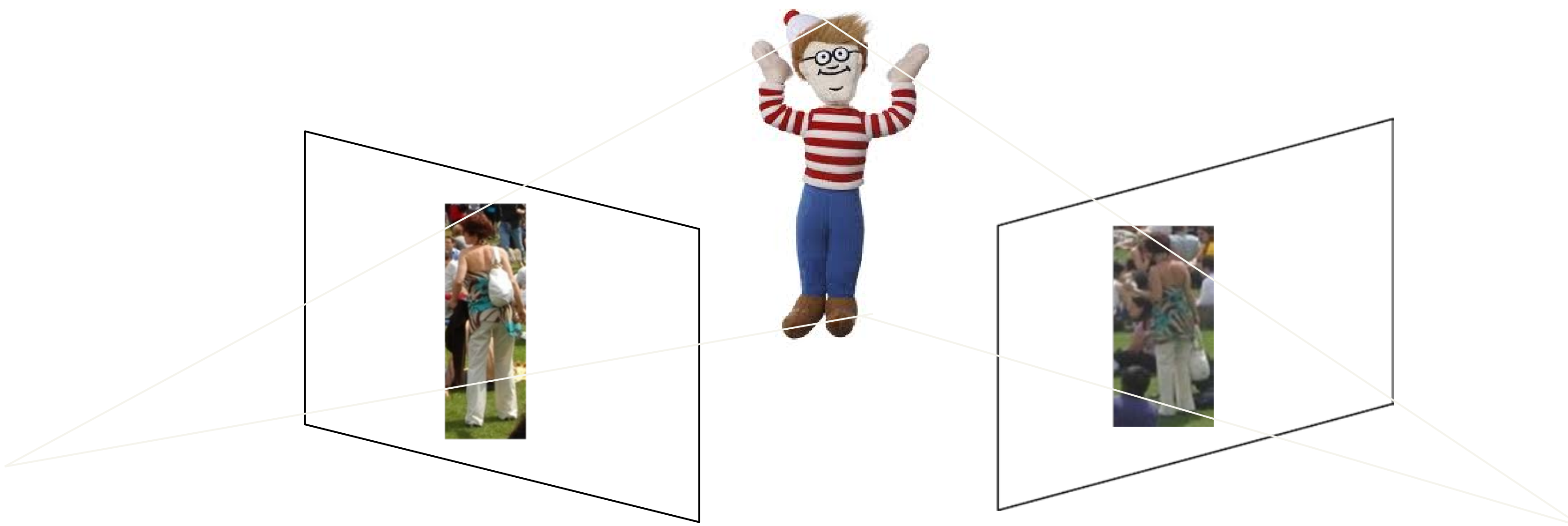



**Scoring a candidate:** Align the candidate with the template. Run the part classifiers and sum the pixel classification weighted using part masks.



Single labeled instance
Learn Appearance Classifier
3D Localization
MRF Refinement
Matches

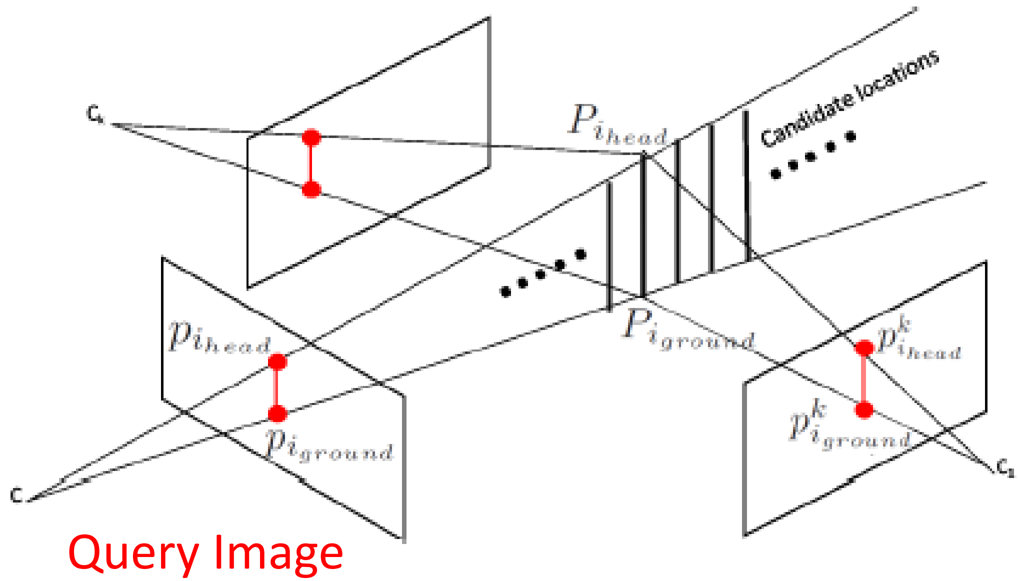
# Key Idea
# Generalization of Multi View Stereo (MVS)



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

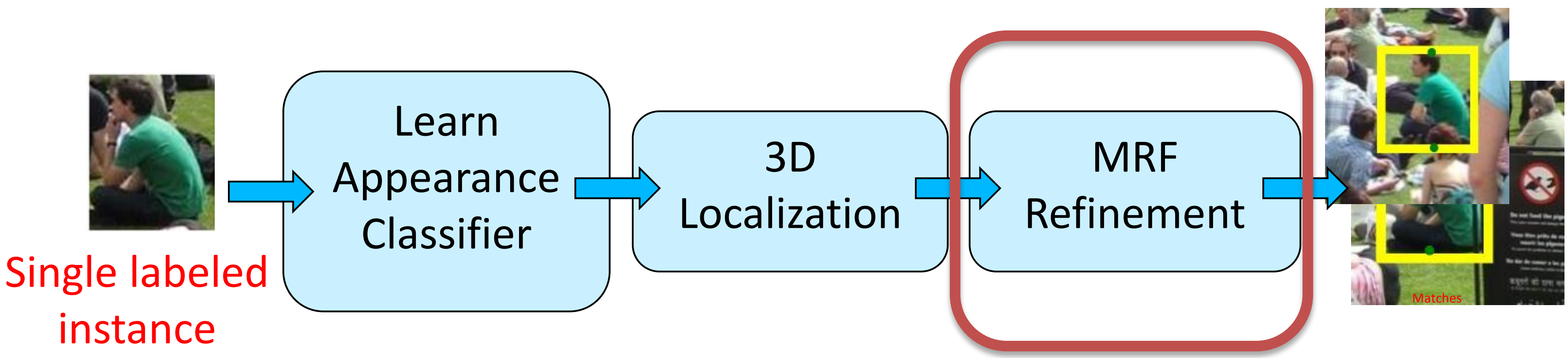
# 3D Localization

Propose candidate locations by back-projecting rays from query image. Project candidate locations into other images and score these images using learnt classifier.




**Height Prior:** Prior on average height of a person.

**Ground Prior:** Encourage back-projection of $p_{ground}$ to be close to the ground plane in 3D.


Single labeled instance
Learn Appearance Classifier
3D Localization
MRF Refinement
Matches

# Markov Random Field 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 people.

# The Markov Random Field Model

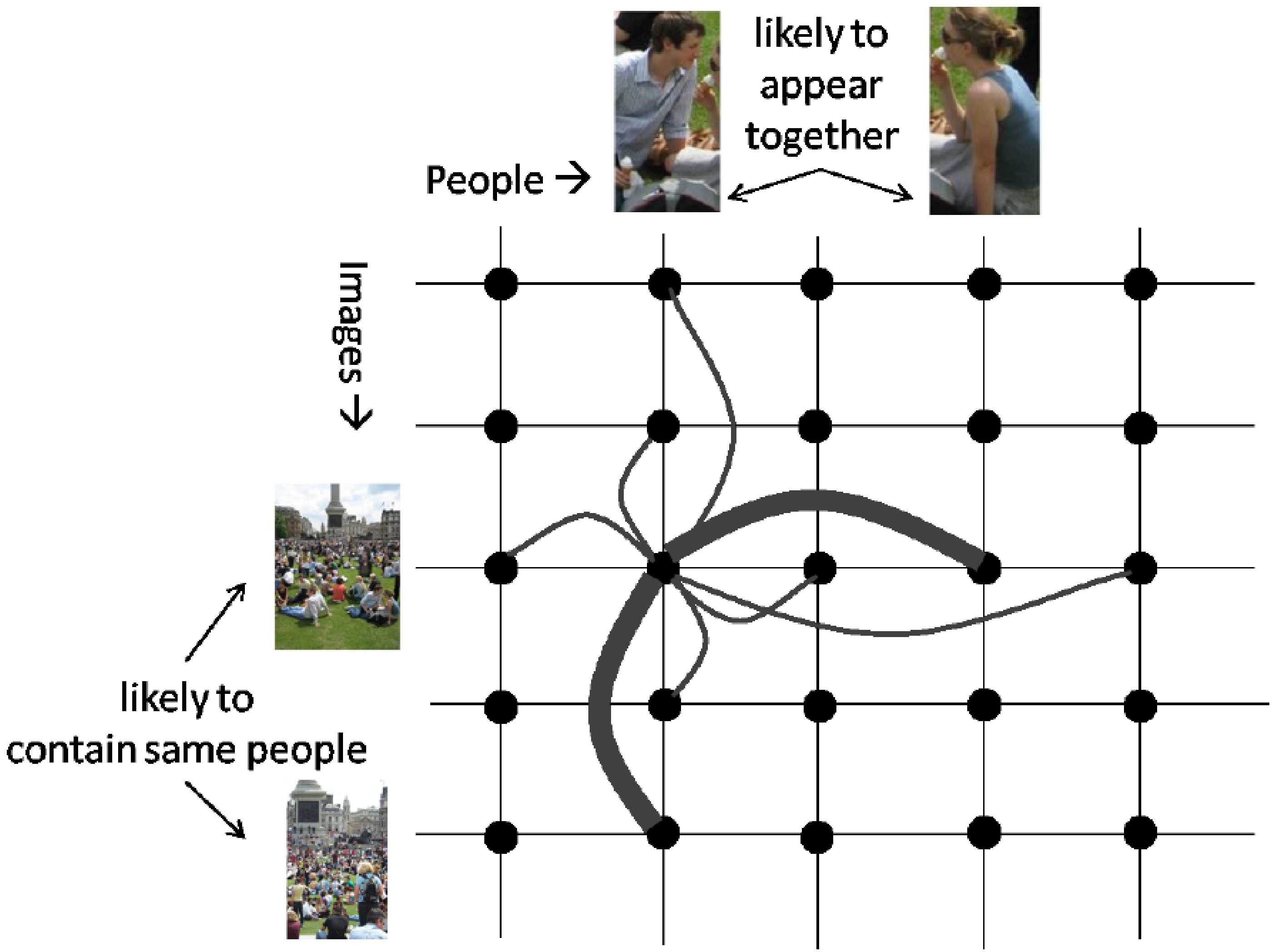
Node for every person-image pair, $(p_i,I_j)$.

Solve for a binary labelling where label = 1 if $p_i$ occurs in $I_{j.}$

Add edges linking people who appear together and between images that are close in time.

Use graph cuts to select the best candidate matches for each person.

# The Markov Random Field Model



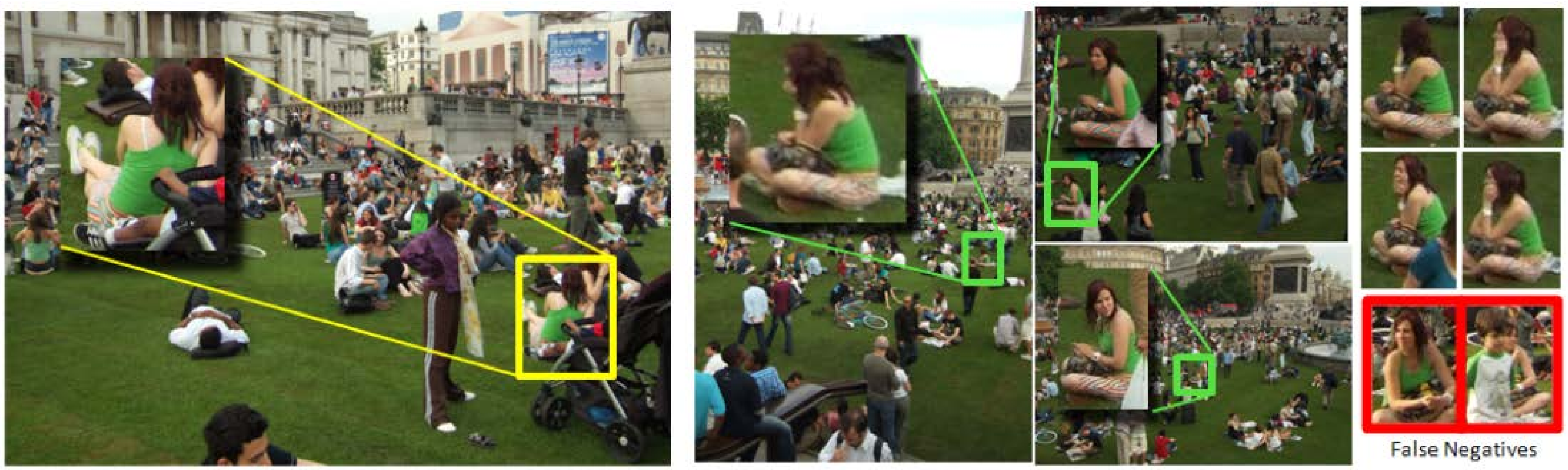
# Results

# Ground Truth

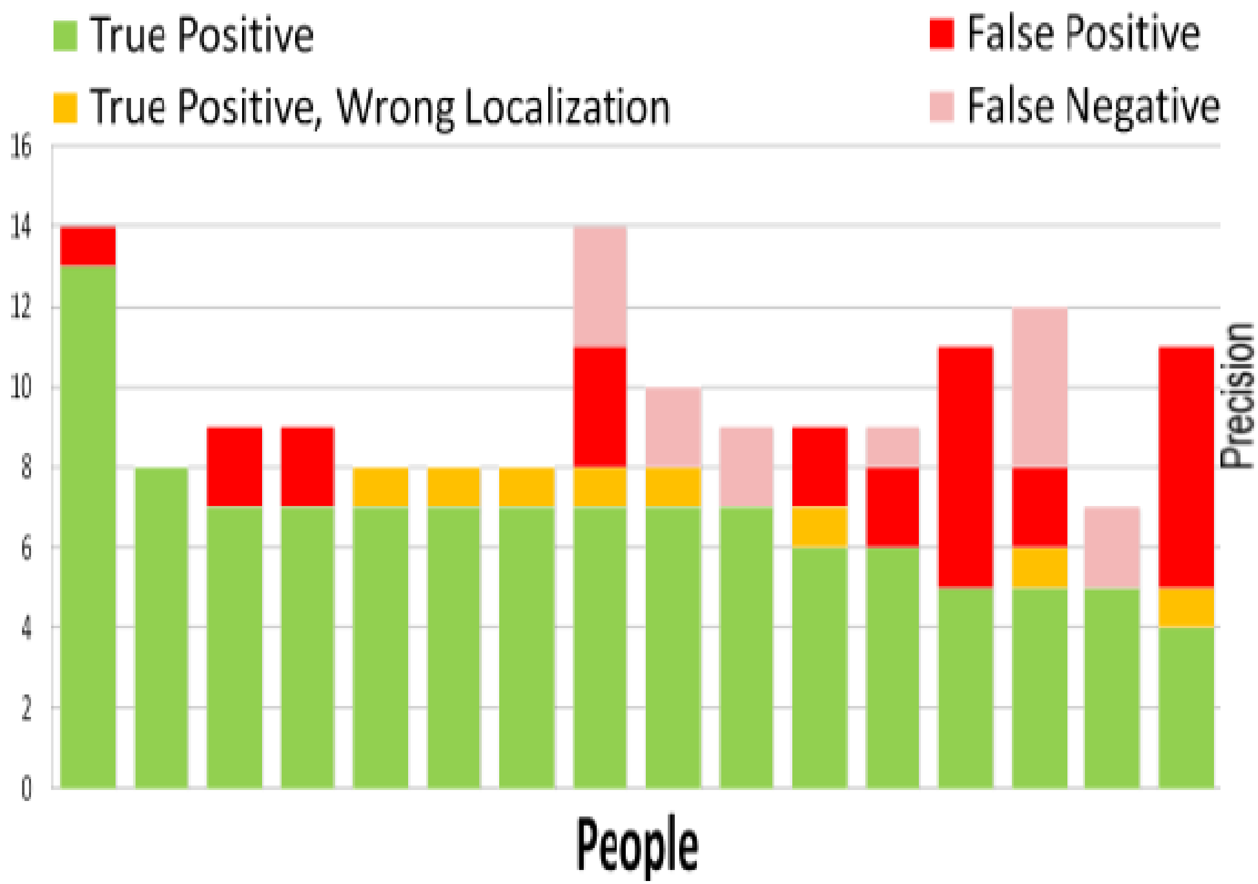
All datasets downloaded from Flickr and manually matched.

# Results: Dataset 1

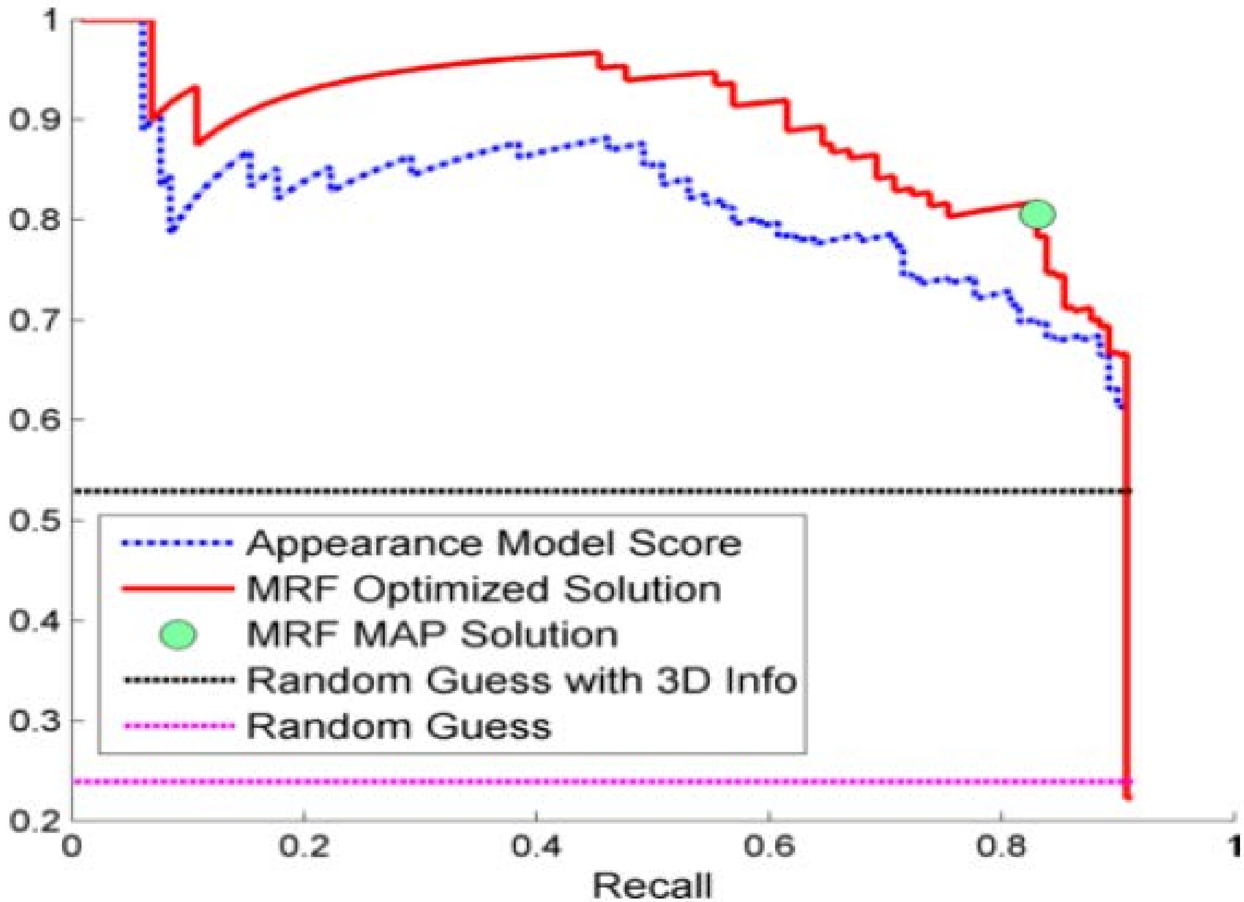
**34 photos** taken by a **single photographer** at Trafalgar Square on a single day. 16 different people to match, 130 total matches.




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.


True Positive
True Positive, Wrong Localization
False Positive
False Negative
16
14
12
10
8
6
4
2
0
Precision
People

precision

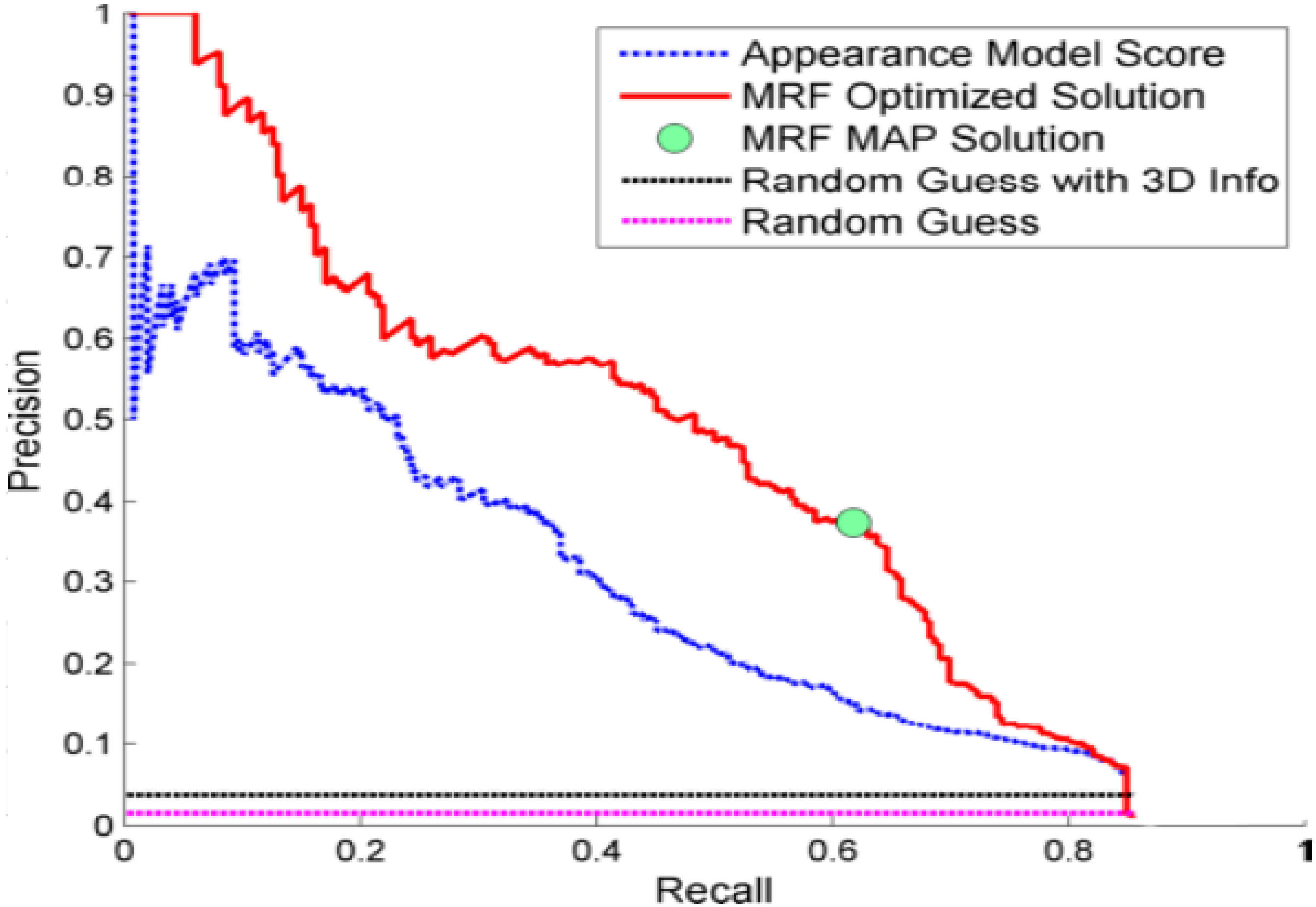


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0
0.2
0.4
0.6
0.8
1
Recall
Appearance Model Score
MRF Optimized Solution
MRF MAP Solution
Random Guess with 3D Info
Random Guess

# Results: Dataset 2

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

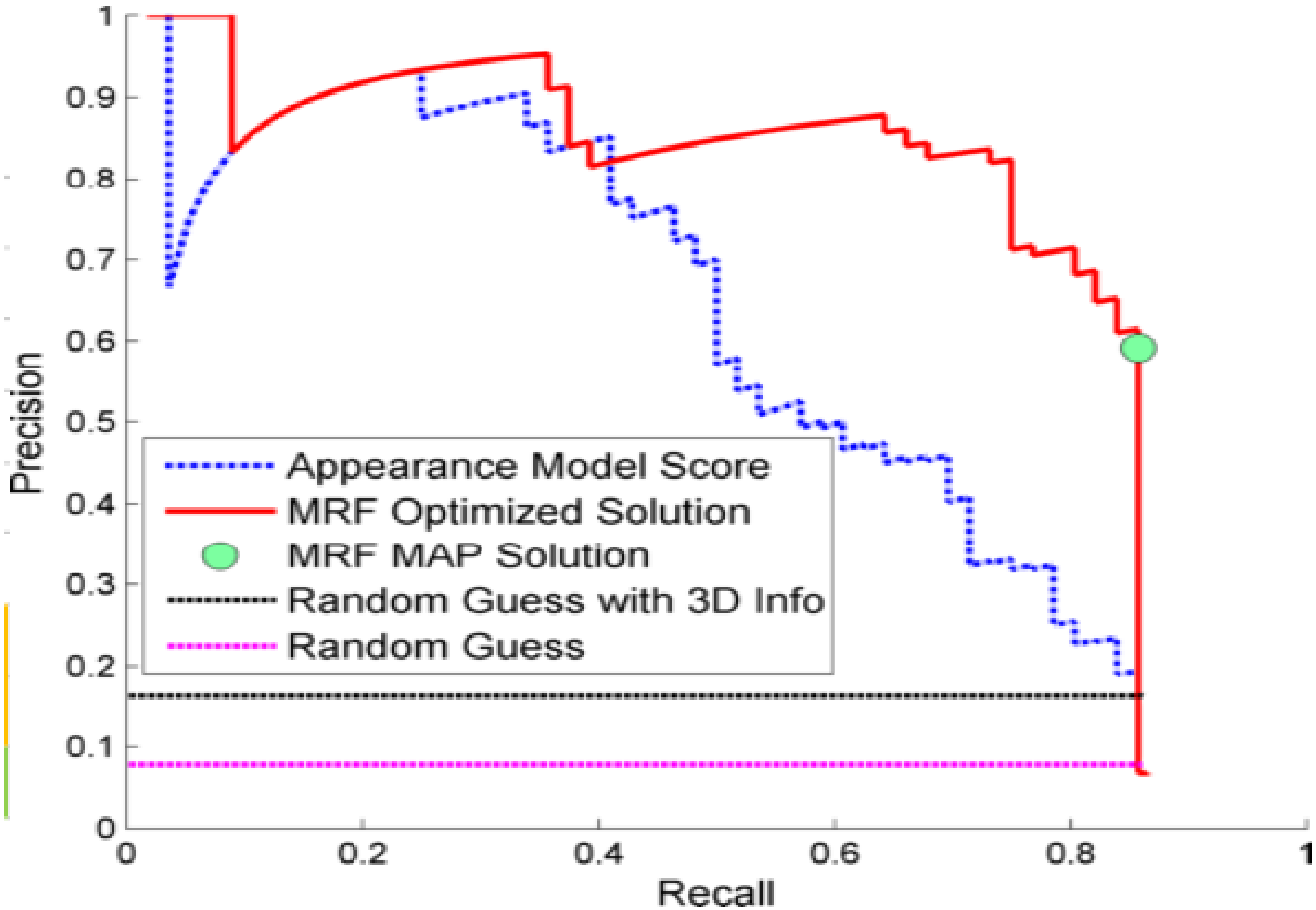

A representative result: 6/7 matches found are correct. One of the missed matches has extreme occlusion and the false positive is due to presence of a similar color.


Appearance Model Score
MRF Optimized Solution
MRF MAP Solution
Random Guess with 3D Info
Random Guess
Precision
Recall
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.2
0.4
0.6
0.8
1

# Results: 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.



Precision
Recall
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.2
0.4
0.6
0.8
1
Appearance Model Score
MRF Optimized Solution
MRF MAP Solution
Random Guess with 3D Info
Random Guess

# Conclusions

- 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

# Discussion

# Relaxing restrictions

How might we relax some of the restrictions in order to work with more diverse image sets?

1. How to handle moving people?
2. Horizontal (sleeping) people?
3. People on stairs, stages, pyramids?
4. Dramatic changes in pose?
5. Changes in clothing?
6. Lighting changes?

# Relaxing restrictions

How might we relax some of the restrictions in order to work with more diverse image sets?

- *Could you use A* or other iterative search techniques to broaden searches adaptively?*
- *Identify a separate set of matches after big changes?*

# Improvements

Are there other features or strategies that might improve the model?

# Improvements

Are there other features or strategies that might improve the model?

- *Augmenting the appearance model with facial recognition*
- *Texture recognition*

# User interactions

What interactions might you build on top of this?

# User interactions

What interactions might you build on top of this?

- *How to navigate a photo set based on correspondences?*
  - *Could you interactively pivot between photos by selecting individual people?*
  - *Center or highlight people of interest in a PhotoSynth-style environment and focus on the images that contain them?*
  - *Display many images of a target individual as small multiples?)*
- *Could you use user feedback to improve results?*

# User interactions

Could you use user feedback to improve results?

# User interactions

Could you use user feedback to improve results?

- *Contextual cues encourage people with high affinities to share detections among them. A side effect is that false positives and false negatives are also shared. More user interaction may be helpful here, i.e., correcting a match for a single person may correct it for a number of other people as well.*

# Other domains

How might you alter this approach to work for other applications?

# Other domains

How might you alter this approach to work for other applications?

- *Cars?*
- *Wildlife tracking?*
- *etc.?*

# Applications

What applications does the paper suggest?

Do these seem useful?

Can you think of other applications?

# Applications

What applications does the paper suggest?

Do these seem useful?

Can you think of other applications?

- *Photo tagging*
- *Navigating photo archives*
- *Surveillance*

| 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 Markov Random Field |

# The Markov Random Field Model

Node for every person-image pair, $(p_i, I_j)$.

Solve for a binary labelling where label = 1 if $p_i$ occurs in $I_j$.

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

**Image Affinity**: $\alpha_I(I_j, I_{j'}) = \lambda_1 e^{\frac{-|t_j - t_{j'}|^2}{2\sigma_t^2}}$

where $t_i$ 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.