PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization

Alex Kendall, Matthew Grimes, and Roberto Cipolla - [ICCV 2015]

Presented by:
Kent Sommer

Outline:

- Motivation / Related work
- Problem Statement / Overview of approach
- Dataset
- Details and issues with approach
- Results
- Conclusion / Quiz

Review and Related Work

Review:

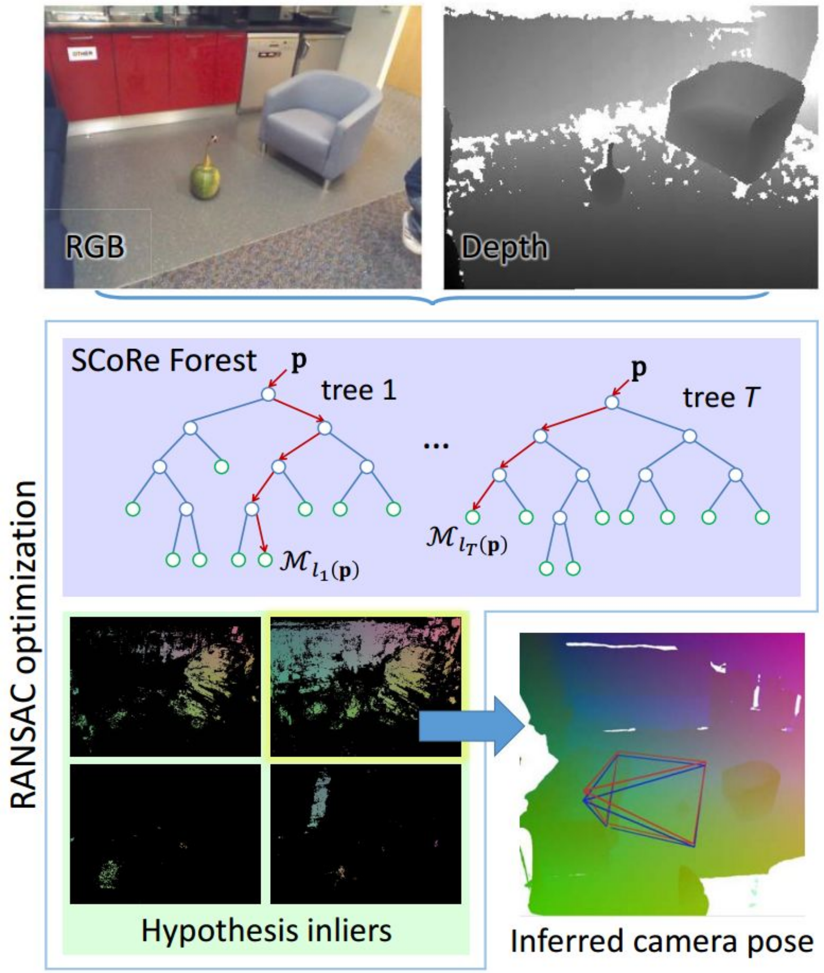
- Two approaches to localization
  - Metric
    - Estimate continuous position
  - Appearance/Topological
    - Classify scene to limited number of discrete locations

What does this have to do with search?

- Appearance/Topological localization can be presented as a search problem!
  - Database of known locations, given an input image, where are we?
    - Efficient retrieval is necessary, usually really large database

Related Work:

- Scene Coordinate Regression Forests
  - Use depth images to map each pixel from camera to global
  - Train a regression forest to regress these labels given an RGB-D image.
  - Limited to indoor use in practice (IR interference)



Related Work:

- Feature extraction and matching as in [1, 2, 3, 4]
  - (Generally) extract various types of image features
    - Match these features with those in the database with tagged known location to return position

[1] J. Wang, H. Zha, and R. Cipolla. Coarse-to-fine vision-based localization by indexing scale-invariant features. Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, 36(2):413–422, 2006.

[2] Y. Li, N. Snavely, D. Huttenlocher, and P. Fua. Worldwide pose estimation using 3d point clouds. In Computer Vision– ECCV 2012, pages 15–29. Springer, 2012.

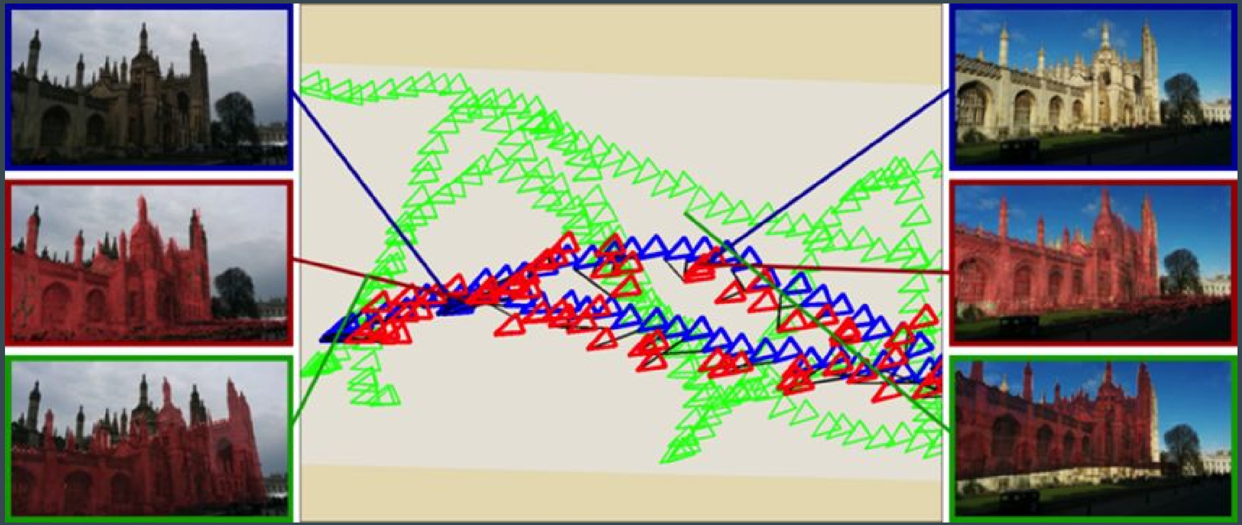
[3] Q. Hao, R. Cai, Z. Li, L. Zhang, Y. Pang, and F. Wu. 3d visual phrases for landmark recognition. In Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, pages 3594–3601. IEEE, 2012.

[4] A. Bergamo, S. N. Sinha, and L. Torresani. Leveraging structure from motion to learn discriminative codebooks for scalable landmark classification. In Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, pages 763– 770. IEEE, 2013.

Problem Statement and Overview of Approach

Problem Statement:

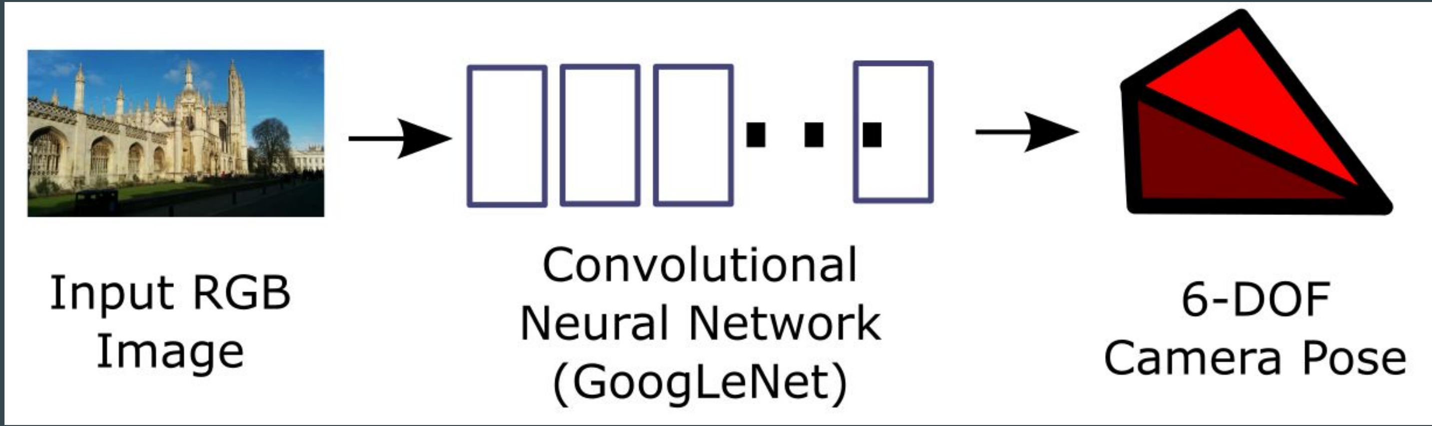
- Estimate the 3D position and orientation of the camera, given a single monocular image taken from a large previously explored area



- Green
  - Training
- Blue
  - Testing
- Red
  - System output

Overview of Approach:

- Perform end-to-end supervised learning with euclidean loss to regress 6-DOF pose.
  - Does not require large landmark database (instead it learns robust high level features to regress 6-DOF pose.)



Dataset

Dataset:

King's College *Street* *Old Hospital* *Shop Façade* *St Mary's Church*

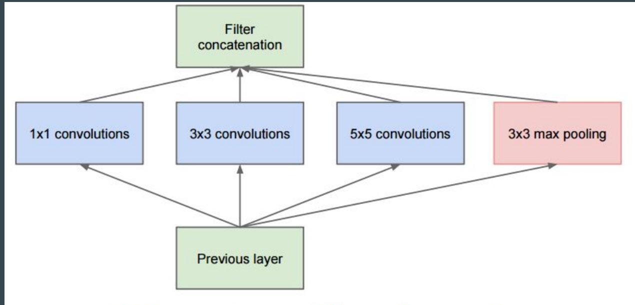
Figure 4: **Map of dataset** showing training frames (green), testing frames (blue) and their predicted camera pose (red). The testing sequences are distinct trajectories from the training sequences and each scene covers a very large spatial extent.

Figure 5: **7 Scenes dataset** example images from left to right; Chess, Fire, Heads, Office, Pumpkin, Red Kitchen and Stairs.

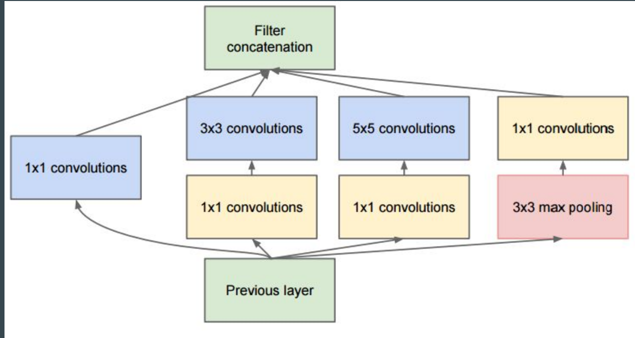
Details and Issues with Approach

Details of Approach (Neural network):

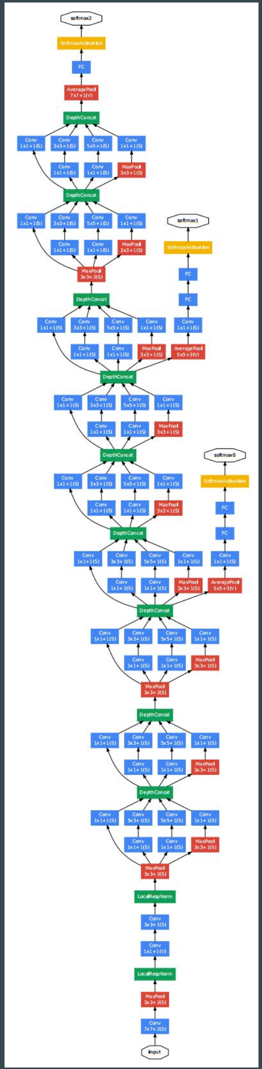
- PoseNet is a modified version of Googles 22 layer Inception Network (27 if counting pooling layers)
    - Includes 6 'inception modules' and 2 additional intermediate classifiers which are discarded during testing



(a) Inception module, naïve version



(b) Inception module with dimension reductions



Details of Approach (Neural network):

- Modifications to LeNet
  - Replace all softmax classifiers with affine regressors
  - Insert another fully connected layer with size 2048 before the final regressor (used for generalization exploration)
  - At test time, normalize quaternion orientation vector to unit length
- Results in a 23 layer (28 layers including pooling) network

Details of Approach (Neural network):

- Euclidean Loss / Affine Regressor layers

```
layer {
  name: "loss3/loss3_xyz"
  type: "EuclideanLoss"
  bottom: "cls3_fc_xyz"
  bottom: "label_xyz"
  top: "loss3/loss3_xyz"
  loss_weight: 1
}
```

```
layer {
  name: "loss3/loss3_wpqr"
  type: "EuclideanLoss"
  bottom: "cls3_fc_wpqr"
  bottom: "label_wpqr"
  top: "loss3/loss3_wpqr"
  loss_weight: 500
}
```

Details of Approach (Neural network):

- Learning location and orientation
  - Train network on Eucliden loss

$$loss(I) = \|\hat{x} - x\|_2 + \beta \left\| \hat{q} - \frac{q}{\|q\|} \right\|_2$$

  - Found that training on just position or orientation performed poorly compared to training on both simultaneously

Details of Approach (Neural network):

- Learning location and orientation

$$loss(I) = \|\hat{x} - x\|_2 + \beta \left\| \hat{q} - \frac{q}{\|q\|} \right\|_2$$

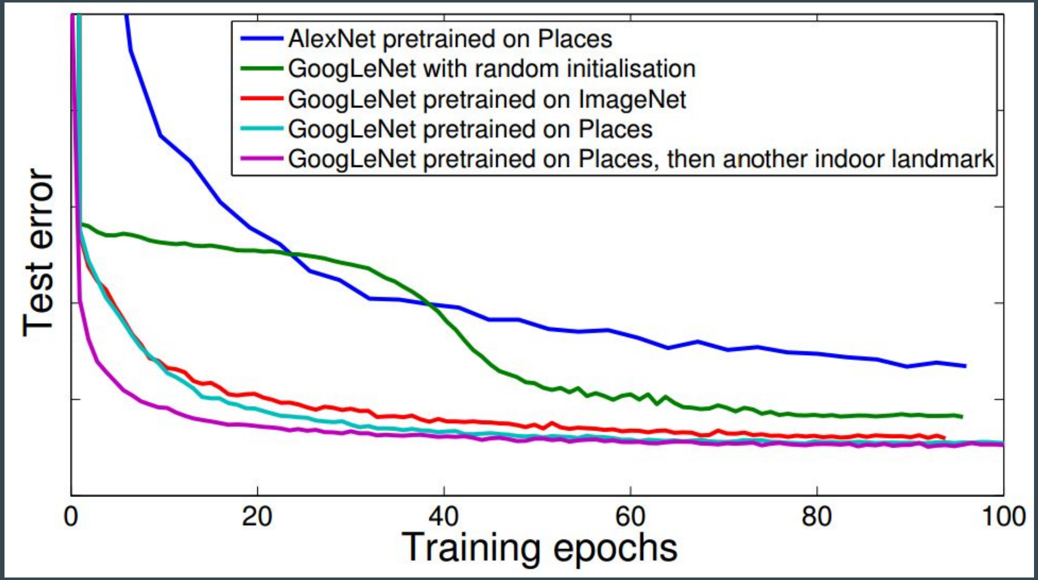
  - Balance β must be struck between orientation and translation penalties.
    - Optimal β given by ratio between expected error of position and orientation at the end of training (not beginning

Details of Approach (Neural network):

- PoseNet model was implemented in Caffe and trained using stochastic gradient descent
  - Base learning rate was 10^-5
    - Reduced by 90% every 80 epochs
  - Momentum of 0.9
  - Batch size of 75
  - Subtract separate image mean for each scene

Issues with Approach:

- Starting network weights (LeNet pretrained on XX) are very important for PoseNet performance



Issues with Approach:

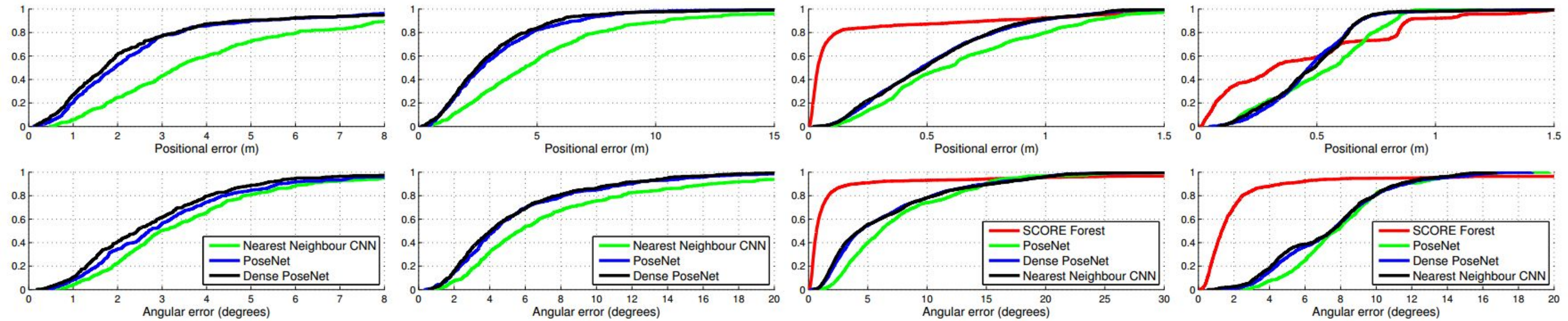
- No output uncertainty produced by network
- Relatively large error compared to SCoRe Forest (indoors - as SCoRe Forest cannot handle the large outdoor datasets)
- Even utilizing transfer learning yields semi-long training times (3-6 hours on Nvidia Titan X)

Results

Results:

Scene	# Frames Train	# Frames Test	Spatial Extent (m)	SCoRe Forest (Uses RGB-D)	Dist. to Conv. Nearest Neighbour	PoseNet	Dense PoseNet
King's College	1220	343	140 x 40m	N/A	3.34m, 2.96°	1.92m, 2.70°	1.66m, 2.43°
Street	3015	2923	500 x 100m	N/A	1.95m, 4.51°	3.67m, 3.25°	2.96m, 3.00°
Old Hospital	895	182	50 x 40m	N/A	5.38m, 4.51°	2.31m, 2.69°	2.62m, 2.45°
Shop Façade	231	103	35 x 25m	N/A	2.10m, 5.20°	1.46m, 4.04°	1.41m, 3.59°
St Mary's Church	1487	530	80 x 60m	N/A	4.48m, 5.65°	2.65m, 4.24°	2.45m, 3.98°
Chess	4000	2000	3 x 2 x 1m	0.03m, 0.66°	0.41m, 5.60°	0.32m, 4.06°	0.32m, 3.30°
Fire	2000	2000	2.5 x 1 x 1m	0.05m, 1.50°	0.54m, 7.77°	0.47m, 7.33°	0.47m, 7.02°
Heads	1000	1000	2 x 0.5 x 1m	0.06m, 5.50°	0.28m, 7.00°	0.29m, 6.00°	0.30m, 6.09°
Office	6000	4000	2.5 x 2 x 1.5m	0.04m, 0.78°	0.49m, 6.02°	0.48m, 3.84°	0.48m, 3.62°
Pumpkin	4000	2000	2.5 x 2 x 1m	0.04m, 0.68°	0.58m, 6.08°	0.47m, 4.21°	0.49m, 4.06°
Red Kitchen	7000	5000	4 x 3 x 1.5m	0.04m, 0.76°	0.58m, 5.65°	0.59m, 4.32°	0.58m, 4.17°
Stairs	2000	1000	2.5 x 2 x 1.5m	0.32m, 1.32°	0.56m, 7.71°	0.47m, 6.93°	0.48m, 6.54°

Results:



(a) King's College (b) St Mary's Church (c) Pumpkin (d) Stairs

Figure 7: **Localization performance.** These figures show our localization accuracy for both position and orientation as a cumulative histogram of errors for the entire testing set. The regression convnet outperforms the nearest neighbour feature matching which demonstrates we regress finer resolution results than given by training. Comparing to the RGB-D SCoRe Forest approach shows that our method is competitive, but outperformed by a more expensive depth approach. Our method does perform better on the hardest few frames, above the 95th percentile, with our worst error lower than the worst error from the SCoRe approach.

Conclusion

Conclusion / Summary:

- PoseNet is an end-to-end 6DOF pose regression convnet
- 5ms run-time, 50MB total storage space
- Large Scale indoor and outdoor relocalization
- Release of public dataset consisting of over 10,000 pose annotated images

Thanks!

Questions?

Quiz

Quiz:

1. PoseNet is able to output uncertainty
    a. True
    b. False
2. PoseNet is based off which of the following models?
    a. VGG16
    b. AlexNet
    c. LeNet
    d. ResNet