Learning to Localize Little Landmarks [1]

Singh, S.[1]; Hoiem, D.[1]; Forsyth, D.[1]

[1] University of Illinois, Urbana-Champaign

June 30th, 2017

DCC
DEPARTAMENTO DE CIÊNCIA DA COMPUTAÇÃO

UFMG

PATREO

Agenda

Agenda

Introduction

- The world is full of tiny but useful objects such as the door handle of a car or the light switch in a room
  - Little landmarks
  - In an image it is barely visible, yet we know where it is
  - They do not have a distinctive appearance of their own
    - Largely defined by their context

Little Landmarks in a Car

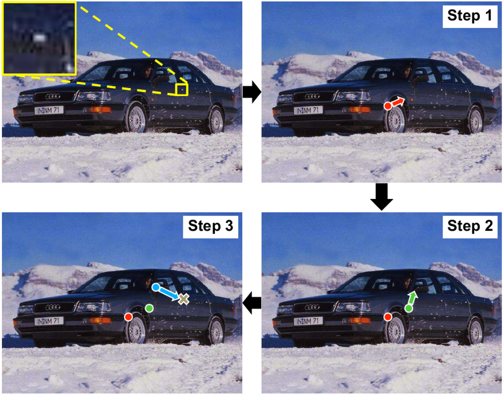


Figure: Several objects of interest are so tiny that they barely occupy few pixels (top-left), yet we interact with them daily. Localizing such objects in images is difficult as they do not have a distinctive local appearance.

Motivations

- Appearance may be similar to many other regions in the image
- May occur in a consistent spatial configuration
  - Location pattern according to other objects
  - Latent Landmark
  - May itself be hard to localize

Overview

- Approach for discovering globally distinctive patterns
- Supervised only by the location of the target
- The first latent landmark in the sequence must be localizable on its own
- Sequence of spatially dependent latent landmarks

Overview

- Handcrafted loss function
  - First latent landmarks must predict the next latent landmark
  - Last latent landmark must predict the target location
- Deep Convolutional Neural Network (CNN)

Contributions

- Novel and intuitive approach to localize little landmarks automatically
- Recurrent architecture using Fully Convolutional Networks
- Spatial information representation for prediction of locations
- Two new little landmark datasets
- Code and datasets are publicly available[1]

[1] http://vision.cs.illinois.edu/projects/litland/

Agenda

Related Works

- Well studied areas
  - Landmark localization
    - Human pose estimation [2, 3, 4]
    - Bird part localization [5, 6, 7]
  - Localization of larger objects [8, 9]
- Practically no work exists for localizing little landmarks

Related Works

- Karlinsky *et al.* [10] is conceptually most related to the paper
  - Keypoint proposals
  - Intermediate set of locations
  - Path from a known landmark to a target
- Current approach
  - Does not use keypoints
  - Learns to find the first landmark

Agenda

Baseline

- Detection
  - Simplest scheme for finding landmarks
  - Direct supervision for locations
  - Do not work for little landmarks

Baseline

- Prediction
  - Single latent landmark to predict the location of the target
  - Target could be far way
  - Hard task because there is no supervision for the latent landmark
  - Outperforms Detection

Approach

- Sequential Prediction
  - Sequential prediction scheme
  - Iteratively uses a latent landmark to predict the location of another latent landmark
  - Outperforms Prediction

Model and Inference

- Fully convolutional network architecture
  - Shared and step-specific layers
  - Step-specific parameters allow the features of a step to quickly adapt
  - Loss function penalizes disagreements between predicted and later detected locations

Architecture

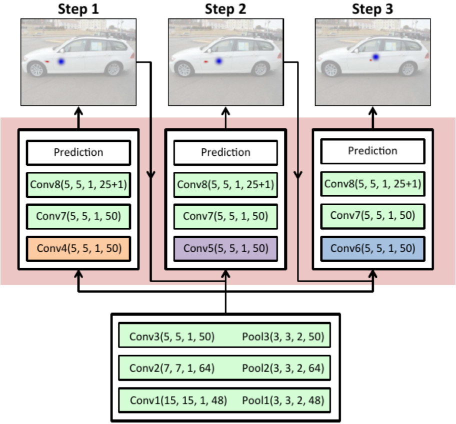


Figure: In each step a latent landmark (red blobs) predicts the location of the latent landmark for the next step. This is encoded as a feature map with radial basis kernel (blue blob) and passed as a feature to the next step.

Architecture

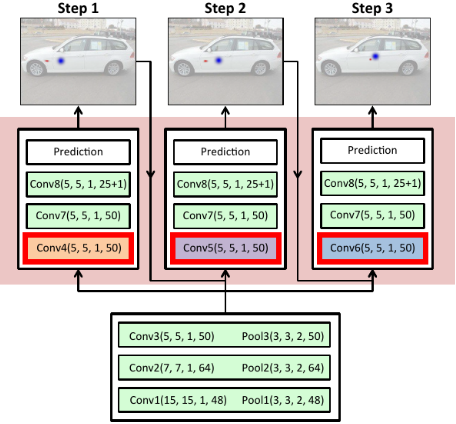


Figure: Orange, purple and blue show step specific layers.

Prediction Scheme

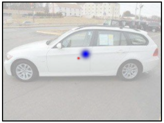
- Image as grid of locations l_i, $i \in \{1, ..., L\}$
- Each step s produces an estimation $P^{(s)}$ of the next latent landmark position
  - Each location l_i produce an estimate $p_i^{(s)}$ for $P^{(s)}$ with confidence $c_i^{(s)}$
  - $P^{(s)} = \sum_{i=1}^{L} c_i^{(s)} p_i^{(s)}$

Prediction Scheme

- $p_i^{(s)}$ is obtained by analysing both the image features and the predicted location in the previous step $P^{(s-1)}$
- $c_i^{(s)}$ is a softmax over all locations
  - $c_i^{(s)} = \frac{e^{z_i^{(s)}}}{\sum_i e^{z_i^{(s)}}}$
  - $z_i^{(s)} \in \mathbb{R}$ is the output from the network at l_i in step s

Prediction Scheme

- $P^{(s)}$ as a feature map
  - A Radial Basis kernel is placed centered in $P^{(s)}$



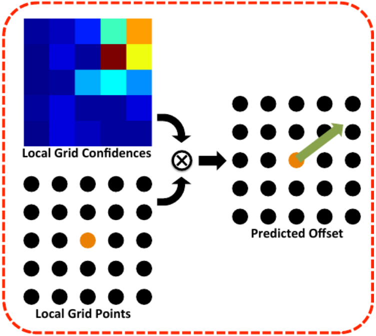
  - Add some “stochasticity” to the process
    - Allow the next step to easily ignore $P^{(s)}$, if needed

Prediction Scheme

- $P^{(s)}$ as a weighted average
  - Robust to individual variances
  - All locations are initialized with non-zero
    - All locations are potential latent landmarks

Location Estimation

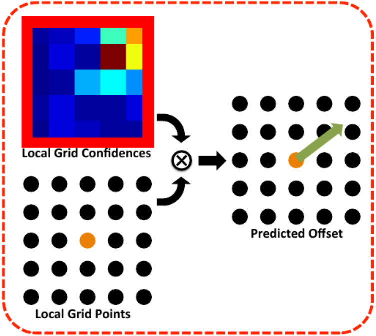
- How to generate $p_i^{(s)}$ at l_i at step s?
  - Simple regression works poorly
  - $g_j(*) \in \{-50, -25, 0, 25, 50\}$
  - Local grid of G points over l_i
    - $o_{j,i}^{(s)}$
    - $g_j^{(s)}$
    - $p_i^{(s)} = l_i + \sum_{j=1}^{G} o_{j,i}^{(s)} g_j$



Offset Prediction

Location Estimation

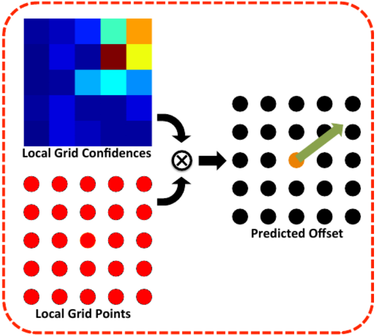
- How to generate $p_i^{(s)}$ at l_i at step s?
  - Simple regression works poorly
  - $g_j(*) \in \{-50, -25, 0, 25, 50\}$
  - Local grid of G points over l_i
    - $o_{j,i}^{(s)}$
    - $g_j^{(s)}$
    - $p_i^{(s)} = l_i + \sum_{j=1}^{G} o_{j,i}^{(s)} g_j$



Offset Prediction

Location Estimation

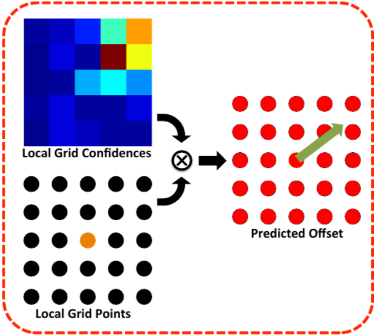
- How to generate $p_i^{(s)}$ at l_i at step s?
  - Simple regression works poorly
  - $g_j(*) \in \{-50, -25, 0, 25, 50\}$
  - Local grid of G points over l_i
    - $o_{j,i}^{(s)}$
    - $g_j^{(s)}$
    - $p_i^{(s)} = l_i + \sum_{j=1}^{G} o_{j,i}^{(s)} g_j$



Offset Prediction

Location Estimation

- How to generate $p_i^{(s)}$ at l_i at step s?
  - Simple regression works poorly
  - $g_j(*) \in \{-50, -25, 0, 25, 50\}$
  - Local grid of G points over l_i
    - $o_{j,i}^{(s)}$
    - $g_j^{(s)}$
    - $p_i^{(s)} = l_i + \sum_{j=1}^{G} o_{j,i}^{(s)} g_j$



Offset Prediction

Loss Function

- L2
  - Requires careful tuning of learning rate
- Huber Loss
  - $\mathcal{H}(x) = \begin{cases} \frac{x^2}{2\delta} & \text{, if } |x| < \delta \\ |x| - \frac{\delta}{2} & \text{, otherwise} \end{cases}$
  - Robustness
    - Gradients are exactly one for large loss values ($|x| > \delta$)
    - Gradients are less than one for smaller loss values ($|x| > \delta$)
    - $\delta = 1$

Loss Function

- $\mathcal{L}^{(s)} = \mathcal{H}\left(P^{(s)} - y*^{(s)}\right) + \gamma \sum_{i=1}^{L} c_i^{(s)} \mathcal{H}\left(p_i^{(s)} - y*^{(s)}\right)$
    - The first term enforces that the prediction $P^{(s)}$ coincides with the target $y*^{(s)}$
    - The scale factor γ (empirically set to 0.1)
    - The second term enforces that the individual predictions for each location also fall on the target, but the **individual losses** are **weighted by their contribution** ($c_i^{(s)}$)

Loss Function

- $\mathcal{L}^{(s)} = \mathcal{H}\left(P^{(s)} - y*^{(s)}\right) + \gamma \sum_{i=1}^{L} c_i^{(s)} \mathcal{H}\left(p_i^{(s)} - y*^{(s)}\right)$
    - The first term enforces that the prediction $P^{(s)}$ coincides with the target $y*^{(s)}$
    - The scale factor γ (empirically set to 0.1)
    - The second term enforces that the individual predictions for each location also fall on the target, but the **individual losses** are **weighted by their contribution** ($c_i^{(s)}$)

Loss Function

- $\mathcal{L}^{(s)} = \mathcal{H}\left(P^{(s)} - y*^{(s)}\right) + \gamma \sum_{i=1}^{L} c_i^{(s)} \mathcal{H}\left(p_i^{(s)} - y*^{(s)}\right)$
  - The first term enforces that the prediction $P^{(s)}$ coincides with the target $y*^{(s)}$
  - The scale factor γ (empirically set to 0.1)
  - The second term enforces that the individual predictions for each location also fall on the target, but the **individual losses** are **weighted by their contribution** ($c_i^{(s)}$)

Loss Function

- $\mathcal{L}^{(s)} = \mathcal{H}\left(P^{(s)} - y*^{(s)}\right) + \gamma \sum_{i=1}^{L} c_i^{(s)} \mathcal{H}\left(p_i^{(s)} - y*^{(s)}\right)$
  - The first term enforces that the prediction $P^{(s)}$ coincides with the target $y*^{(s)}$
  - The scale factor γ (empirically set to 0.1)
  - The second term enforces that the individual predictions for each location also fall on the target, but the **individual losses** are **weighted by their contribution** ($c_i^{(s)}$)

Agenda

Two New Datasets

- Light Switch Dataset (LSD)
- Car Door Handle Dataset (CDHD)
  - Based on the Stanford Car Dataset [2]

[2] http://ai.stanford.edu/~jkrause/cars/car_dataset.html

Repurposed Datasets

- Caltech UCSD Birds Dataset (CUBS) [3]
  - Beak location
- Leeds Sports Dataset (LSP) [4]
  - Wrist location

[3] http://www.vision.caltech.edu/visipedia/CUB-200.html
[4] http://www.comp.leeds.ac.uk/mat4saj/lsp.html

Evaluation Metrics

- LSD, CDHD, LSP
  - 2D Plot
    - y-axis = Detection Rate
    - x-axis = Normalized Distance from Ground Truth
- CUBS
  - PCP as used in [5]

Results and Discussion

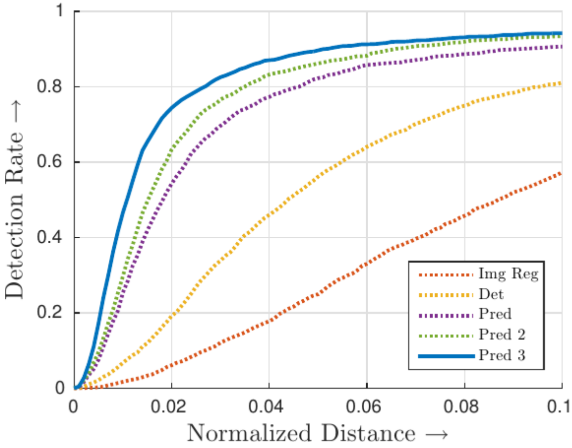
- CDHD
  - Img Reg
    ↓
    VGG-16

Table: Detection Rates for CDHD. Values for normalized distance of 0.02.

				Seq Prediction	
Method	**Img Reg**	**Det**	**Pred**	**Pred 2**	**Pred 3**
Detection Rate	6.1	19.2	54.3	63.3	**74.4**

Results and Discussion

- CDHD
  - Img Reg
    ↓
    VGG-16



Results and Discussion

- CDHD
  - Img Reg
    ↓
    VGG-16

Figure: Step 1 - Red. Step 2 - Green. Step 3 - Blue. The system finds the wheel as the first latent landmark and then moves towards the door handle.

Results and Discussion

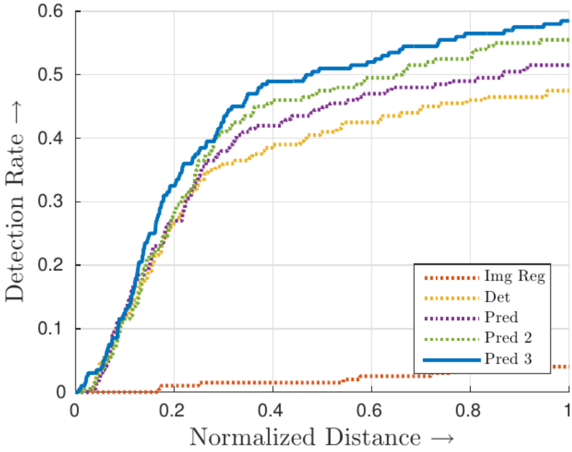
- LSD
  - Img Reg
    $\downarrow$
    VGG-16

Table: Detection Rates for LSD. Values for normalized distance of 0.5.

				Seq Prediction	
Method	**Img Reg**	**Det**	**Pred**	**Pred 2**	**Pred 3**
Detection Rate	1.5	41.0	44.5	47.5	**51.0**

Results and Discussion

- LSD
  - Img Reg
    ↓
    VGG-16



Results and Discussion

- LSD
  - Img Reg

    $\downarrow$

    VGG-16

Figure: Step 1 - Red. Step 2 - Green. Step 3 - Blue. The system relies on finding the edge of the door first.

Results and Discussion

- UCSD
  - Img Reg

    ↓

    VGG-16

Table: PSP for UCSD.

Methods	PCP
Liu *et al.* [5]	49.0
Liu *et al.* [6]	61.2
Shih *et al.* [7]	51.8
Proposed	**64.1**

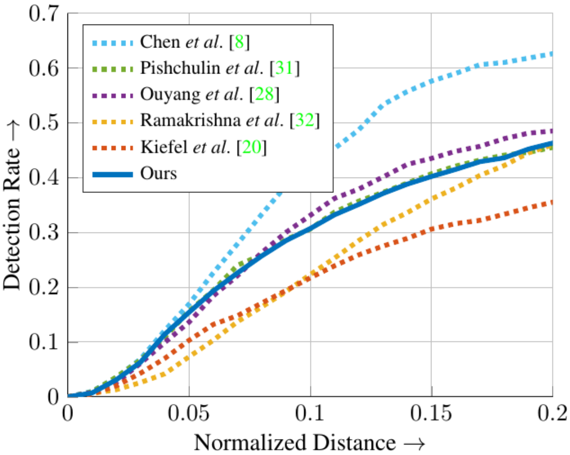
Results and Discussion

- UCSD
  - Img Reg
    ↓
    VGG-16

Figure: Step 1 - Red. Step 2 - Green. Step 3 - Blue. The first landmark tends to be on the neck, followed by one near the eye and the last tends to be outside at the curve of neck and beak

Results and Discussion

- LSP
  - Img Reg
    ↓
    VGG-16



Agenda

Conclusions

- Recognizable atterns emerged solely from the supervision of the target landmark
  - Adapt to the evidence in the image
  - The method does not impose any hard constraints
  - Later steps can choose to ignore the evidence from earlier steps

Conclusions

- Strong performance in the tasks
  - Success attributed from the spatial prediction scheme
- Future work
  - Multiple targets
  - Directed Graphs of latent landmarks
  - Accumulation from features of previous steps

[1] Saurabh Singh, Derek Hoiem, and David Forsyth.
Learning to localize little landmarks.
In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 260–269, 2016.

[2] Mykhaylo Andriluka, Stefan Roth, and Bernt Schiele.
Pictorial structures revisited: People detection and articulated pose estimation.
In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, pages 1014–1021. IEEE, 2009.

[3] Matthias Dantone, Juergen Gall, Christian Leistner, and Luc Van Gool.
Human pose estimation using body parts dependent joint regressors.
In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3041–3048, 2013.

[4] Marcin Eichner, Vittorio Ferrari, and S Zurich.
Better appearance models for pictorial structures.
In BMVC, volume 2, page 5, 2009.

[5] Jiongxin Liu and Peter N Belhumeur.
Bird part localization using exemplar-based models with enforced pose and subcategory consistency.
In Proceedings of the IEEE International Conference on Computer Vision, pages 2520–2527, 2013.

[6] Jiongxin Liu, Yinxiao Li, and Peter N Belhumeur.
Part-pair representation for part localization.
In European Conference on Computer Vision, pages 456–471. Springer, 2014.

[7] Kevin J Shih, Arun Mallya, Saurabh Singh, and Derek Hoiem.
Part localization using multi-proposal consensus for fine-grained categorization.
arXiv preprint arXiv:1507.06332, 2015.

[8] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan.
Object detection with discriminatively trained part-based models.
IEEE transactions on pattern analysis and machine intelligence, 32(9):1627–1645, 2010.

[9] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik.
Rich feature hierarchies for accurate object detection and semantic segmentation.
In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 580–587, 2014.

[10] Leonid Karlinsky, Michael Dinerstein, Daniel Harari, and Shimon Ullman.

The chains model for detecting parts by their context.
In Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, pages 25–32. IEEE, 2010.