Learning to Localize Objects with Structured Output Regression

Mathew B. Blaschko and Christoph H. Lampert
(Best Student Paper Award – ECCV'08)

Presented by
Yimeng Zhang and Adarsh Kowdle

Introduction

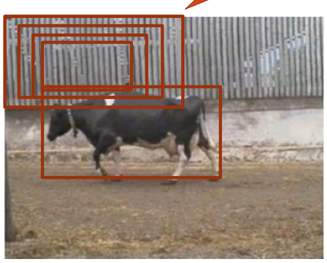
- What is object localization or object detection?

2

Introduction

- Sliding window approach

Trained Cow detector



3

Introduction

- Sliding window approach - disadvantages
  - Computationally inefficient
    - Addressed by earlier work on efficient sub-window search (CVPR '08) – Branch and bound optimization
  - Not clear how to optimally train a discriminant function for localization – **this paper**
    - Propose a training strategy that specifically optimizes localization accuracy
    - Structured learning
      - Output space is the space of all bounding boxes – parameterized by 4 numbers i.e. corners of the box

4

Algorithm Overview

Apply structured SVM algorithm to object localization

$$g : X \to Y$$

Input X: the space of all images
Output Y: the space of all bounding boxes (rectangles)

Input x

Output y: [top, left, bottom, right]

5

Structured SVM

$$\min_{w} \; \|w\|^2 + C\sum_{i=1}^{n}\xi_i$$

$$\text{s.t.} \;\; \langle w, \underbrace{\boxed{\varphi(x_i, y_i)}}_{\text{Feature vector}}\rangle - \langle w, \varphi(x_i, y)\rangle \geq \underbrace{\boxed{\Delta(y_i, y)}}_{\text{Loss function}} - \xi_i \quad \forall y \in \mathcal{Y} \setminus \{y_i\}$$

 −  $\geq \Delta(y_i, y) - \xi_i$

 −  $\geq \Delta(y_i, y) - \xi_i$

...

6

Loss Function

$$\Delta(y_i, y) = \begin{cases} 1 - \frac{\text{Area}(y_i \bigcap y)}{\text{Area}(y_i \bigcup y)} & \text{if } y_{i\omega} = y_\omega = 1 \\ 1 - \left(\frac{1}{2}(y_{i\omega} y_\omega + 1)\right) & \text{otherwise} \end{cases}$$

7

Feature Vector

- Feature vector extracted from the image restricted to the box region X | y

$\psi($  $) = \phi($  $)$

$$\psi(x, y) = \phi(x|_y)$$

8

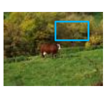
Joint Kernel

- Structured SVM can also be written in terms of kernels

$$< w, \psi(x_i, y_i) >= \underbrace{\sum_x \sum_y \alpha_{xy}}_{\text{Support vectors}} \underbrace{< \psi(x, y), \psi(x_i, y_i) >}_{\text{Joint Kernel}}$$

$$k_{joint}((x, y), (x', y')\ ) = k_{image}(x|_y, x'|_{y'},)$$

Linear Case

$K_{joint}($  ,  $) =< \phi($  $)\ \phi($  $)>$

Non-linear Kernels: Polynomial Kernels, Gaussian Kernels

9

Joint Kernel Examples

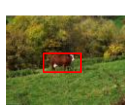
is large.

is small.

could also be large.

10

Maximization steps



- Most violated constraints

$$\max_{y \in \mathcal{Y} \setminus y_i} \langle w, \varphi(x_i, y) \rangle + \Delta(y_i, y)$$

- Testing

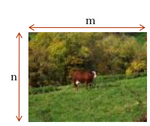
$$g(x) = \operatorname{argmax}_{y \in \mathcal{Y}} \langle w, \varphi(x, y) \rangle$$

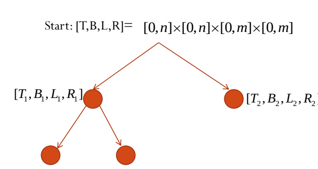
- Efficient Algorithm: Branch and Bound

11

Branch and Bound

- Branch: divide the output space into subspaces
- Bound: pruning the subspaces whose upper bound is lower than some guaranteed score in other subspaces





12

Comparison with Sliding Window Approach

- Same:
  - feature vectors,
  - model parameters
  - Inference steps

$w^T\psi($  $) = w^T\phi($  $)$

- Different:
  - loss function
  - Training steps

  (sliding widow: sample negative boxes,
  this paper: cutting plane)

13

Experiments

- TU Darmstadt cow dataset

- PASCAL VOC 2006 dataset

14

Experiments

- Bag-of-visual-words approach
  - Extract local SURF descriptors*
    - 10000 descriptors – K means clustered into 3000 entry codebook
  - Every bounding box is now described by a histogram of these features
- Binary training – benchmark binary classifier
  - Ground truth boxes are positive samples
  - Randomly sampled boxes (<20% overlap with ground truth) are negative samples

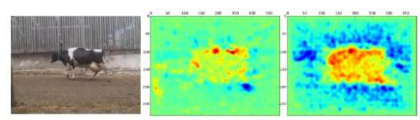
* Herbert Bay, Andreas Ess, Tinne Tuytelaars, Luc Van Gool, "SURF: Speeded Up Robust Features", Computer Vision and Image Understanding (CVIU), 2008

15

Results

- TU Darmstadt cow dataset – all images contain a cow

Localization performance



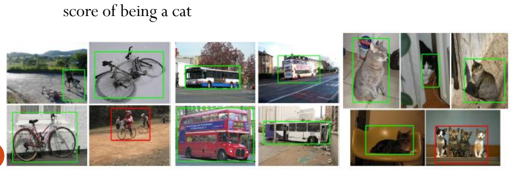
Binary training Structured learning

- Well distributed scores over the cow and negative weights for the background

16

Experiments

- PASCAL VOC 2006 dataset
  - Strongly unbalanced
    - Images may not contain object being detected
    - Separate SVM to rank the bounding boxes.
- The framework does not allow for detecting multiple objects
  - Group of cats image – the bigger bounding box will have a high score of being a cat



17

Summary

- Structured learning makes better use of training data
  - More sensible negative examples are added to the training data in structured learning
  - Focusing training on locations where mistakes would otherwise be made
- The loss function in the structured learning framework allows to suitably incorporate partial detections into the training which are not possible with binary training.

18

Thank you

Questions?

19