Bayesian Joint Topic Modelling for Weakly Supervised Object Localisation

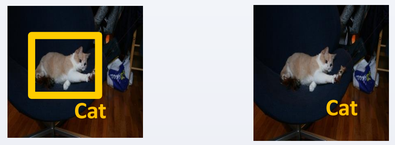
Zhiyuan Shi, Timothy M. Hospedales, Tao Xiang

Queen Mary, University of London, London E1 4NS, UK

{zhiyuan.shi,tmh,txiang}@eecs.qmul.ac.uk

International Conference on Computer Vision Sydney 2013

Task



Fully annotated **Weakly annotated**

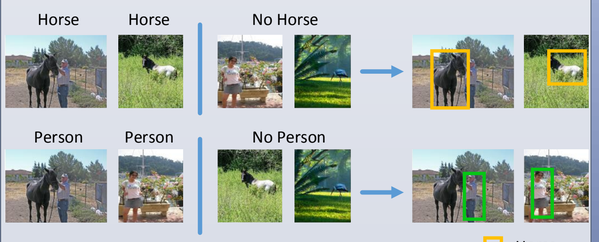
- Many computer vision tasks require fully annotated data, but
  - ✓ Time-consuming, Laborious, Human various
- More and more online media sharing websites (e.g. Flickr) provide weakly annotated data, However,
  - ✓ Weaker supervision, Ambiguity (background clutter, occlusion…)
- Challenge: Weakly Supervised Object Localisation (WSOL).

Existing Approaches vs. Ours

Three types of cues are exploited in existing WSOL:

- ✓ **Object-saliency**: A region containing the object should look different from background in general.
- ✓ **Intra-class**: The region should look similar to other regions containing the object of interest in other training images.
- ✓ **Inter-class**: The region should look dissimilar to any regions that are known to not contain the object of interest.

However, they are independently trained:



Independent learning ignores the fact that:

- ✓ The knowledge that multiple objects co-exist within each image is not exploited.
- ✓ The background is relevant to different foreground object classes.

Our contributions:

- ✓ We propose the novel concept of joint modelling of all object classes and backgrounds for weakly supervised object localisation.
- ✓ We formulate a novel Bayesian topic model suitable for localization of objects and utilizing various types of prior knowledge available.
- ✓ We provide a solution for exploiting unlabeled data for semi+weakly supervised learning of object localisation.

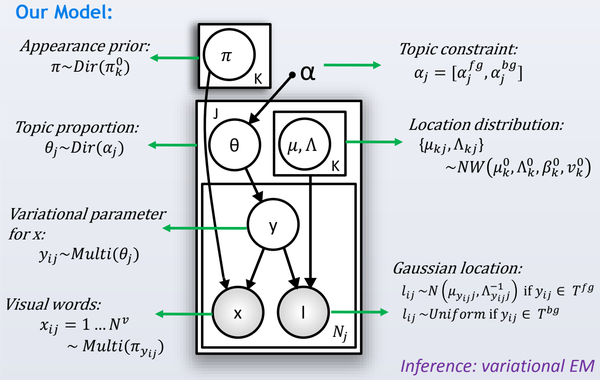
Methodology

Preprocessing and Representation:

- ✓ Regular Grid SIFT Descriptors. Sampled every 5 pixels.
- ✓ Quantising using $N_v = 2000$ word codebook.
- ✓ Words and corresponding locations:

$$\{x_i, l_{xi}, l_{yi}\}_{i=1}^{N_j}$$

Our Model:



Inference: variational EM

Observed variables:

- ✓ $O = \{x_j, l_j\}_{j=1}^{J}$ Low-level feature words and corresponding location

Latent variables:

- ✓ $H = \{\{\pi_k\}_{k=1}^{K}, \{y_j, \mu_{kj}, \Lambda_{kj}, \theta_j\}_{k=1,j=1}^{K,J}\}$ For each topic k and image j

Given parameters:

- ✓ $\Pi = \left\{\{\pi_k^0, \mu_k^0, \Lambda_k^0, \beta_k^0, \nu_k^0\}_{k=1}^{K}, \{\alpha_j\}_{j=1}^{J}\right\}$ Label information and prior

Joint distribution:

$$p(O,H|\Pi) = \prod_j^J \prod_k^K \left[p(\mu_{jk}, \Lambda_{jk}|\mu_k^0, \Lambda_k^0, \beta_k^0, \nu_k^0) p(\theta_j|\alpha_j) \left(\prod_i^{N_j} p(x_{ij}|y_{ij}, \theta_j) p(y_{ij}|\theta_j) \right) \right] p(\pi_k|\pi_k^0)$$

Prior Knowledge:

- ➢ *Human knowledge* objects and their relationships with backgrounds
  - ✓ Objects are compact whilst background spread across the image.
  - ✓ Objects stand out against background.

$$\pi_k^0 = |\frac{1}{N_c} \sum_{j,c_j=k} h(x_j) - \frac{1}{N_j} \sum_j h(x_j)|_+$$

- ➢ Transferred knowledge
  - ✓ Appearance and Geometry information from existing dataset.

Object Localisation:

- ✓ *Our-Gaussian* Aligning a window to the ellipse obtained from $q(\mu, \Lambda)$
- ✓ *Our-Sampling* Non-maximum suppression sampling over heat-map

Results

Dataset: PASCAL VOC 2007. Three variants are used:

- ✓ VOC07-6×2 : 6 classes with Left and Right poses, 12 classes in total.
- ✓ VOC07-14: 14 classes, other 6 were used as annotated auxiliary data
- ✓ VOC07-20: all 20 classes, each class contain all pose data.

PASCAL criterion:

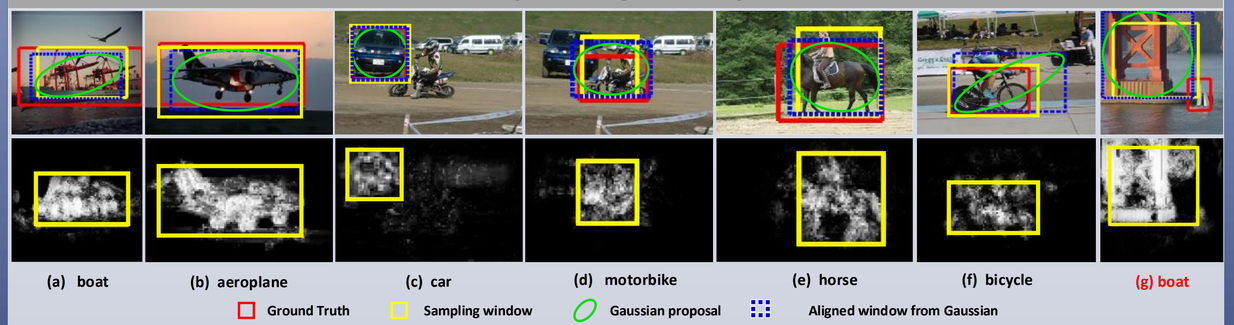
- ✓ intersection-over-union > 0.5 between Ground-Truth and predicted box

Comparison with state-of-the-art

- ✓ Initialisation: Localising object of interest in weakly labelled images.
- ✓ Refined by detector: A conventional object detector can be trained using initial annotation. Then it can be used to refine object location.

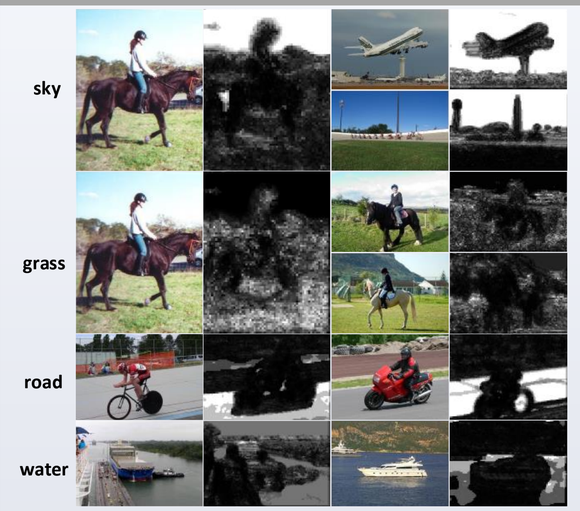
Method	Initialisation			Refined by detector		
	6×2	14	20	6×2	14	20
Deselares et al [1]	39	22	-	50	28	-
Pandey and Lazebnik [2]	43.7	23.0	-	61.1	30.3	-
Siva and Xiang [3]	40	-	28.9	49	-	30.4
Siva et al. [4]	37.1	-	29.0	46	-	-
Our-Sampling	50.8	**32.2**	**34.1**	65.5	**33.8**	**36.2**
Our-Gaussian	**51.5**	30.5	31.2	**66.1**	32.5	33.4

Example: Foreground Topics



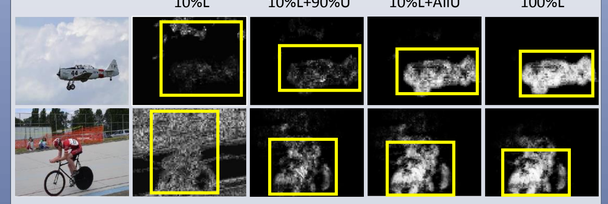
- ✓ Figs. (c) and (d) illustrate that the object of interest "explain away" other objects of no interest.
- ✓ A car is successfully located in Fig. (c) using the heat map of car topic.
- ✓ Fig. (d) shows that the motorbike heat map is quite accurately selective, with minimal response obtained on the other vehicular clutter.
- ✓ Fig. (e) indicates how the Gaussian can sometimes give a better location.
- ✓ Fig. (f) shows that the single Gaussian assumption is not ideal when the foreground topic has a less compact response.
- ✓ A failure case is shown in Fig. (g), where a bridge structure resembles the boat in Fig (a) resulting strong response from the foreground topic, whilst the actual boat topic is small and overwhelmed.

Example: Background Topics



- ➢ Background non-annotated data has been modelled in our framework.
- ➢ Irrelevant pixels will be explained to reduce confusion with object.
- ➢ Automatically learned background topics have clear semantic meanings, corresponding to common components as shown in the Figure.
- ➢ Some background components are mixed, e.g. the water topic gives strong response to both water and sky. But this is understandable since water and sky are almost visually indistinguishable in the image.

Example: Semi-supervised Learning



- ➢ Unknown image can set as α_j^{fg}=0.1. (soft constraint)
- ➢ 10% labelled data + 90% unlabeled data (relevant) or unrelated data
- ➢ Evaluating on (1) initially annotated 10% data (standard WSOL).
  (2) testing part dataset (localize objects in new images
- ➢ The figure clearly shows unlabeled data helps to learn a better object model.

References

[1] T. Deselaers, B. Alexe, and V. Ferrari. Weakly supervised localization and learning with generic knowledge. IJCV. 2012.

[2] M. Pandey and S. Lazebnik. Scene recognition and weakly supervised object localization with deformable part-based models. In ICCV, 2011

[3] P. Siva and T. Xiang. Weakly supervised object detector learning with model drift detection. In ICCV, 2011.

[4] P. Siva, C. Russell, and T. Xiang. In defence of negative mining for annotating weakly labelled data. In ECCV, 2012.