Learning the structure of objects from Web supervision

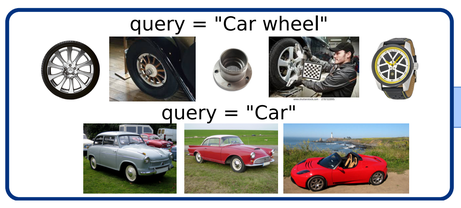
David Novotný[1,2], *Diane Larlus*[2], *Andrea Vedaldi*[1]

[1]*Visual Geometry Group, University of Oxford*
[2]*Computer Vision Group, Xerox Research Centre Europe*

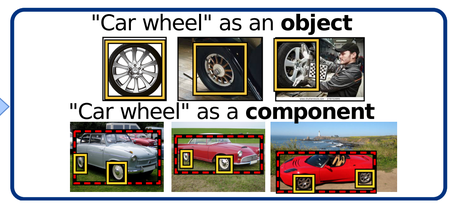
Objective

Learn about objects, semantic parts and their geometric relationships from Web images

input
Noisy Web results



our method



output
Learned concept

Main goal

Learn the **structure** of an object category
→ By decomposition to **semantic (nameable) parts**

Full supervision too expensive → collect annotations from the **Web**

Related work:

LEVAN [1], NEIL [2] - Webly supervised learning of visual concepts
→ Focus on large scale learning
→ Less emphasis on parsing objects into parts and discovering geometric relationships

Discovering anchors in deep networks

Anchors are linear appearance models operating on R-CNN [3] features

A novel mid-level element discovery formulation:

$$\min_{\omega_1,\ldots,\omega_K} \sum_{k=1}^{K} \Big[\underbrace{\frac{\lambda}{2}\|\omega_k\|^2}_{\text{Regularizer}} + \underbrace{\sum_{i=1}^{n} \ell(y_i, x_i, \omega_k)}_{\text{Discriminability term}} \Big] + \underbrace{\gamma \sum_{k\neq q} \left\langle \frac{\omega_k}{\|\omega_k\|}, \frac{\omega_q}{\|\omega_q\|} \right\rangle^2}_{\text{Diversity term}}$$

→ Results in **diverse** and **discriminative** linear anchor models $\{\omega_1, \ldots \omega_K\}$

Method

Multiple Instance Learning

Detection of parts in Web images formalized as a **Multiple Instance Learning** problem:

For each semantic part class, we learn a linear scoring function $\langle \underbrace{\mathbf{w}}_{\text{linear model}}, \underbrace{\phi(R)}_{\text{embedding of region } R} \rangle$

$\langle \mathbf{w}, \phi(R) \rangle$ associates a high score to image regions R that contain the semantic part

Main contribution: **Robust geometric embedding $\phi^{\mathbf{g}}(\mathbf{R})$**

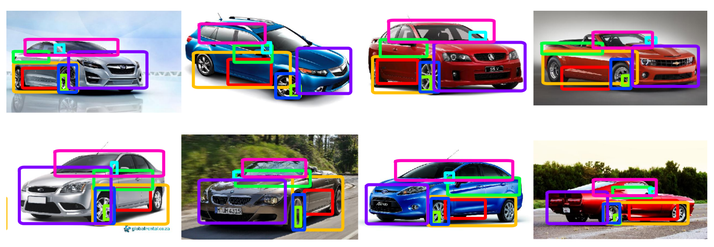
Anchors - robust geometric reference

Semantic parts are hard to detect

Leverage non-semantic **anchors** to construct $\phi^{\mathbf{g}}(\mathbf{R})$

Anchors:

- Visually stable mid-level elements
- Robust to noise and geometric transformations
- Interpretability not required

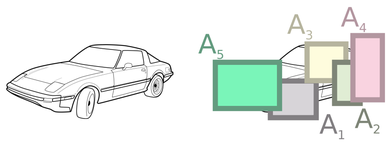


Geometric embedding

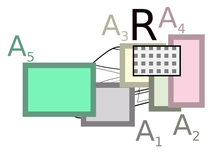
Anchors serve as a robust geometric reference

→ *An arbitrary region R can be* **localized relative to the anchors**
→ *encoding of the relative locations defines* $\phi^{\mathbf{g}}(\mathbf{R})$

1. Anchors ($A_1, \ldots, A_5$) define a robust geometric reference



2. Relative position of R and $(A_1, \ldots, A_5)$ encodes position w.r.t. the object



3. A geometric embedding $\phi^{\mathbf{g}}(\mathbf{R})$ is a stacking of similarity measures between anchors and R

$$\phi^g(\boxdot) = \begin{bmatrix} 0 \\ 0.3 \\ 0.5 \\ 0.1 \\ 0 \end{bmatrix}$$

As a region similarity measure we select the Intersection-over-Union, i.e.: $\phi^{\mathbf{g}}(\mathbf{R}) = \begin{bmatrix} IoU(R, A_1) \\ \vdots \\ IoU(R, A_K) \end{bmatrix}$

$IoU(\cdot,\cdot)$ ... Intersection-over-Union

Theoretical properties:
IoU is a positive definite kernel (proof in the paper)
→ The anchors correspond to a basis of a Hilbert Space with the *IoU* measure as its valid dot product
→ *IoU* and the anchor-induced coordinate frame provide a natural measure of geometric similarity between regions

Geometry aware MIL

Appearance-geometry embedding $\phi^{\mathbf{ag}}(\mathbf{R})$ of a region R is defined as:

$$\phi^{ag}(R) = \phi^{\mathbf{g}}(\mathbf{R}) \otimes \phi^{\mathbf{a}}(\mathbf{R})$$

$\otimes$ ... Kronecker product
$\phi^{\mathbf{a}}(\mathbf{R})$ ... appearance embedding (CNN feature)

Using $\phi^{ag}(R)$ in the MIL objective function results in a robust geometry-aware model

Experiments

Part detection experiments conducted on two datasets:
PascalParts - "car" and "bus" parts **Labeled Faces in The Wild (LFPW)** - "face" parts

Webly supervised detection of object parts

Evaluation of part detector performance on PascalParts and LFPW

Supervision	Method	mAP {face}	mAP {car}	mAP {bus}
Web	Cho et al. [4]	16.6	16.9	12.4
	Bilen & Vedaldi [5]	2.7	12.0	4.7
	MIL	20.6	29.1	22.7
	MIL + geom. embedding (ours)	**44.9**	**34.4**	**23.0**
Web + Single annotation	MIL	27.3	33.3	26.9
	MIL + geom. embedding (ours)	**43.0**	**36.4**	**30.1**
Full supervision	R-CNN	53.7	51.2	48.2
	R-CNN + geom. embedding (ours)	**61.4**	**60.3**	**54.1**

Geometric embedding improves part detection performance in all cases

Mid-level element discovery

Evaluation of discriminative power of anchors

- Comparison to mid-level element discovery baselines on the MIT Places dataset

features	accuracy [%] SVM baseline	BoE[6]	ours
Decaf	57.5	69.7	**71.5**
VGG-VD	68.9	77.6	**77.8**

Anchors outperform other mid-level element mining methods in terms of discriminative power.

Semantic matching

Evaluation of the ability to establish semantic matches

- Given a ground truth part annotation R_S in the source image
  → find the best matching region R_T in the target image

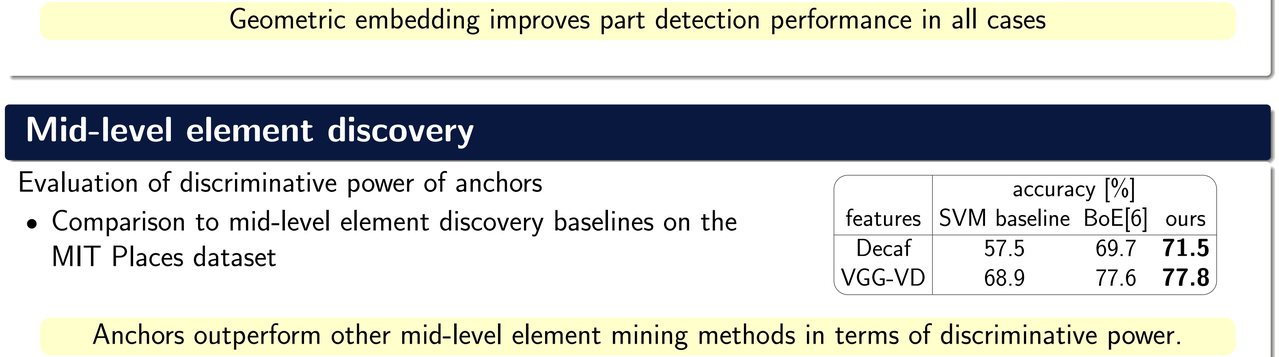
- Matches established by maximizing similarity between appearance-geometry embeddings $\phi^{ag}(R)$

Parent class	Matching performance [average IoU over parts] App.-Geom. embedding (ours)	Appearance embedding	Flowweb [7]	DSP [8]
{car}	**0.36**	0.31	0.34	0.23
{bus}	**0.37**	0.31	0.31	0.22
{face}	0.41	0.33	**0.43**	0.19

Appearance-Geometry embedding outperforms competition on classes with large viewpoint variations.

Visual semantic atlas

- A byproduct is a dataset of web images annotated with object and part locations
- Our method allows intuitive browsing of the image collection



References

[1] S. K. Divvala, A. Farhadi, and C. Guestrin, "Learning everything about anything: Webly-supervised visual concept learning," in *Proc. CVPR*, 2014.

[2] X. Chen, A. Shrivastava, and A. Gupta, "Neil: Extracting visual knowledge from web data," in *Proc. ICCV*, 2013.

[3] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. CVPR*, 2014.

[4] M. Cho, S. Kwak, C. Schmid, and J. Ponce, "Unsupervised Object Discovery and Localization in the Wild: Part-based Matching with Bottom-up Region Proposals," in *Proc. CVPR*, 2015.

[5] H. Bilen and A. Vedaldi, "Weakly supervised deep detection networks," *arXiv preprint arXiv:1511.02853*, 2015.

[6] Y. Li, L. Liu, C. Shen, and A. van den Hengel, "Mid-level deep pattern mining," in *Proc. CVPR*, 2015.

[7] T. Zhou, Y. Jae Lee, S. X. Yu, and A. A. Efros, "Flowweb: Joint image set alignment by weaving consistent, pixel-wise correspondences," in *Proc. CVPR*, 2015.

[8] J. Kim, C. Liu, F. Sha, and K. Grauman, "Deformable spatial pyramid matching for fast dense correspondences," in *Proc. CVPR*, 2013.