Concept Detection with Blind Localization Proposals

Hichem SAHBI

CNRS TELECOM ParisTech, Paris

ImageCLEF/CLEF 2015

Sep 9th 2015

Outline

Visual Category Recognition

Goal : recognize categories (sky, sea, cars, trees, roads, persons,...)

Concept Detection

car, building, etc.

Concept Localization

Object Class Segmentation

Related work : Pascal VOC 05-12, ImageCLEF 04-., ImageNET 09-., MS COCO 14-., etc.

Motivation and Contribution (I)

Related Work (concept detection and localization)

- Sliding window generation and scoring (Viola & Jones, 2001).
- Pixel and superpixel-based partitioning (Shotton et al, 2006, Batra et al., 2008, Yang et al., 2007, Ladicky et al., 2009, etc.)
- Scoring : SVMs, decision forests, deep networks + graphical models for pixel/superpixel/segment interactions (Farag et al, 2006, Kohli et al. 2008, etc.)

Limitations

- Sliding window scoring is very expensive.
- Pixel-based partitioning not expressive.
- Segments (or superpixels) do not always correspond to relevant objects.

Motivation and Contribution (II)

Related Work (cont.)

- Multiple segmentations (Pantofaru et al., 2008, etc.).
- More recently : segmentation proposals (Malik et al., 2012, etc.)

Limitations (cont.)

- Multiple (and proposal) segmentations are computationally expensive and may be incomplete.
- Segmentation is an ill-posed problem.

Proposed Solution (2 steps)

- Achieve first image annotation : train classifiers to detect concepts.
- Many concept locations (sky, sun) are highly predictable : use multiple object localization proposals based on a priori statistical trained model instead of image segmentation.

Outline

Outline

Concept Detection with SVMs

- We trained "one-versus-all" SVM classifiers for each concept c ; we use many random folds (taken from training data) for multiple SVM training and we use these SVMs in order to predict the concepts on image (depending on the sign)

$$f_c(x) = \sum_{\ell=1}^{N} 1_{\{g_{\ell(x)} \geq 0\}} - \sum_{\ell=1}^{N} 1_{\{g_{\ell(x)} < 0\}}, \qquad (N = 10 \text{ in practice})$$

$$g_\ell(x) = \sum_{x'} \alpha_{\ell,x'} \mathbf{K}_{x,x'} + b_\ell$$

- We used only the 9 (provided) visual features.
- We build 10 gram matrices (9 visual + 1 textual), based on efficient histogram intersection kernel. We linearly combine those matrices into a single one.

Kernel Map Evaluation (I)

- About kernels : $\mathbf{K}_{x,x'}$, when (p.s.d) $\mathbf{K}_{x,x'} = \Phi'_x \Phi_{x'}$
- Linear kernel map : $\mathbf{K}_{x,x'} = \langle x, x' \rangle$ (just identity map).
- Polynomial kernel map : $\mathbf{K}_{x,x'} = \langle x, x' \rangle^p = \Phi'_x \Phi_{x'}$ with $\Phi_x = x \otimes ... \otimes x$ (p times).
- Histogram intersection map : $\mathbf{K}_{x,x'} = \sum_{d=1}^{s} \min(x^d, x'^d)$. Each dimension x^d of x is mapped using

$$\psi(x^d) = 2^0 + 2^1 + \cdots + 2^{k(x^d)}$$

$$k(x^d) = \left\lfloor Q \frac{x^d - \ell_d}{u_d - \ell_d} \right\rfloor$$

$\psi(.)$ is a "decimal-to-unary" map ; $\psi(x^d)$ is a Q dimensional vector with its $k(x^d)$ first dimensions equal to 1 and the remaining $Q - k(x^d)$ to 0, e.g., with $Q = 4$, 1 is mapped to 0001, 2 is mapped to 0011, 3 is mapped to 0111, and so on.

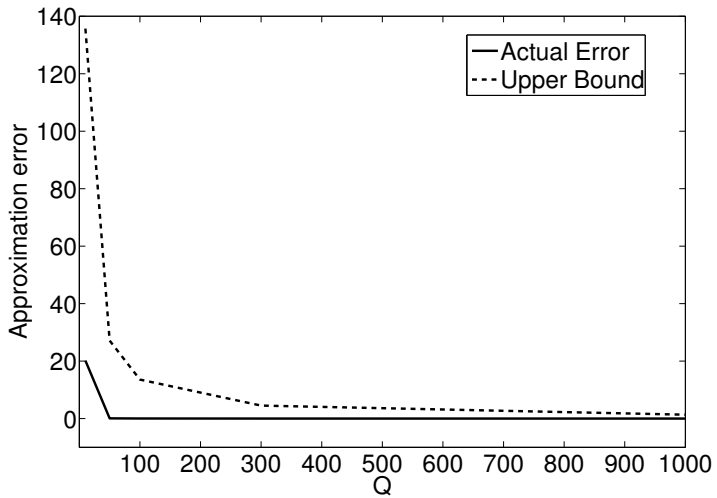
Kernel Map Evaluation (II)

Proposition (3)

Given x, x' in $\mathcal{X}$, for sufficiently large Q, the inner product $\langle \Phi_x, \Phi_{x'} \rangle$, with $\Phi_x = \left(\psi(x^1)' \sqrt{\frac{u_1 - \ell_1}{Q}}, \sqrt{u_1}, \ldots, \psi(x^s)' \sqrt{\frac{\ell_s - u_s}{Q}}, \sqrt{u_s} \right)'$, *approximates the histogram intersection kernel* $\sum_{d=1}^{s} \min(x^d, x'^d)$.

Proof shows that $\left| \langle \Phi_x, \Phi_{x'} \rangle - HI(x, x') \right| \leq \frac{1}{Q} \sum_{d=1}^{s} u_d - \ell_d \rightsquigarrow 0$ as $Q \nearrow$

(Sahbi, ICPR 2014)



Outline

Blind Localization Proposals (I)

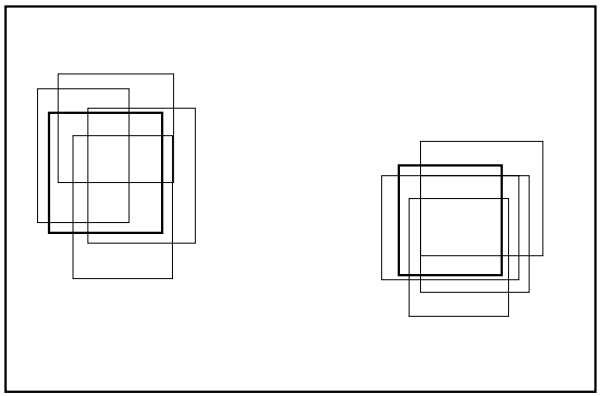
A two step process

- Given concept detection results :
  - Several heuristics are tried in order to suggest multiple concept localization proposals
  - Concept localization is achieved blindly, i.e., without consulting the content of the test image (but only its detected concepts)
  - Bounding boxes (BBs) are either fixed (using test image dimensions) or based on statistics estimated offline on the training/dev set (in "imageclef2015.dev.bbox.v20150226")

Blind Localization Proposals (II)

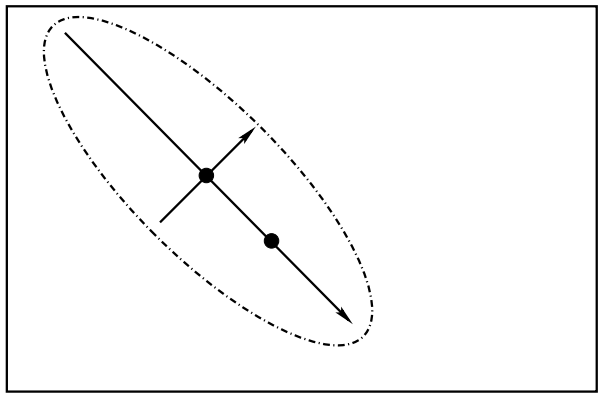
- **Heuristic 1 (fixed BB) :** for a detected concept c in a given test image, its bounding box p_c is set to $(W/2, H/2, W, H)$.

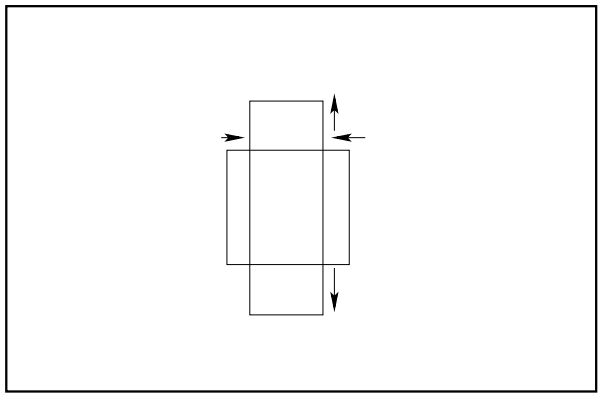
- **Heuristic 2 (concept dependent BBs) :** for a detected concept c in a given test image, we generate N_c bounding boxes whose coordinates correspond to the cluster centers obtained after applying k-means on $\mathcal{T}_c$ (set of BBs).
  - N_c is the average number of bounding boxes (par image) associated to c (evaluated offline from the training set)



Blind Localization Proposals (III)

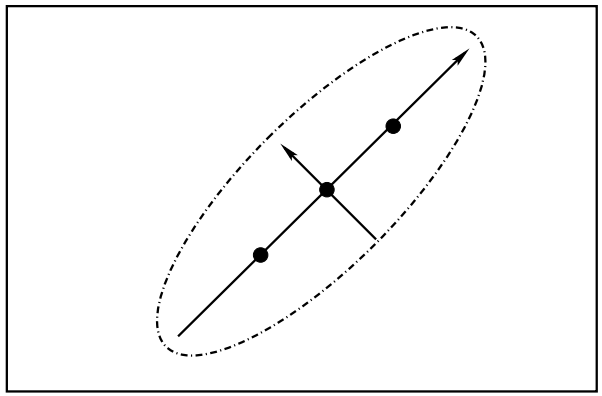
- **Heuristic 3 (rescaled concept-dependent BBs) :** each bounding box $p_c = (x, y, w, h)$ generated in heuristic 2, is replaced by re-scaled BB.
  - First, PCA is applied offline to the BB dimensions $\{(w_i, h_i)\}_i$ in the training set that also belong to concept c,
  - Then, (w, h) of p_c are moved towards the first principal component of PCA (i.e., the eigenvector with the largest eigenvalue), with an amplitude proportional to its eigenvalue
  - this corresponds a re-scale of the dimensions of p_c.
  - In this heuristic (x, y) remains unchanged.

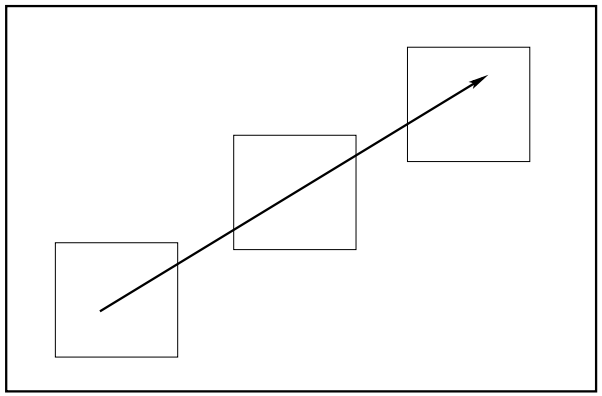




Blind Localization Proposals (IV)

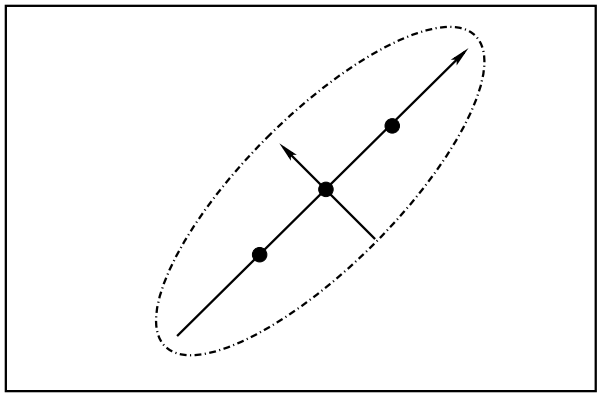
- **Heuristic 4 (shifted concept-dependent BBs) :** for each bounding box $p_c = (x, y, w, h)$ generated in heuristic 2, we generate two extra BBs, with shifted coordinates.
- Again, PCA is applied offline to the BB coordinates $\{(x_i, y_i)\}_i$ in the training set that also belong to concept c,
- Then, (x, y) of p_c are shifted towards two opposite directions corresponding to the first principal component of PCA.
- In this heuristic (w, h) remains unchanged.

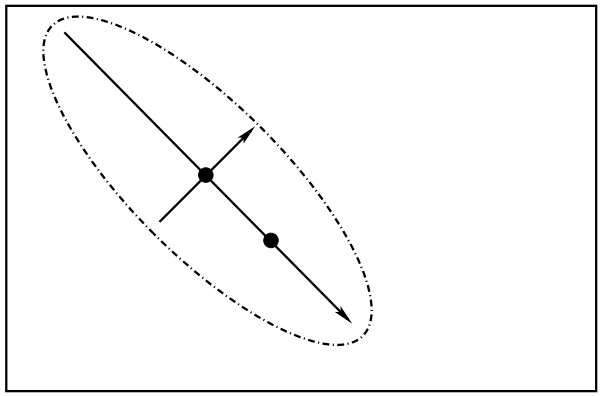


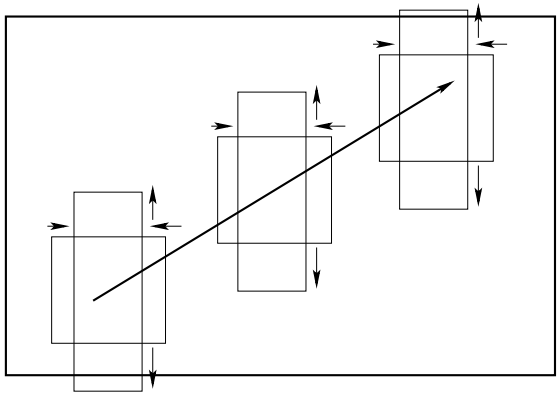


Blind Localization Proposals (V)

- **Heuristic 5 (shifted and re-scaled concept-dependent BBs) :** this heuristic corresponds to the combination of the two heuristics 3 and 4.



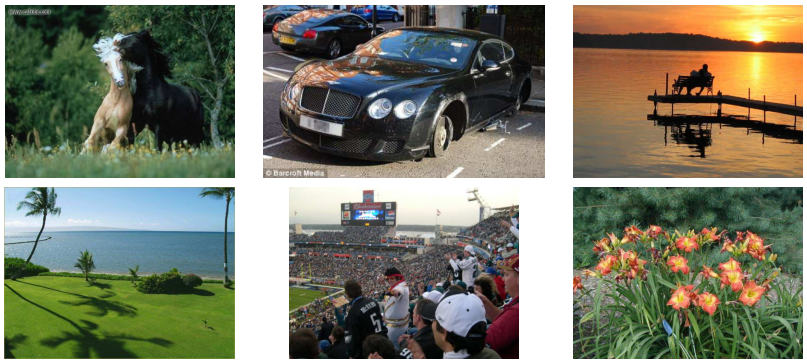




Outline

ImageCLEF 2015 Benchmark

- 500k images, 251 categories, only 2k labeled.



- Images are described using 9 visual features provided.
- Extra textual feature (normalized vector space model) ; first, a vocabulary of keywords $\mathcal{V}$ is defined to query the associated meta-data. For each keyword $\omega \in \mathcal{V}$, only images whose textual descriptions include ω have their ω vector entry set to non-zeros.
- Performance measured using MAP based on different percentages of bounding box overlaps.

ImageCLEF 2015 : Training data generation

- A keyword ω (in "concept.txt") is added to $\mathcal{V}$ iff (1/IDF) score is high. ω is also called attribute.

  n02691156 airplane.n.01 airplane,aeroplane,plane ["an aircraft that has a fixed wing and is powered by propellers or jets"]

- We applied some very simple morphological expansions to words in $\mathcal{V}$: leaf -> leaves, etc.
- We define a matrix of relations "concepts/attributes" $\mathbf{A}$, with $\mathbf{A}_{c,\omega} = 1$ iff the keyword $\omega \in \mathcal{V}$ exists in the definition of the concept c in the file "concept.txt".
- For a given concept (without training data), we extract a training set, by collecting among the 500k images those which include its attributes, in their meta-data files, i.e.,

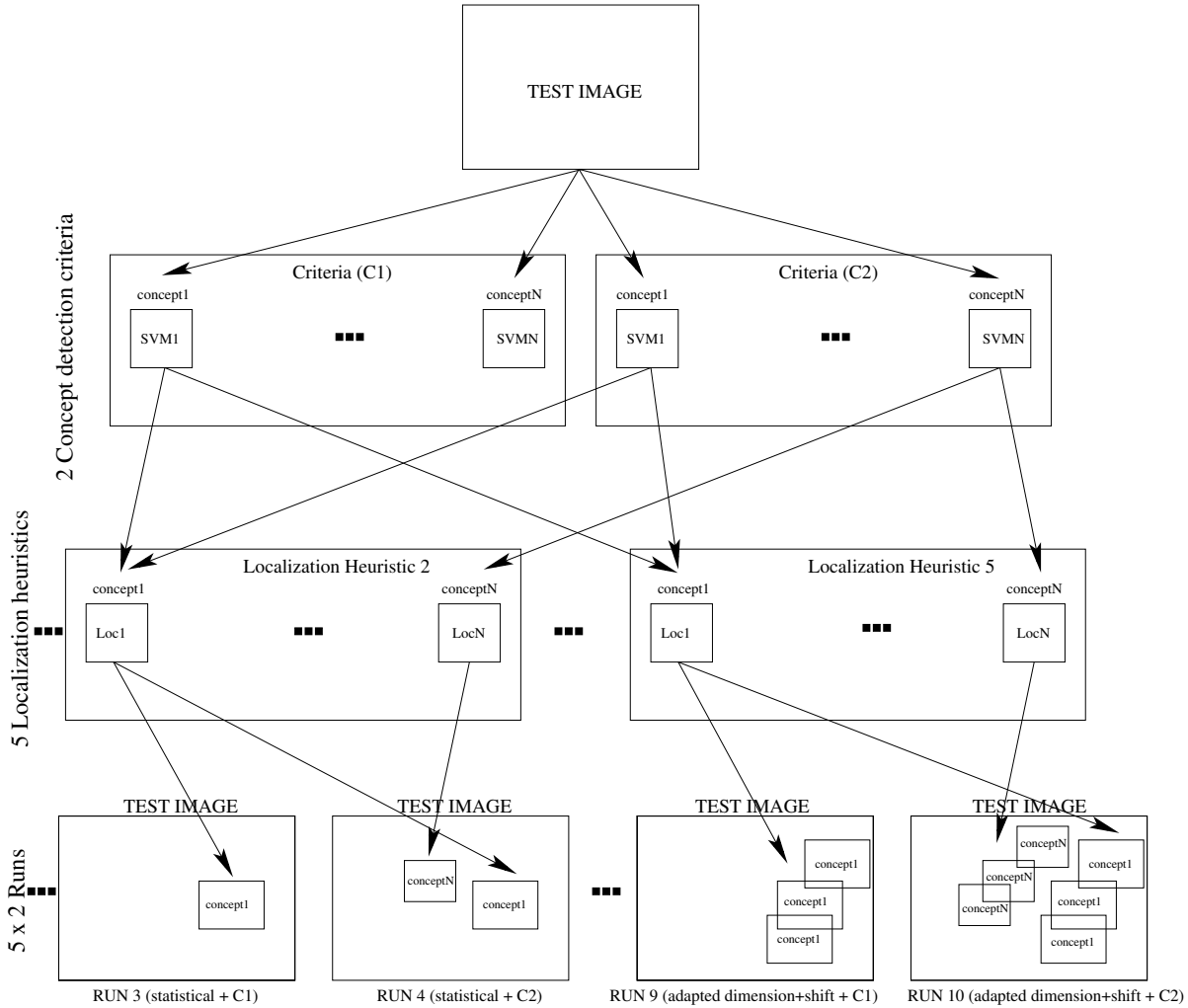
$$\mathbf{Y}_{c,i} = 1_{\{\sum_{\omega} \mathbf{A}_{c,\omega}\mathbf{M}_{\omega,i} \geq \tau\}}.$$

ImageCLEF 2015 : Submitted Runs (I)

- Nbr of submitted runs : 10 based on a combination of 2 concept detection criteria (SVMs + HI kernel) and 5 localization heuristics
  - **Criterion 1 (C1) :** concept detection results are obtained as described earlier.
  - **Criterion 2 (C2) :** if an image has no detected concepts, then we select the top 3 concepts (i.e., with the highest negative SVM scores) as annotations.

	Heuristic 1	**Heuristic 2**	**Heuristic 3**	**Heuristic 4**	**Heuristic 5**
Criterion 1 (C1)	Run 1	Run 3	Run 5	Run 7	Run 9
Criterion 2 (C2)	Run 2	Run 4	Run 6	Run 8	Run 10

ImageCLEF 2015 : Submitted Runs (II)



ImageCLEF 2015 : Performance (I)

Runs # \ Overlap	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%
5 (heuristic 3+C1)	**30.73**	**27.64**	**25.11**	**22.47**	**19.80**	**16.92**	**14.68**	**12.46**	**10.05**	**07.85**
9 (heuristic 5+C1)	**30.73**	27.21	24.87	21.69	19.01	16.58	14.08	11.65	09.32	07.62
3 (heuristic 2+C1)	**30.73**	26.13	24.48	21.52	18.72	15.82	13.01	10.30	08.03	06.32
7 (heuristic 4+C1)	**30.73**	25.73	23.50	20.58	17.77	14.61	11.80	09.38	07.38	05.55
1 (heuristic 1+C1)	**30.73**	26.11	20.79	16.81	13.29	10.49	08.53	06.81	04.99	03.17
6 (heuristic 3+C2)	19.40	17.39	15.83	14.08	12.44	10.63	09.25	08.00	06.56	05.20
10(heuristic 5+C2)	19.40	17.21	15.71	13.83	12.10	10.38	08.90	07.52	06.10	05.01
4 (heuristic 2+C2)	19.40	16.35	15.31	13.56	11.98	10.11	08.33	06.87	05.40	04.15
8 (heuristic 4+C2)	19.40	16.23	14.91	13.17	11.41	09.48	07.83	06.16	04.92	03.68
2 (heuristic 1+C2)	19.40	16.30	13.05	10.53	08.44	06.73	05.53	04.51	03.34	02.21

- Obviously, better decoupled concept detection provides better localization (further better concept detection should further improve the results).
- BB rescaling (heuristic 3) provides the best overall performances ; even though shifting is important, it has less impact on performances : this is due to the non-rigidity of many concepts (such as animals) while shifting is already captured by the statistical model.

ImageCLEF 2015 : Performance (II)

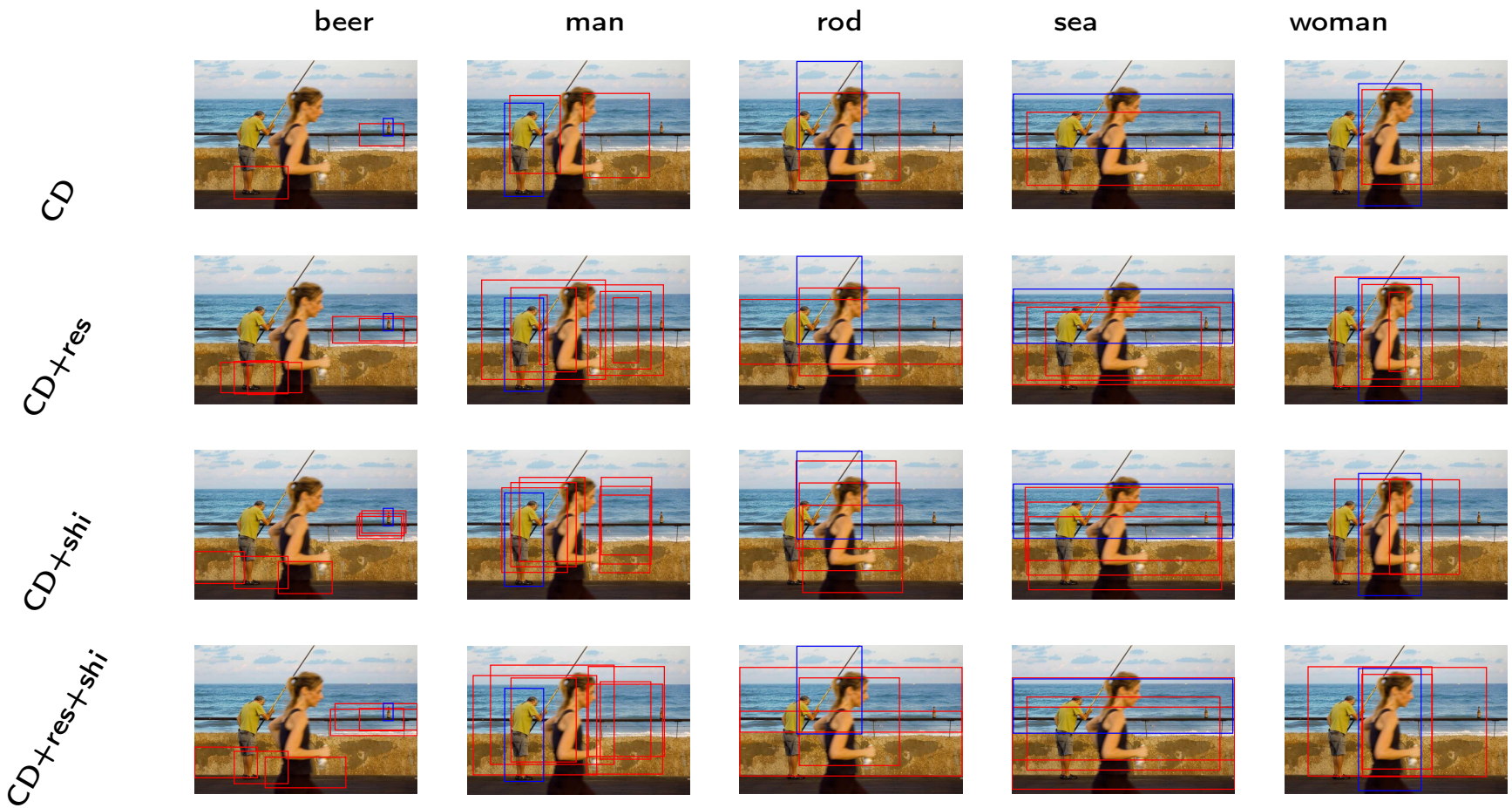
concepts	description	(run 5) statistical+rescale	(run 9) statistical+rescale+shift	(run 3) statistical	(run 7) statistical+shift	(run 1) fixed	(run 6) statistical+rescale	(run 10) statistical+rescale+shift	(run 4) statistical	(run 8) statistical+shift	(run 2) fixed
n01639765	frog	18.18	**36.36**	**36.36**	27.27	18.18	18.18	27.27	45.45	27.27	18.18
n01896031	feather	20.00	20.00	**40.00**	**40.00**	20.00	20.00	20.00	**40.00**	**40.00**	20.00
n02084071	dog	**50.00**	**50.00**	**50.00**	**50.00**	**50.00**	33.33	33.33	33.33	33.33	33.33
n02114100	wolf	22.86	20.00	**28.57**	25.71	20.00	22.86	20.00	**28.57**	25.71	20.00
n02129165	lion	0	0	0	0	0	**02.22**	**02.22**	**02.22**	**02.22**	**02.22**
n02131653	bear	0	0	0	0	**50.00**	0	0	0	0	**50.00**
n02206856	bee	**66.67**	**66.67**	**66.67**	**66.67**	**66.67**	50.00	50.00	50.00	50.00	50.00
n02330245	mouse	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	50.00	50.00	50.00	50.00	50.00
n02395406	hog	40.00	40.00	**52.00**	**52.00**	36.00	40.00	40.00	**52.00**	**52.00**	36.00
n02411705	sheep	**52.94**	**52.94**	35.29	41.18	47.06	**52.94**	**52.94**	23.53	29.41	47.06
n02416519	goat	**50.00**	**50.00**	0	0	**50.00**	33.33	33.33	0	0	33.33
n02430045	deer	**25.00**	**25.00**	**25.00**	**25.00**	0	**25.00**	**25.00**	**25.00**	**25.00**	0
n02484322	monkey	57.14	57.14	**64.29**	**64.29**	50.00	52.94	52.94	58.82	58.82	47.06
n02503517	elephant	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	28.57	28.57	28.57	28.57	14.29
n02512053	fish	60.00	60.00	**70.00**	60.00	60.00	46.67	46.67	53.33	46.67	40.00
n02691156	airplane	0	**100.0**	0	**100.0**	0	0	0	33.33	33.33	0
n02709367	anchor	63.16	**73.68**	42.11	47.37	52.63	63.64	63.64	45.45	59.09	54.55
n02774152	bag	**10.00**	**10.00**	0	0	**10.00**	07.69	07.69	0	0	07.69
n02778669	ball	**16.67**	**16.67**	**16.67**	**16.67**	0	12.50	12.50	12.50	12.50	0
n02782093	balloon	0	**05.26**	0	**05.26**	0	0	**05.26**	0	**05.26**	0
n02800213	baseball	0	0	0	0	0	**01.49**	**01.49**	**01.49**	**01.49**	**01.49**
n02828884	bench	20.00	20.00	**25.00**	**25.00**	20.00	20.00	20.00	**25.00**	**25.00**	20.00
n02834778	bicycle	10.00	05.00	10.00	05.00	05.00	**13.16**	05.26	10.53	05.26	07.89
n02839910	bin	30.43	30.43	30.43	30.43	30.43	30.43	30.43	30.43	30.43	30.43
n02883344	box	0	0	0	0	0	**04.35**	**04.35**	0	**04.35**	0
n02909870	bucket	46.67	**53.33**	**53.33**	**53.33**	**53.33**	41.18	47.06	47.06	47.06	47.06
n02933112	cabinet	60.00	60.00	**80.00**	**80.00**	60.00	24.24	21.21	27.27	24.24	21.21
n02942699	camera	0	05.26	0	05.26	0	05.41	05.41	02.70	**08.11**	05.41
n02984061	cathedral	36.76	**37.50**	13.97	08.82	34.56	36.76	**37.50**	15.44	16.18	34.56
n02990373	ceiling	**36.36**	**36.36**	**36.36**	**36.36**	**36.36**	23.26	23.26	23.26	25.58	18.60
n03001627	chair	0	0	0	0	0	10.26	10.26	10.26	**15.38**	0
n03046257	clock	**11.32**	09.43	07.55	07.55	05.66	09.43	09.43	09.43	07.55	05.66
n03135532	cross	**25.00**	0	**25.00**	0	0	20.00	0	20.00	0	0

ImageCLEF 2015 : Performance (III)

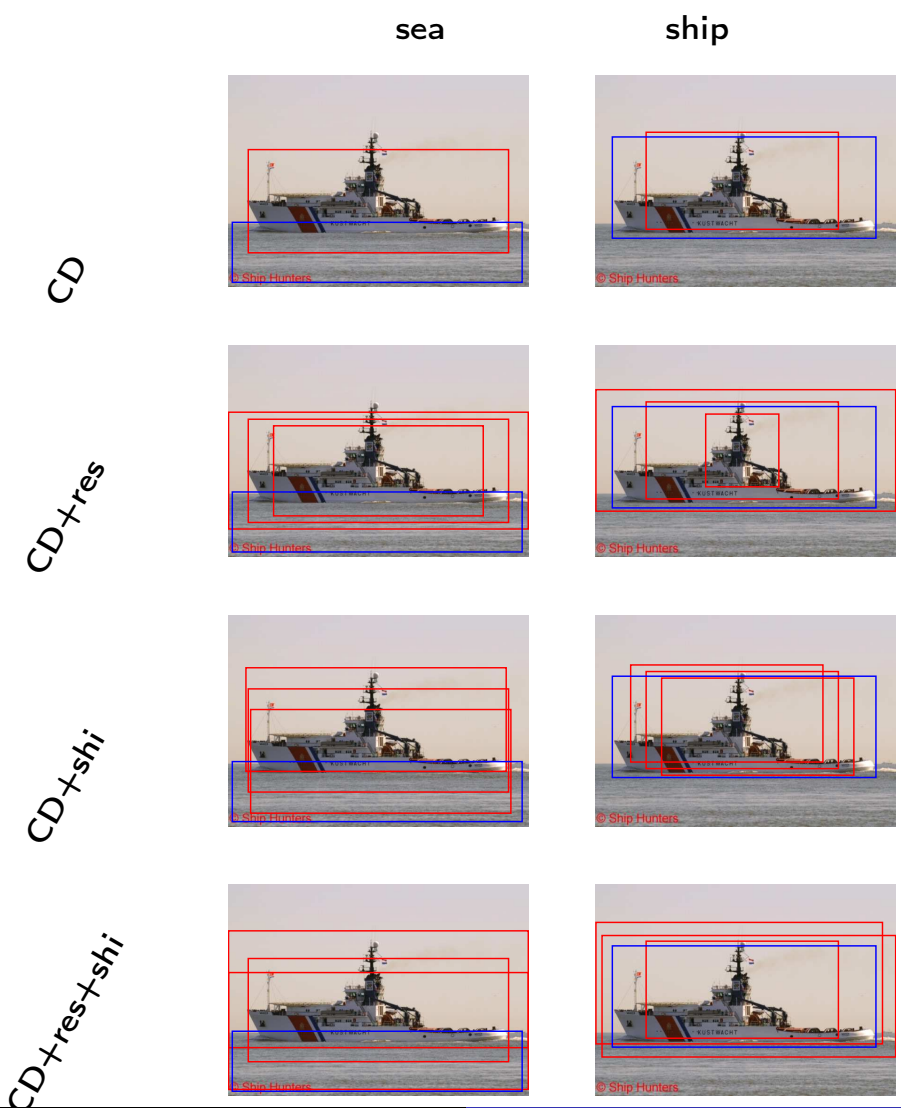
- for some concepts such as "frog", re-scaling and shifting are important, as this category is highly non-rigid while for other categories such as "bear" localization is less predictable.
- for rigid (and man-made) objects, such as "cathedral" and "bicycle", re-scaling is more important than shifting as the proportions of the w-h dimensions are very changing.
- while for others ("airplane", "balloon", "bucket", "camera"), the adaptation of shift is more important than scale ; as the variability of w-h proportions is small.
- In sum, BB re-scaling and shifting is important for some concepts. This suggests to mix heuristics for some concepts (we already observe gain in "concept-by-concept" results).

Concept Localization : Examples (I)

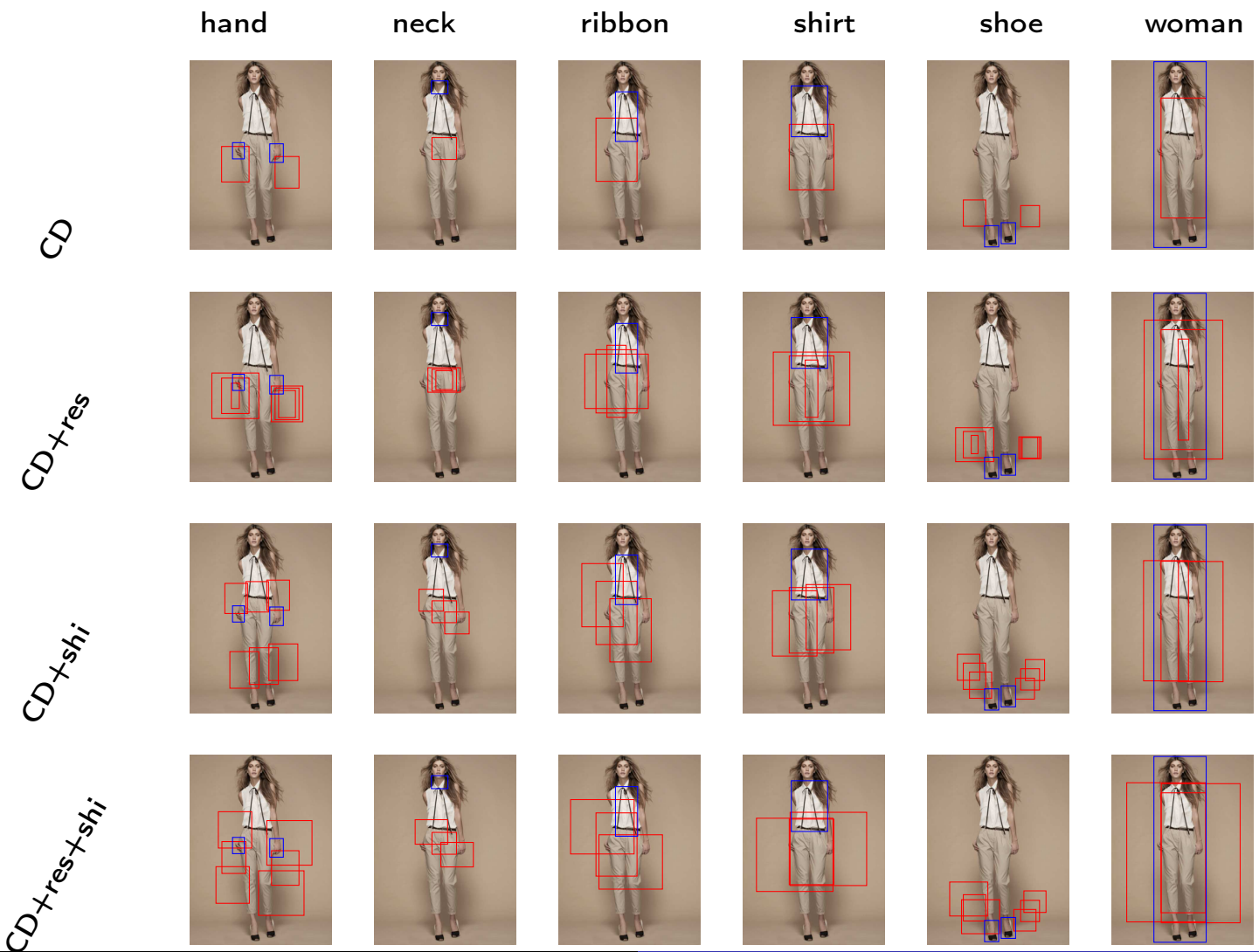
Concept Localization : Examples (II)



Concept Localization : Examples (III)



Concept Localization : Examples (IV)



Outline

Conclusion and Extensions

- Our runs are based on a two-step process that decouples concept detection from localization.
- The former is achieved using SVMs trained with linear combination of HIK, while the latter is accomplished blindly using a simple statistical model that allows us to generate multiple localization proposals (without image segmentation).
- Observed results show that i) the accuracy of concept detection has an impact on the performance of localization, and ii) the adaptation of scale and shift of concept localization is essential to improve performances mainly for some concepts.
- A future extension, how to make concept localization non-blind and also coupled with concept detection, consider interaction statistics. Another possible extension is to mix and select different localization heuristics for different concepts.