Segmentation Driven Object Detection with Fisher Vectors

Camille Brasseur

20 décembre 2013

Aim of the work

Object detection :

The aim is to determine for an object :

- its location (bounding box)
  and
- its category

Aim of the work

Object detection :

The aim is to determine for an object :

- its location (bounding box)
  and
- its category

Used tools :

- Ficher Vector
- SIFT descriptor
- color descriptor

Tests on datasets :

- PASCAL VOC 2007
- PASCAL VOC 2010

Object detection

Sliding Window approaches

Detection windows of various scale and aspect ratios evaluated at many positions access the image.

- (Viola and Jones) : cascade $\Rightarrow$ less windows to examine
- two or three-stage approaches : windows are discarded at each stage + richer features
- branch and bound scheme (non-exhaustive search)
- prune the set of candidate windows without using class specific information by relying on low-level contours and image segmentation

The last idea is used there.

Contributions

Fisher Vector

They were already used in previous approaches.
But here, normalization of the FVs.

Segmentation

- image segmentation created for the detection
- computation of a mask with a weight for each pixel linked with its contribution to the descriptors.
- suppression of the background

Segmentation

State of the art

- extraction of explicit segmentation for each object detection hypothesis
- scoring superpixels individually and then assemble them into object detections
- use of the output from a semantic segmentation to improve object detection.

Here :

segmentation incorporated into the feature extraction step

Segmentation

Steps

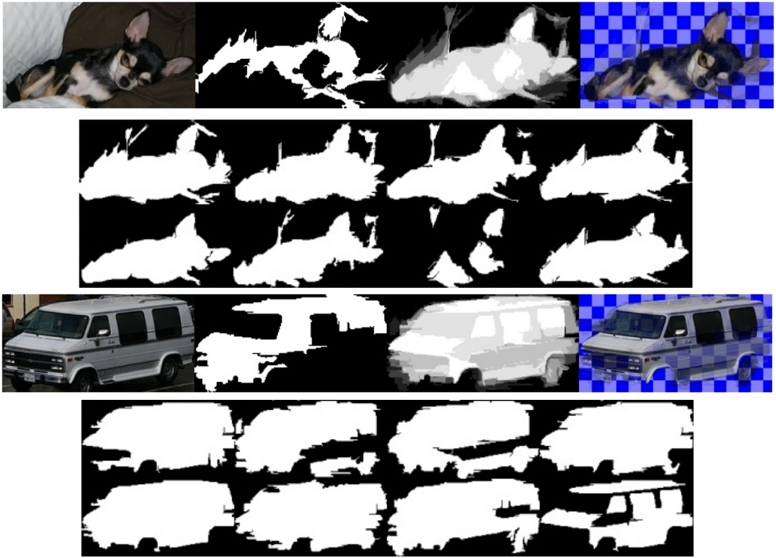
1. partition of the image into superpixels
2. hierarchically group the superpixel into a segmentation tree (merging neighboring and visually similar segments)

This is repeated eight times with

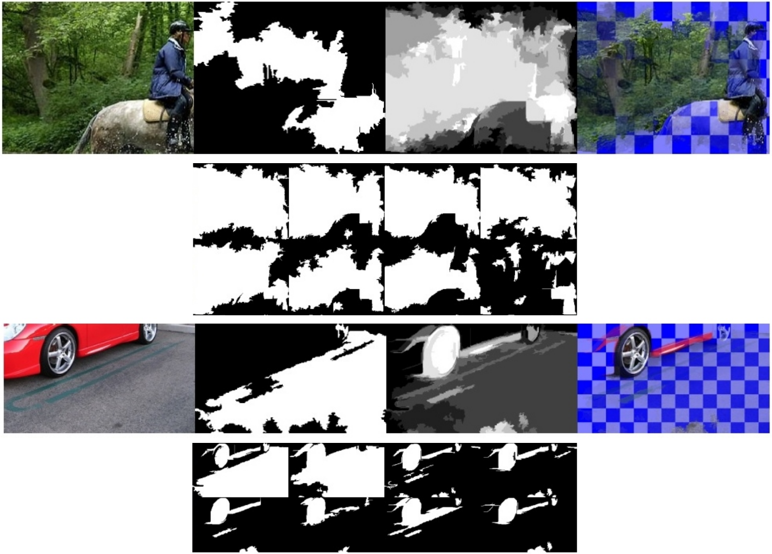
- 4 different color spaces and
- 2 different scale parameters

to compure the superpixels.
$\Rightarrow$ rich set of segments of varying sizes and shapes (around 1500 object windows per image)
It is far less windows than in a sliding window approach.

Correct examples



Incorrect examples



Feature extraction

local features :

- SIFT
- color descriptor

Aggregation

Using Fisher vector representation

Fisher vector

Normalized gradients

$$\frac{\partial \ln p(x)}{\partial \mu_{kd}} = \frac{p(k|x)}{\sqrt{\pi_k}} \left(\frac{x_d - \mu_{kd}}{\sigma_{kd}} \right) \quad (1)$$

$$\frac{\partial \ln p(x)}{\partial \sigma_{kd}} = \frac{p(k|x)}{\sqrt{\pi_k}} \left(\frac{(x_d - \mu_{kd})^2}{\sigma_{kd}^2} - 1 \right) \quad (2)$$

x local descriptor

μ_{kd} and σ_{kd} mean and standard derivation of the k-th Gaussian in dimension d

π_k mixing weight of the k-th Gaussian

$p(k|x)$ soft assignment of x to the k-th Gaussian

Candidate window

Representation :

1. sum the normalized gradients
2. weight the contribution of local descriptors by the averaged segmentation masks

Final window descriptor :

- concatenation of FV obtained over color and SIFT
- FV over the full image to capture global scene context

Compression

used tools

- Product Quantization
- Blosc compression

First test

Performance on the development set with different descriptors regions and with/without SPM

Desc.	Regions	Norm.	SPM	bus	cat	mbike	sheep	mAP
S	W	object	no	22.2	35.8	26.3	16.6	25.2
S	W	object	yes	47.6	45.0	54.2	30.0	44.2
S	W	cell	yes	48.0	47.2	53.0	32.0	45.0
S	G (train on W)	cell	yes	35.7	46.3	43.2	17.0	35.5
S	M (train on W)	cell	yes	41.1	47.8	52.7	19.2	40.2
S	M	cell	yes	44.0	48.8	51.4	30.8	43.8
S	W+M	cell	yes	48.5	49.2	54.3	33.8	46.4
S+C	W	cell	yes	47.3	48.2	54.4	35.8	46.4
S+C	W+M	cell	yes	48.1	51.1	55.5	40.0	48.7
S+C	W+M+F	cell	yes	50.3	51.6	54.8	41.9	49.6

Second test

Performance on VOC07 with different descriptors and regions.

		aero	bicy	bird	boa	bot	bus	car	cat	cha	cow	dtab	dog	hors	mbik	pers	plnt	she	sofa	trai	tv	mAP
S	W	46.7	48.7	14.1	19.4	15.7	45.0	54.6	36.3	11.4	36.2	37.4	24.3	37.1	52.4	25.8	14.7	35.3	30.4	47.2	48.2	34.0
S	W+M	50.2	49.4	16.6	21.3	15.7	45.5	55.3	39.8	14.8	36.3	39.5	25.4	42.4	50.4	**30.6**	15.8	34.3	35.5	48.3	49.7	35.8
S+C	W	47.7	50.1	16.5	19.2	15.9	45.1	55.1	37.2	13.0	37.3	40.8	25.5	40.7	51.8	26.4	18.2	35.5	30.6	47.7	49.6	35.2
S+C	W+M	50.5	51.2	18.8	23.8	17.8	47.2	56.4	41.6	14.7	38.6	40.7	27.1	47.3	52.4	29.7	19.6	38.3	35.0	49.3	52.8	37.6
S+C	W+F	49.9	51.6	16.4	21.7	16.5	45.9	55.6	38.4	15.3	**42.1**	42.0	25.3	41.2	52.2	26.8	18.8	36.2	35.8	48.5	51.6	36.6
S+C	W+M+F	52.6	52.6	19.2	25.4	18.7	47.3	56.9	42.1	**16.6**	41.4	41.9	27.7	47.9	51.5	29.9	20.0	**41.1**	36.4	48.6	53.2	38.5
S+C	W+M+F+Context	**56.1**	**56.4**	**21.8**	**26.8**	**19.9**	**49.5**	**57.9**	**46.2**	16.4	41.4	**47.1**	**29.2**	**51.3**	**53.6**	28.6	**20.3**	40.5	**39.6**	**53.5**	**54.3**	**40.5**

Third test

Comparison of this detector with and without context with the state-of-the-art object detectors on VOC 2007.

	aero	bicy	bird	boa	bot	bus	car	cat	cha	cow	dtab	dog	hors	mbik	pers	plnt	she	sofa	trai	tv	mAP
	methods without inter-class contextual cues																				
SUGS'11 [34]	43.3	46.4	11.2	11.9	9.3	49.3	53.7	39.2	12.5	36.8	42.0	26.4	47.0	52.1	23.5	11.9	29.7	36.1	42.0	48.7	33.7
HMR'12 [19]	23.3	41.0	9.9	11.0	17.0	37.8	38.4	11.5	11.8	14.5	12.2	10.2	44.8	27.9	22.4	3.1	16.3	8.9	30.3	28.8	21.0
VZ'12 [36]	27.9	55.2	9.5	10.4	16.4	47.6	52.0	16.0	13.5	18.6	20.7	10.7	53.4	39.7	37.3	10.4	12.7	19.7	41.7	40.9	27.7
GFM'12 [16]	33.2	60.3	10.2	16.1	27.3	54.3	58.2	23.0	20.0	24.1	26.7	12.7	58.1	48.2	43.2	12.0	21.1	36.1	46.0	43.5	33.7
KAWBVL'12 [22]	34.5	61.1	11.5	19.0	22.2	46.5	58.9	24.7	21.7	25.1	27.1	13.0	59.7	51.6	44.0	19.2	24.4	33.1	48.4	49.7	34.8
SWJZ'13 [32]	35.3	60.2	16.6	**29.5**	**53**	57.1	49.9	**48.5**	11	23	27.7	13.1	58.9	22.4	41.4	16	22.9	28.6	37.2	42.4	34.7
Ours, without context	52.6	52.6	19.2	25.4	18.7	47.3	56.9	42.1	16.6	**41.4**	41.9	27.7	47.9	51.5	29.9	20.0	**41.1**	36.4	48.6	53.2	38.5
	methods using inter-class contextual cues																				
GFM'12 context [16]	36.6	62.2	12.1	17.6	28.7	54.6	60.4	25.5	21.1	25.6	26.6	14.6	60.9	50.7	44.7	14.3	21.5	38.2	49.3	43.6	35.4
CDXH'13 [5]	41.0	**64.3**	15.1	19.5	33.0	**57.9**	**63.2**	27.8	23.2	28.2	29.1	16.9	**63.7**	**53.8**	**47.1**	18.3	28.1	**42.2**	53.1	49.3	38.7
Ours, with context	**56.1**	56.4	**21.8**	26.8	19.9	49.5	57.9	46.2	16.4	**41.4**	**47.1**	**29.2**	51.3	53.6	28.7	**20.3**	40.5	39.6	53.5	**54.3**	**40.5**

Fourth test

Comparison of our detector with and without context with the state-of-the-art object detectors on VOC 2010.

	aero	bicy	bird	boa	bot	bus	car	cat	cha	cow	dtab	dog	hors	mbik	pers	plnt	she	sofa	trai	tv	mAP
	methods without inter-class contextual cues																				
SUGS'11 [34]	58.2	41.9	19.2	14.0	14.3	44.8	36.7	48.8	12.9	28.1	28.7	39.4	44.1	52.5	25.8	**14.1**	38.8	34.2	43.1	42.6	34.1
GFM'12 [16]	45.6	49.0	11.0	11.6	27.2	50.5	43.1	23.6	17.2	23.2	10.7	20.5	42.5	44.5	41.3	8.7	29.0	18.7	40.0	34.5	29.6
SWJZ'13 [32]	44.6	48.5	12.9	**26.3**	47.5	41.6	45.3	39	10.8	21.6	23.6	22.9	40.9	30.4	37.9	9.6	17.3	11.5	25.3	31.2	29.4
Ours, without context	61.3	46.4	21.1	21.0	18.1	49.3	45.0	46.9	12.8	29.2	26.1	38.9	40.4	53.1	31.9	13.3	39.9	33.4	43.0	45.3	35.8
	methods using inter-class contextual cues																				
NLPR 2010 *	53.3	**55.3**	19.2	21.0	30.0	54.4	46.7	41.2	**20.0**	31.5	20.7	30.3	48.6	55.3	**46.5**	10.2	34.4	26.5	50.3	40.3	36.8
SCHHY'11 [33]	53.1	52.7	18.1	13.5	**30.7**	53.9	43.5	40.3	17.7	31.9	28.0	29.5	**52.9**	**56.6**	44.2	12.6	36.2	28.7	**50.5**	40.7	36.8
GFM'12 context [16]	48.2	52.2	14.8	13.8	28.7	53.2	44.9	26.0	18.4	24.4	13.7	23.1	45.8	50.5	43.7	9.8	31.1	21.5	44.4	35.7	32.2
Ours, with context	**65.9**	50.1	**23.7**	24.1	20.4	52.6	**47.1**	**50.9**	13.2	**32.8**	**31.8**	**41.4**	43.9	55.3	29.8	**14.1**	**41.7**	**35.6**	46.7	**46.9**	**38.4**
	uncomparable methods using additional training data																				
FMYU'13 [15] **	56.4	48.0	24.3	21.8	31.3	51.3	47.3	48.2	16.1	29.4	19.0	37.5	44.1	51.5	44.4	12.6	32.1	28.8	48.9	39.1	36.6
FMYU'13 context [15] **	61.4	53.4	25.6	25.2	35.5	51.7	50.6	50.8	19.3	33.8	26.8	40.4	48.3	54.4	47.1	14.8	38.7	35.0	52.8	43.1	40.4