Fisher and VLAD with FLAIR

Koen E. A. van de Sande[12], Cees G. M. Snoek[12], Arnold W. M. Smeulders[123]

[1]University of Amsterdam, [2]Euvision Technologies, [3]Centrum Wiskunde & Informatica

UNIVERSITEIT VAN AMSTERDAM

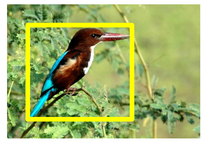
EUVISION technologies

http://koen.me/research/flair

Object Recognition

Object recognition seeks to answer 2 questions:

- What is it?
- **Where is it?**



Box Encodings

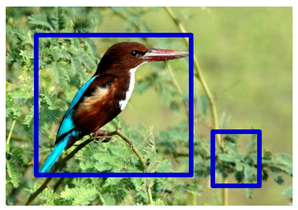
State-of-the-art encodings:
- VLAD [JegouTPAMI12]
- Fisher [PerronninECCV10]

Current state-of-the-art for recognition:
- VOC'10 [CinbisICCV13]
- VOC'12 [SandeVOC12]
- ImageNet-DET'13 (this work)

Drawback: Computationally expensive

Key Problems

Encoding a box depends on the #pixels inside



Large boxes are slower to encode than small

Existing approaches:
- Integral images compute the sum of an area in O(1)
- Multidimensional integral images [PorikliCVPR05] extend this to histograms
- Approximate at significant accuracy costs

For VLAD encoding, integral histograms require 14GB of memory and minutes to compute per image: **unusable in practice**.

For Fisher, there are additional complexities with multiple assignment, power and L_2 normalizations

FLAIR solves these problems. It was our secret ingredient to winning the 2013 IMAGENET 200 Object Detection Challenge

References
- Cinbis et al. *Segmentation driven object detection with fisher vectors* (ICCV2013)
- Jegou et al. *Aggregating local image descriptors into compact codes* (TPAMI2012)
- Perronnin et al. Improving the fisher kernel for large-scale image classification (ECCV2010)
- Porikli. *Integral histogram: A fast way to extract histograms in cartesian spaces* (CVPR2005)
- Uijlings et al. *Selective Search for Object Recognition* (IJCV2013)
- Van de Sande et al. *Hybrid Coding for Selective Search* (PASCAL VOC2012)
- Wang et al. *Regionlets for Generic Object Detection* (ICCV2013)

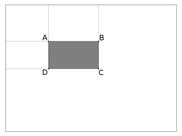
Fast Local Area Independent Representation

Goal: Construct feature vector for *B* boxes with *S* spatial pyramid cells
- *N* descriptors of length *D* in the image of size (*W*,*H*)

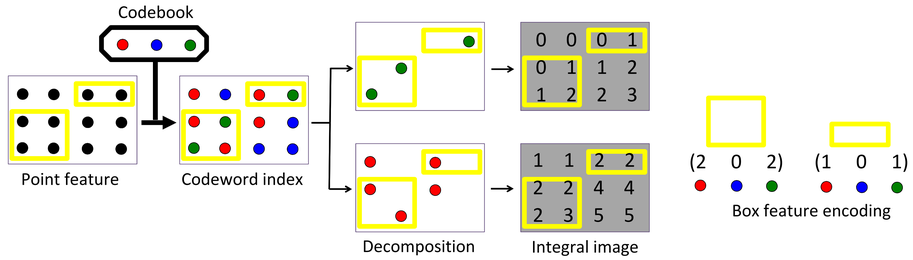
$$\phi(\vec{v}_n, \vec{c}_k) = \begin{cases} 1 & \text{if } k \text{ is the closest codeword for } \vec{v}_n \\ 0 & \text{otherwise} \end{cases}$$

The *k*-th feature vector element: $F_k = \sum_{n=1}^{N} \phi(\vec{v}_n, \vec{c}_k)$

FLAIR for standard Bag-of-Words:
- Decompose over *K* codewords
- *K* integral images to create feature vector in *O(K)*
- Complexity: from *O(NBS)* to *O(NK+KBS)*
- Overhead of making integral images limits practical use
- SIFT-only VOC2007: meanAP ~ 32.3



Finding the sum of a box:
C+A-B-D

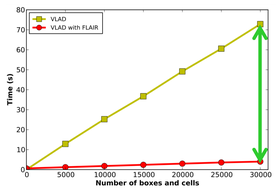


VLAD with FLAIR

One descriptor now affects *D* elements of feature vector:

$$\vec{F}_k = \sum_{n=1}^{N} (\vec{v}_n - \vec{c}_k)\phi(\vec{v}_n, \vec{c}_k)$$

VLAD with FLAIR:
- Uses multidimensional integral images
- Exploits sparsity due to decomposition
  - 79% of rows/columns is empty
  - Memory usage down to 1GB from 14GB
- Support power and *Lp* normalizations
- **Evaluate VLAD independent of box area**: *O(KWHD+KDBS)*



18x speedup
VOC'07 meanAP ~ 28.2

Fisher with FLAIR

Fisher is the normalized gradient of the loglikelihood under a Mixture-of-Gaussians.

For codeword (Gaussian) *k*:

$$\frac{\Delta \ln p}{\Delta \vec{\mu}_k} = \frac{1}{N} \sum_{n=1}^{N} \frac{p(k|\vec{v}_n)}{\sqrt{\pi_k}} \left(\frac{\vec{v}_n - \vec{\mu}_k}{\vec{\sigma}_k} \right),$$

$$\frac{\Delta \ln p}{\Delta \vec{\sigma}_k} = \frac{1}{N} \sum_{n=1}^{N} \frac{p(k|\vec{v}_n)}{\sqrt{\pi_k}} \left(\frac{(\vec{v}_n - \vec{\mu}_k)^2}{\vec{\sigma}_k^2} - 1 \right)$$

Decomposition over *k* still possible, even with multiple assignment. Consider:

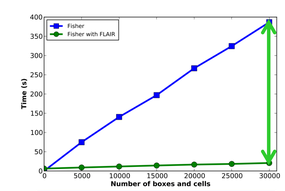
$$S_0(k) = \sum_{n=1}^{N} p(k|\vec{v}_n), \qquad \vec{S}_1(k) = \sum_{n=1}^{N} p(k|\vec{v}_n)\vec{v}_n, \qquad \vec{S}_2(k) = \sum_{n=1}^{N} p(k|\vec{v}_n)\vec{v}_n^2.$$

Rewrite using these to:

$$\frac{\Delta \ln p}{\Delta \vec{\mu}_k} = \frac{1}{N\sqrt{\pi_k}\vec{\sigma}_k} \left(\vec{S}_1(k) - \vec{\mu}_k \cdot S_0(k) \right),$$

$$\frac{\Delta \ln p}{\Delta \vec{\sigma}_k} = \frac{1}{N\sqrt{\pi_k}} \left(\frac{\vec{S}_2(k) - 2\vec{\mu}_k\vec{S}_1(k) + \vec{\mu}_k^2 \cdot S_0(k)}{\vec{\sigma}_k^2} - S_0(k) \right)$$

The scalar integral image $S_0(k)$ and the multidimensional integral images $S_1(k)$ and $S_2(k)$ are supplemented by a scalar integral image holding the number of descriptors *N* in an area. With these four integral images the gradients for a single codeword evaluate in *O(D)* and **independent of box area**, similar to VLAD



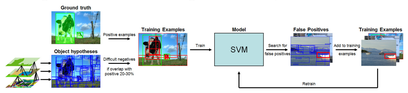
18x speedup
VOC'07 meanAP ~ 33.3

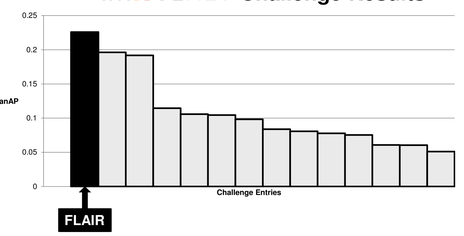
ImageNet 2013 Detection Challenge

Challenge has 200 object classes

Recognition system uses:
- Fisher with FLAIR
- Selective Search
- ColorSIFT

Using approved training data, outperforms CNNs



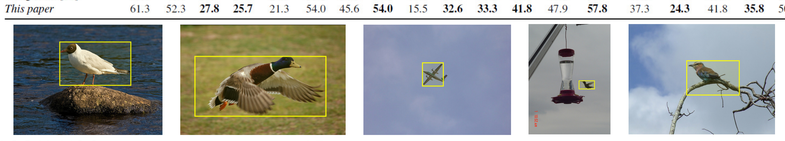


Elephant *Volleyball* *Accordion*

PASCAL VOC Object Detection

PASCAL VOC2010 test set (through independent evaluation server):
- We improve the state-of-the-art and perform best for 9 out of 20 objects
- Current #1 in the online leaderboard

System	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mAP
Felzenszwalb *et al.* [10]	52.4	54.3	13.0	15.6	**35.1**	54.2	49.1	31.8	15.5	26.2	13.5	21.5	45.4	51.6	**47.5**	09.1	35.1	19.4	46.6	38.0	33.7
Uijlings *et al.* [24]	56.2	42.4	15.3	12.6	21.8	49.3	36.8	46.1	12.9	32.1	30.0	36.5	43.5	52.9	32.9	15.3	41.1	31.8	47.0	44.8	35.1
Cinbis *et al.* [5]	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
NLPR	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
Wang *et al.* [31]	**65.0**	48.9	25.9	24.6	24.5	**56.1**	**54.5**	51.2	17.0	28.9	30.2	35.8	40.2	55.7	43.5	14.3	**43.9**	32.6	**54.0**	45.9	39.7
This paper	61.3	52.3	**27.8**	**25.7**	21.3	54.0	45.6	**54.0**	15.5	**32.6**	**33.3**	**41.8**	47.9	**57.8**	37.3	**24.3**	41.8	**35.8**	50.4	**47.3**	**40.4**



PASCAL VOC2012 test set: 40.6 meanAP

CUB200-2011 Bird Species Categorization

- Improved the state-of-the-art by over 20% for the real-world task where there are no ground truth bounding boxes on the test set
- Fisher with FLAIR opens doors to fine-grained object detection *without domain-specific feature optimization.*

System	Accuracy	
	No test boxes	With test boxes
Wah *et al.* [30]	10.3	17.3
Zhang *et al.* [32]	28.2	-
Berg *et al.* [2]	-	56.8
This paper	52.2	55.5

Conclusion

- FLAIR is the new data representation which enables fast encoding of arbitrary boxes in an image with the powerful VLAD and Fisher vectors:
  - Area independent
  - Supports spatial pyramids, power norm, L2 norm
  - No approximation
- Allows for large-scale detection
- State-of-the-art detection results on PASCAL VOC and ImageNet challenges