Large-scale visual recognition
Novel patch aggregation mechanisms

Florent Perronnin, XRCE

Hervé Jégou, INRIA

CVPR tutorial on Large-Scale Visual Recognition

June 16, 2012

Motivation

For large-scale visual recognition, we need image signatures which contain **fine-grained information**:

- in retrieval: the larger the dataset size, the higher the probability to find another similar but irrelevant image to a given query
- in classification: the larger the number of other classes, the higher the probability to find a class which is similar to any given class

BOV answer to the problem: increase visual vocabulary size

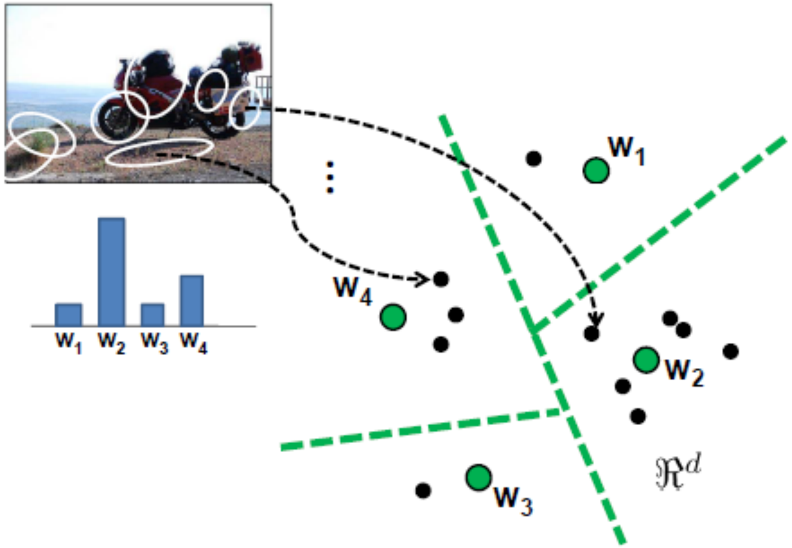
→ see previous part on scaling visual vocabularies

How to increase amount of information **without increasing the visual vocabulary size**?

Motivation

BOV is only about **counting** the number of local descriptors assigned to each Voronoi region

Why not including **other statistics**?



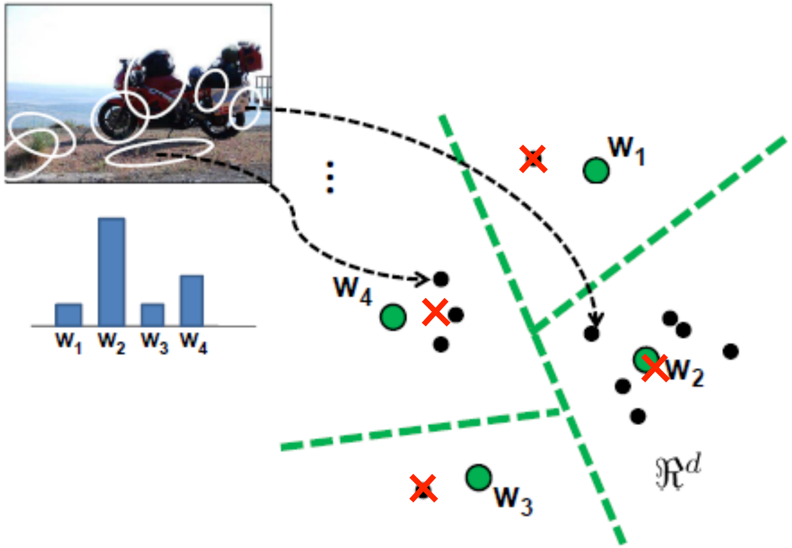
http://www.cs.utexas.edu/~grauman/courses/fall2009/papers/bag_of_visual_words.pdf

Motivation

BOV is only about **counting** the number of local descriptors assigned to each Voronoi region

Why not including **other statistics**? For instance:

- mean of local descriptors ×



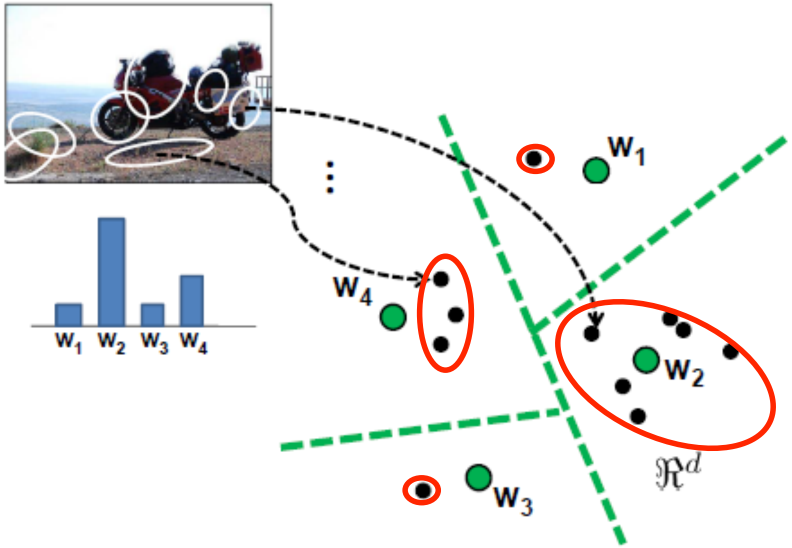
http://www.cs.utexas.edu/~grauman/courses/fall2009/papers/bag_of_visual_words.pdf

Motivation

BOV is only about **counting** the number of local descriptors assigned to each Voronoi region

Why not including **other statistics**? For instance:

- mean of local descriptors
- (co)variance of local descriptors



http://www.cs.utexas.edu/~grauman/courses/fall2009/papers/bag_of_visual_words.pdf

Outline

A first example: the VLAD

The Fisher Vector

Other higher-order representations

Example results

Outline

A first example: the VLAD

The Fisher Vector

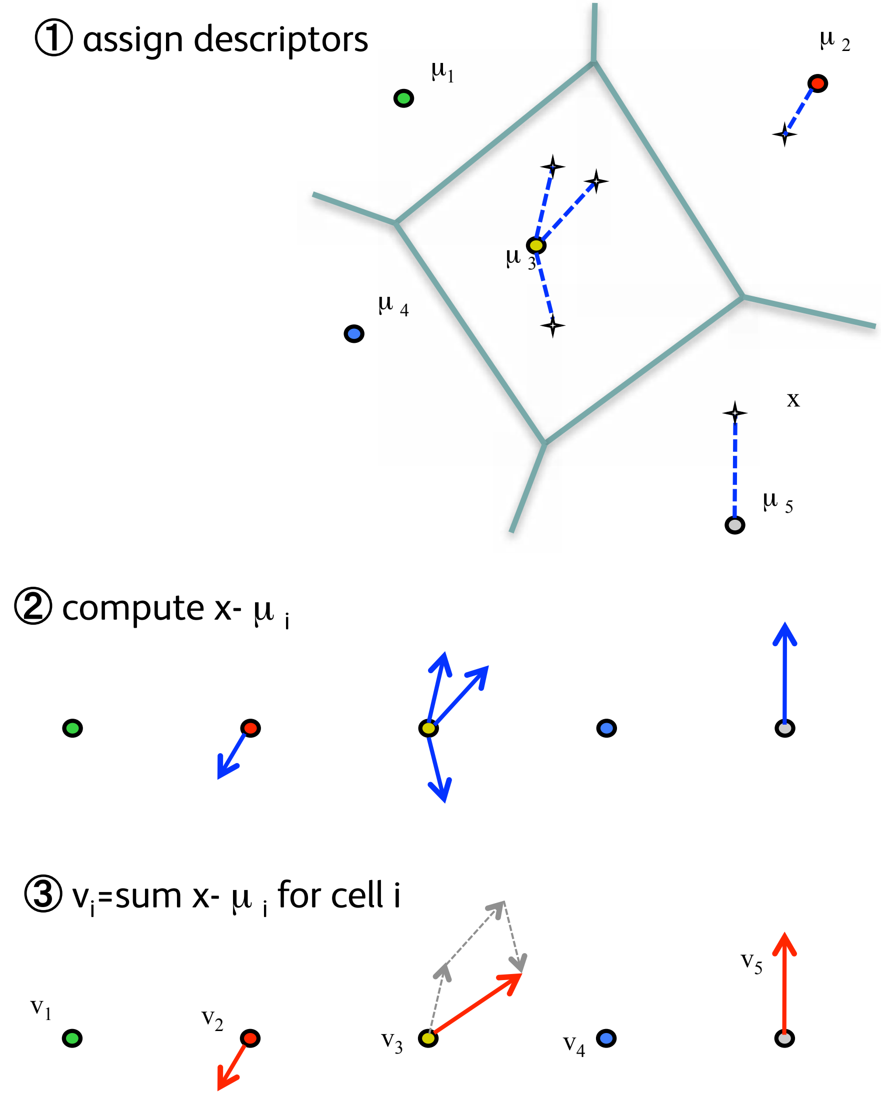
Other higher-order representations

Example results

A first example: the VLAD

Given a codebook $\{\mu_i, i = 1 \dots N\}$,
e.g. learned with K-means, and a set of local descriptors $X = \{x_t, t = 1 \dots T\}$:

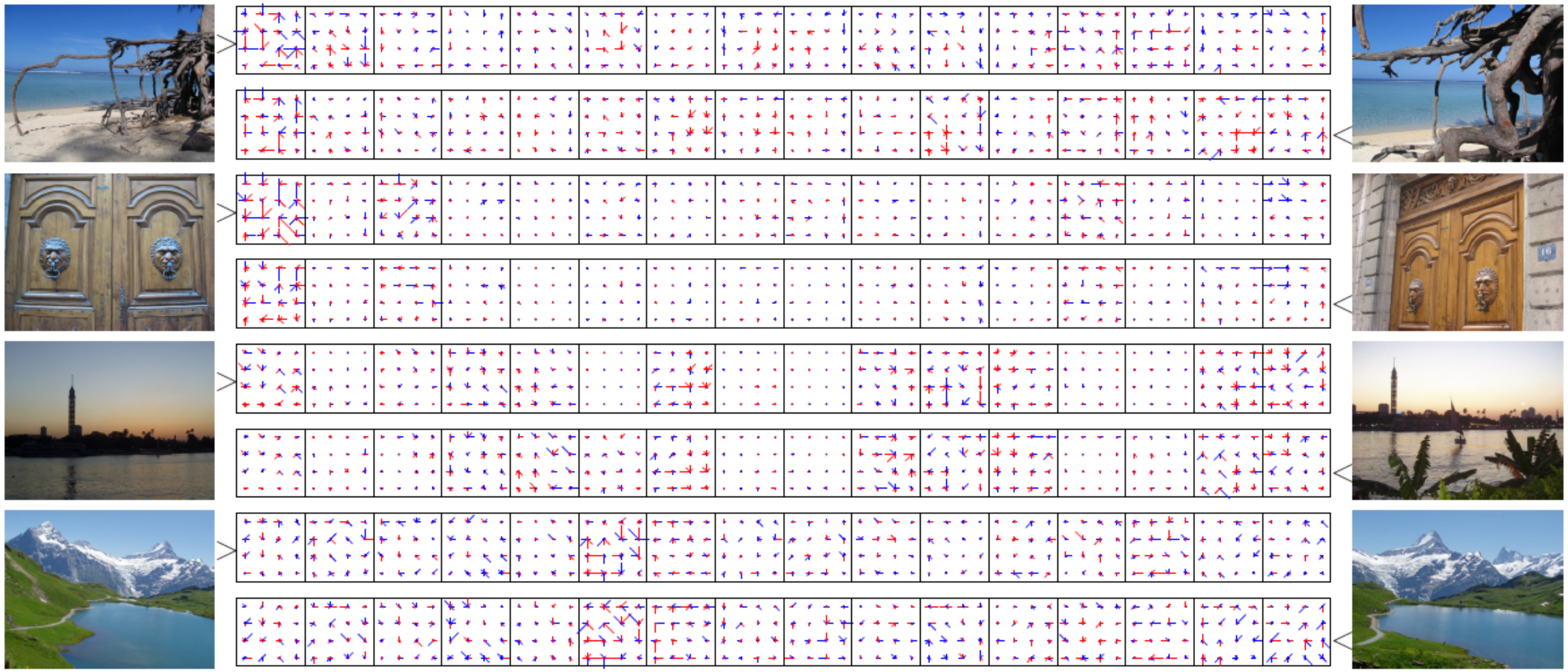
- ① assign: $\mathrm{NN}(x_t) = \arg\min_{\mu_i} ||x_t - \mu_i||$
- ②③ compute: $v_i = \sum_{x_t : \mathrm{NN}(x_t) = \mu_i} x_t - \mu_i$
- concatenate v_i's + ℓ_2 normalize



Jégou, Douze, Schmid and Pérez, "Aggregating local descriptors into a compact image representation", CVPR'10.

A first example: the VLAD

A graphical representation of $v_i = \sum_{x_t : \mathrm{NN}(x_t) = \mu_i} x_t - \mu_i$



Jégou, Douze, Schmid and Pérez, “Aggregating local descriptors into a compact image representation”, CVPR’10.

A first example: the VLAD

But in which sense is the VLAD optimal?

Could we add other (higher-order) statistics?

Outline

A first simple example: the VLAD

The Fisher Vector

Other higher-order representations

Example results

The Fisher vector

Score function

Given a likelihood function u_λ with parameters λ, the **score function** of a given sample X is given by:

$$\boxed{G_\lambda^X = \nabla_\lambda \log u_\lambda(X)}$$

→ Fixed-length vector whose **dimensionality depends only on # parameters**.

Intuition: direction in which the parameters λ of the model should we modified to better fit the data.

The Fisher vector
Fisher information matrix

Fisher information matrix (FIM) or negative Hessian:

$$F_\lambda = E_{x \sim u_\lambda} \left[\nabla_\lambda \log u_\lambda(x) \nabla_\lambda \log u_\lambda(x)' \right]$$

Measure similarity between using the **Fisher Kernel (FK)**:

$$\boxed{K(X,Y) = G_\lambda^{X\prime} F_\lambda^{-1} G_\lambda^Y}$$

Jaakkola and Haussler, "Exploiting generative models in discriminative classifiers", NIPS'98.

$\longrightarrow$ can be interpreted as a score whitening

As the FIM, is PSD, it can be decomposed as: $F_\lambda^{-1} = L'_\lambda L_\lambda$

and the FK can be rewritten as a dot product between **Fisher Vectors** (FV):

$$\boxed{\mathcal{G}_\lambda^X \quad = \quad L_\lambda G_\lambda^X}$$

The Fisher vector
Application to images

$X = \{x_t, t = 1 \dots T\}$ is the set of T i.i.d. D-dim local descriptors (e.g. SIFT) extracted from an image:

$$\boxed{G_\lambda^X = \frac{1}{T}\sum_{t=1}^{T} \nabla_\lambda \log u_\lambda(x_t)}$$

→ **average pooling** is a direct consequence of independence assumption

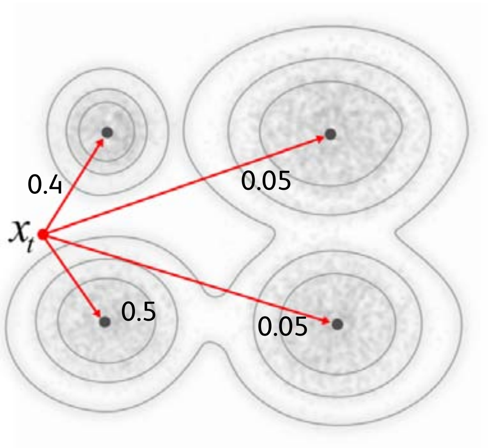
$u_\lambda(x) = \sum_{i=1}^{K} w_i u_i(x)$ is a Gaussian Mixture Model (GMM) with parameters $\lambda = \{w_i, \mu_i, \Sigma_i, i = 1 \dots N\}$ trained on a large set of local descriptors → a probabilistic **visual vocabulary**

Perronnin and Dance, "Fisher kernels on visual categories for image categorization", CVPR'07.

The Fisher vector

Relationship with the BOV

FV formulas:



Perronnin and Dance, “Fisher kernels on visual categories for image categorization”, CVPR’07.

The Fisher vector

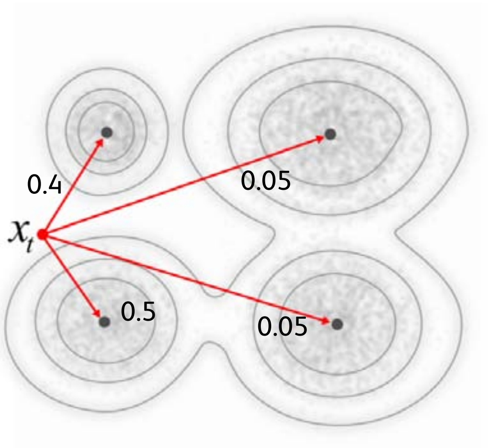
Relationship with the BOV

FV formulas:

- gradient wrt to w

$$\approx \boxed{\frac{1}{T}\sum_{t=1}^{T}\gamma_t(i)}$$

$\rightarrow$ **soft BOV**



$\gamma_t(i)$ = soft-assignment of patch t to Gaussian i

Perronnin and Dance, "Fisher kernels on visual categories for image categorization", CVPR'07.

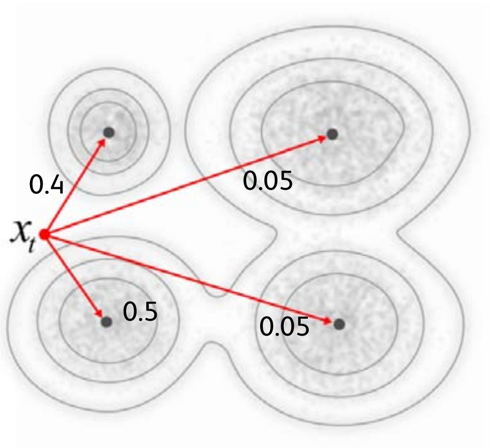
The Fisher vector
Relationship with the BOV

FV formulas:

- gradient wrt to w

≈ $\boxed{\frac{1}{T}\sum_{t=1}^{T}\gamma_t(i)}$

→ **soft BOV**



- gradient wrt to μ and σ

$$\boxed{\begin{aligned}\mathcal{G}_{\mu,i}^X &= \frac{1}{T\sqrt{w_i}}\sum_{t=1}^{T}\gamma_t(i)\left(\frac{x_t-\mu_i}{\sigma_i}\right)\\ \mathcal{G}_{\sigma,i}^X &= \frac{1}{T\sqrt{2w_i}}\sum_{t=1}^{T}\gamma_t(i)\left[\frac{(x_t-\mu_i)^2}{\sigma_i^2}-1\right]\end{aligned}}$$

$\gamma_t(i)$ = soft-assignment of patch t to Gaussian i

→ compared to BOV, include **higher-order statistics** (up to order 2)

Let us denote: D = feature dim, N = # Gaussians

- BOV = N-dim
- FV = 2DN-dim

Perronnin and Dance, "Fisher kernels on visual categories for image categorization", CVPR'07.

The Fisher vector

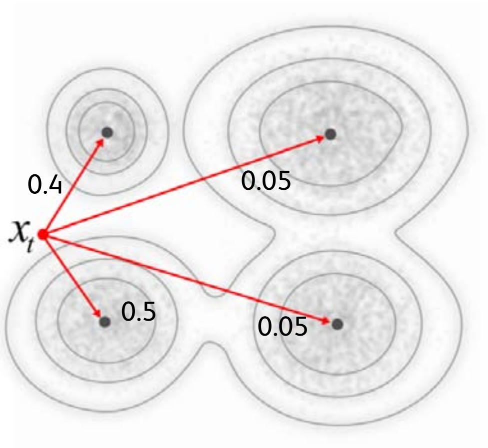
Relationship with the BOV

FV formulas:

- gradient wrt to w

$\approx \boxed{\frac{1}{T}\sum_{t=1}^{T}\gamma_t(i)}$

→ **soft BOV**



- gradient wrt to μ and σ

$$\mathcal{G}^X_{\mu,i} = \frac{1}{T\sqrt{w_i}}\sum_{t=1}^{T}\gamma_t(i)\left(\frac{x_t-\mu_i}{\sigma_i}\right)$$

$$\mathcal{G}^X_{\sigma,i} = \frac{1}{T\sqrt{2w_i}}\sum_{t=1}^{T}\gamma_t(i)\left[\frac{(x_t-\mu_i)^2}{\sigma_i^2}-1\right]$$

$\gamma_t(i)$ = soft-assignment of patch t to Gaussian i

→ compared to BOV, include **higher-order statistics** (up to order 2)

→ FV **much higher-dim** than BOV for a **given visual vocabulary size**

→ FV **much faster to compute** than BOV for a **given feature dim**

Perronnin and Dance, "Fisher kernels on visual categories for image categorization", CVPR'07.

The Fisher vector

Dimensionality reduction on local descriptors

Perform PCA on local descriptors:

→ uncorrelated features are more consistent with diagonal assumption of covariance matrices in GMM

→ FK performs whitening and enhances low-energy (possibly noisy) dimensions

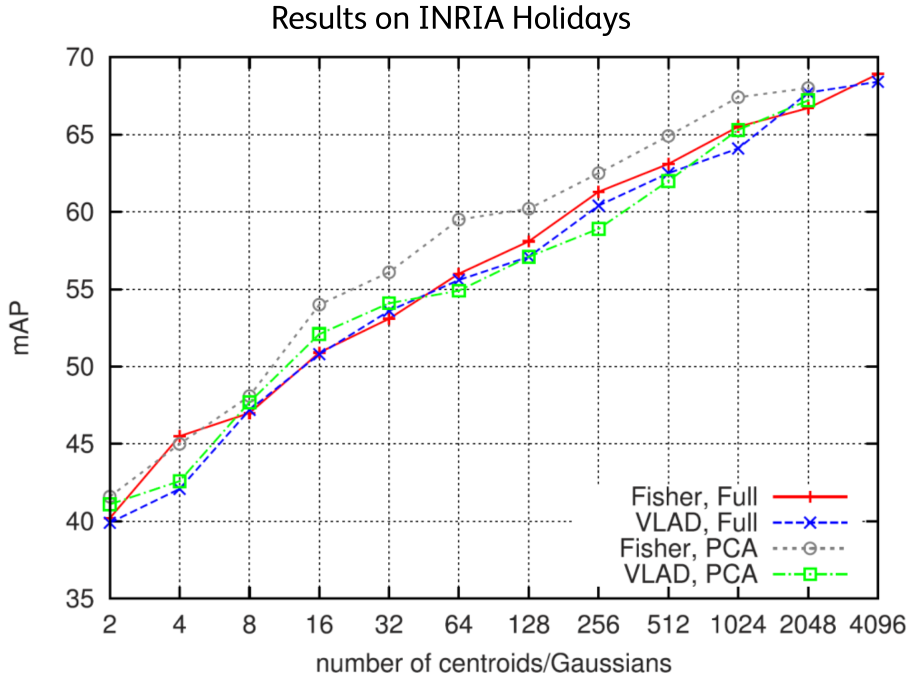
The Fisher vector

Dimensionality reduction on local descriptors

Perform PCA on local descriptors:

→ uncorrelated features are more consistent with diagonal assumption of covariance matrices in GMM

→ FK performs whitening and enhances low-energy (possibly noisy) dimensions



Jégou, Perronnin, Douze, Sánchez, Pérez and Schmid, “Aggregating local descriptors into compact codes”, TPAMI’11.

The Fisher vector
Normalization: TF-IDF effect

Assuming that the x_t's are iid drawn from a distribution p, we have:

$$G_\lambda^X = \frac{1}{T}\sum_{t=1}^{T} \nabla_\lambda \log u_\lambda(x_t) \approx \nabla_\lambda E_{x\sim p} \log u_\lambda(x) = \nabla_\lambda \int_x p(x) \log u_\lambda(x) dx.$$

If we assume that p is a mixture of image-dependent and image-independent information:

$$p(x) = \omega q(x) + (1-\omega) u_\lambda(x)$$

Then we have:

$$G_\lambda^X \approx \omega \nabla_\lambda \int_x q(x) \log u_\lambda(x) dx + (1-\omega) \underbrace{\nabla_\lambda \int_x u_\lambda(x) \log u_\lambda(x) dx}_{\approx 0 \text{ (MLE)}}$$

→The FV depends only (approximately) on image-specific content (**TF-IDF**)

→ ℓ_2 normalization removes dependence on ω

Perronnin, Sánchez and Mensink, "Improving the Fisher kernel for large-scale image classification", ECCV'10.

The Fisher vector

Normalization: variance stabilization

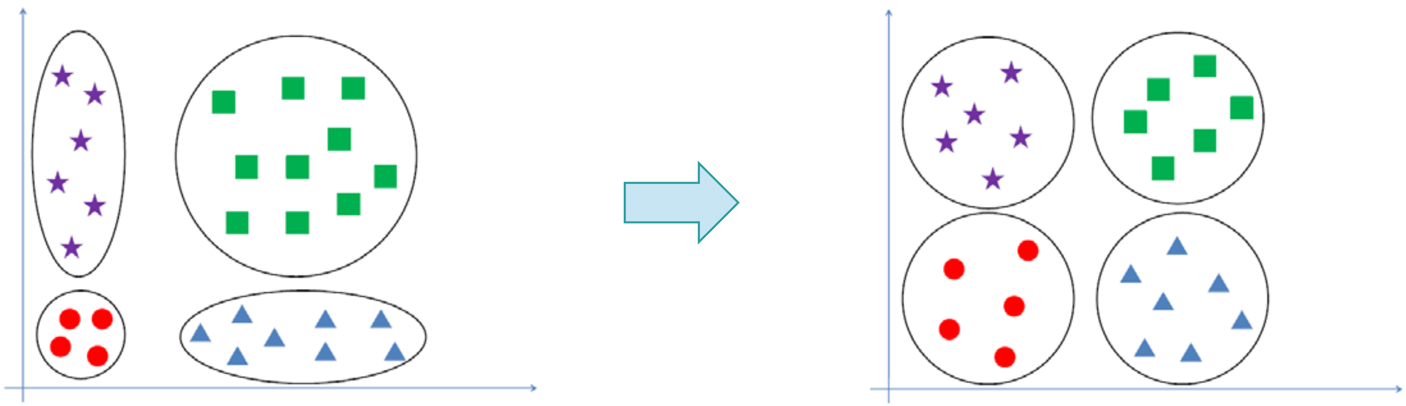
FVs can be (approximately) viewed as emissions of a compound Poisson: a sum of N iid random variables with N~Poisson.
☹ variance depends on mean

Jégou, Perronnin, Douze, Sánchez, Pérez and Schmid, "Aggregating local descriptors into compact codes", TPAMI'11.

→ **Variance stabilizing transforms** of the form:

$$\boxed{f(z) = \text{sign}(z)|z|^{\alpha} \text{ with } 0 \leq \alpha \leq 1}$$ (with α=0.5 by default)

can be used on the FV (or the VLAD).



Perronnin, Sánchez and Mensink, "Improving the Fisher kernel for large-scale image classification", ECCV'10.

The Fisher vector

Normalization: variance stabilization

FVs can be (approximately) viewed as emissions of a compound Poisson: a sum of N iid random variables with N~Poisson.
☹ variance depends on mean

Jégou, Perronnin, Douze, Sánchez, Pérez and Schmid, "Aggregating local descriptors into compact codes", TPAMI'11.

→ **Variance stabilizing transforms** of the form:

$$\boxed{f(z) = \mathrm{sign}(z)|z|^{\alpha} \text{ with } 0 \leq \alpha \leq 1}$$ (with α=0.5 by default)

can be used on the FV (or the VLAD).

→ Reduce impact of bursty visual elements

Jégou, Douze, Schmid, "On the burstiness of visual elements", ICCV'09.

Outline

A first example: the VLAD

The Fisher Vector

Other higher-order representations

Example results

Other higher-order representations

Revisiting the VLAD

But in which sense is the VLAD optimal?

Could we add other (higher-order) statistics?

Jégou, Douze, Schmid and Pérez, “Aggregating local descriptors into a compact image representation”, CVPR’10.

Other higher-order representations
Revisiting the VLAD

But in which sense is the VLAD optimal?

→ The VLAD can be viewed as a non-probabilistic version of the FV:

- gradient with respect to mean only
- replace GMM clustering by k-means

$$\mathcal{G}^X_{\mu,i} = \frac{1}{T\sqrt{w_i}} \sum_{t=1}^{T} \gamma_t(i) \left(\frac{x_t - \mu_i}{\sigma_i} \right) \qquad \rightarrow \qquad v_i = \sum_{x_t : \mathrm{NN}(x_t) = \mu_i} x_t - \mu_i$$

Could we add other (higher-order) statistics?

→ extension of the VLAD to include 2nd order statistics: VLAT

Picard and Gosselin, "Improving image similarity with vectors of locally aggregated tensors", ICIP '11.

Other higher-order representations
Super-Vector (SV) coding

$f : \mathbb{R}^D \rightarrow \mathbb{R}$ is Lipschitz smooth if $\forall (x, y) \in \mathbb{R}^D \times \mathbb{R}^D$:

$$|f(x) - f(y) - \nabla f(y)'(x - y)| \leq \frac{\beta}{2}||x - y||^2$$

Given a codebook $\{\mu_i, i = 1 \dots N\}$ and a patch x_t we have:

$$f(x_t) \approx f(\mu_i) + \nabla f(\mu_i)'(x_t - \mu_i) = w' \varphi_{SV}(x_t)$$

with $\varphi_{SV}(x_t) = \left[0, \dots, 0, \overbrace{s, (x_t - \mu_i)}^{(D+1) \text{ non-zero dim}}, 0, \dots, 0 \right]$

and $w = \left[0, \dots, 0, \frac{f(\mu_i)}{s}, \nabla f(\mu_i), 0, \dots, 0 \right]$ (to be learned)

Zhou, Yu, Zhang and Huang, "Image classification using super-vector coding of local image descriptors", ECCV'10.

Other higher-order representations

Super-Vector (SV) coding

$f : \mathbb{R}^D \rightarrow \mathbb{R}$ is Lipschitz smooth if $\forall (x, y) \in \mathbb{R}^D \times \mathbb{R}^D$:

$$|f(x) - f(y) - \nabla f(y)'(x - y)| \leq \frac{\beta}{2} ||x - y||^2$$

Given a codebook $\{\mu_i, i = 1 \ldots N\}$ and a patch x_t we have:

$$f(x_t) \approx f(\mu_i) + \nabla f(\mu_i)'(x_t - \mu_i) = w' \varphi_{SV}(x_t)$$

with $\varphi_{SV}(x_t) = \left[0, \ldots, 0, \overbrace{s, (x_t - \mu_i)}^{(D+1) \text{ non-zero dim}}, 0, \ldots, 0 \right]$

Average pooling → **SV ≈ BOV + VLAD**

Bound in Lipschitz smooth inequality provides argument for k-means.

Zhou, Yu, Zhang and Huang, “Image classification using super-vector coding of local image descriptors”, ECCV’10.
See also: Ladický and Torr, “Locally linear support vector machines”, ICML’11.

Outline

A first example: the VLAD

The Fisher Vector

Other higher-order representations

Example results

Examples
Retrieval

Example on Holidays:

From: Jégou, Perronnin, Douze, Sánchez, Pérez and Schmid, "Aggregating local descriptors into compact codes", TPAMI'11.

Descriptor	K	D	Holidays (mAP)					
			$D' = D$	$\rightarrow D'$=2048	$\rightarrow D'$=512	$\rightarrow D'$=128	$\rightarrow D'$=64	$\rightarrow D'$=32
BOW	1 000	1 000	40.1		43.5	44.4	43.4	40.8
	20 000	20 000	43.7	41.8	44.9	45.2	44.4	41.8
Fisher (μ)	16	1 024	54.0		54.6	52.3	49.9	46.6
	64	4 096	59.5	60.7	61.0	56.5	52.0	48.0
	256	16 384	62.5	62.6	57.0	53.8	50.6	48.6
VLAD	16	1 024	52.0		52.7	52.6	50.5	47.7
	64	4 096	55.6	57.6	59.8	55.7	52.3	48.4
	256	16 384	58.7	62.1	56.7	54.2	51.3	48.1

Examples
Retrieval

Example on Holidays:

From: Jégou, Perronnin, Douze, Sánchez, Pérez and Schmid, “Aggregating local descriptors into compact codes”, TPAMI’11.

Descriptor	K	D	Holidays (mAP)					
			$D' = D$	$\rightarrow D'$=2048	$\rightarrow D'$=512	$\rightarrow D'$=128	$\rightarrow D'$=64	$\rightarrow D'$=32
BOW	1 000	1 000	40.1		43.5	44.4	43.4	40.8
	20 000	20 000	43.7	41.8	44.9	45.2	44.4	41.8
Fisher (μ)	16	1 024	54.0		54.6	52.3	49.9	46.6
	64	4 096	59.5	60.7	61.0	56.5	52.0	48.0
	256	16 384	62.5	62.6	57.0	53.8	50.6	48.6
VLAD	16	1 024	52.0		52.7	52.6	50.5	47.7
	64	4 096	55.6	57.6	59.8	55.7	52.3	48.4
	256	16 384	58.7	62.1	56.7	54.2	51.3	48.1

$\rightarrow$ second order statistics are not essential for retrieval

Examples
Retrieval

Example on Holidays:

From: Jégou, Perronnin, Douze, Sánchez, Pérez and Schmid, “Aggregating local descriptors into compact codes”, TPAMI’11.

Descriptor	K	D	Holidays (mAP)					
			$D'=D$	$\rightarrow D'$=2048	$\rightarrow D'$=512	$\rightarrow D'$=128	$\rightarrow D'$=64	$\rightarrow D'$=32
BOW	1 000	$\boxed{1\,000}$	$\boxed{40.1}$		43.5	44.4	43.4	40.8
	20 000	20 000	43.7	41.8	44.9	45.2	44.4	41.8
Fisher (μ)	16	$\boxed{1\,024}$	$\boxed{54.0}$		54.6	52.3	49.9	46.6
	64	4 096	59.5	60.7	61.0	56.5	52.0	48.0
	256	16 384	62.5	62.6	57.0	53.8	50.6	48.6
VLAD	16	$\boxed{1\,024}$	$\boxed{52.0}$		52.7	52.6	50.5	47.7
	64	4 096	55.6	57.6	59.8	55.7	52.3	48.4
	256	16 384	58.7	62.1	56.7	54.2	51.3	48.1

$\rightarrow$ second order statistics are not essential for retrieval

$\rightarrow$ even for the same feature dim, the FV/VLAD can beat the BOV

Examples
Retrieval

Example on Holidays:

From: Jégou, Perronnin, Douze, Sánchez, Pérez and Schmid, "Aggregating local descriptors into compact codes", TPAMI'11.

Descriptor	K	D	Holidays (mAP)					
			$D' = D$	$\rightarrow D'$=2048	$\rightarrow D'$=512	$\rightarrow D'$=128	$\rightarrow D'$=64	$\rightarrow D'$=32
BOW	1 000	1 000	40.1		43.5	44.4	43.4	40.8
	20 000	20 000	43.7	41.8	44.9	45.2	44.4	41.8
Fisher (μ)	16	1 024	54.0		54.6	52.3	49.9	46.6
	64	4 096	59.5	60.7	61.0	56.5	52.0	48.0
	256	16 384	62.5	62.6	57.0	53.8	50.6	48.6
VLAD	16	1 024	52.0		52.7	52.6	50.5	47.7
	64	4 096	55.6	57.6	59.8	55.7	52.3	48.4
	256	16 384	58.7	62.1	56.7	54.2	51.3	48.1

$\rightarrow$ second order statistics are not essential for retrieval

$\rightarrow$ even for the same feature dim, the FV/VLAD can beat the BOV

$\rightarrow$ soft assignment + whitening of FV helps when number of Gaussians $\uparrow$

Examples
Retrieval

Example on Holidays:

From: Jégou, Perronnin, Douze, Sánchez, Pérez and Schmid, “Aggregating local descriptors into compact codes”, TPAMI’11.

Descriptor	K	D	Holidays (mAP)					
			$D' = D$	$\rightarrow D'$=2048	$\rightarrow D'$=512	$\rightarrow D'$=128	$\rightarrow D'$=64	$\rightarrow D'$=32
BOW	1 000	1 000	40.1		43.5	44.4	43.4	40.8
	20 000	20 000	43.7	41.8	44.9	45.2	44.4	41.8
Fisher (μ)	16	1 024	54.0		54.6	52.3	49.9	46.6
	64	4 096	59.5	60.7	61.0	56.5	52.0	48.0
	256	16 384	62.5	62.6	57.0	53.8	50.6	48.6
VLAD	16	1 024	52.0		52.7	52.6	50.5	47.7
	64	4 096	55.6	57.6	59.8	55.7	52.3	48.4
	256	16 384	58.7	62.1	56.7	54.2	51.3	48.1

→ second order statistics are not essential for retrieval

→ even for the same feature dim, the FV/VLAD can beat the BOV

→ soft assignment + whitening of FV helps when number of Gaussians ↑

→ after dim-reduction however, the FV and VLAD perform similarly

Examples
Classification

Example on PASCAL VOC 2007:

From: Chatfield, Lempitsky, Vedaldi and Zisserman, “The devil is in the details: an evaluation of recent feature encoding methods”, BMVC’11.

	Feature dim	mAP
VQ	25K	55.30
KCB	25K	56.26
LLC	25K	57.27
SV	41K	58.16
FV	132K	61.69

Examples
Classification

Example on PASCAL VOC 2007:

From: Chatfield, Lempitsky, Vedaldi and Zisserman, “The devil is in the details: an evaluation of recent feature encoding methods”, BMVC’11.

	Feature dim	mAP
VQ	25K	55.30
KCB	25K	56.26
LLC	25K	57.27
SV	41K	58.16
FV	132K	61.69

→ FV outperforms BOV-based techniques including:

- VQ: plain vanilla BOV
- KCB: BOV with soft assignment
- LLC: BOV with sparse coding

Examples
Classification

Example on PASCAL VOC 2007:

From: Chatfield, Lempitsky, Vedaldi and Zisserman, “The devil is in the details: an evaluation of recent feature encoding methods”, BMVC’11.

	Feature dim	mAP
VQ	25K	55.30
KCB	25K	56.26
LLC	25K	57.27
SV	41K	58.16
FV	132K	61.69

→ FV outperforms BOV-based techniques including:

- VQ: plain vanilla BOV
- KCB: BOV with soft assignment
- LLC: BOV with sparse coding

→ including 2nd order information is important for classification

Packages

The INRIA package:

http://lear.inrialpes.fr/src/inria_fisher/

The Oxford package (soon to be released):

http://www.robots.ox.ac.uk/~vgg/research/encoding_eval/

Questions?