Mixtures of Gaussians and Advanced Feature Encoding

Computer Vision

CS 143, Brown

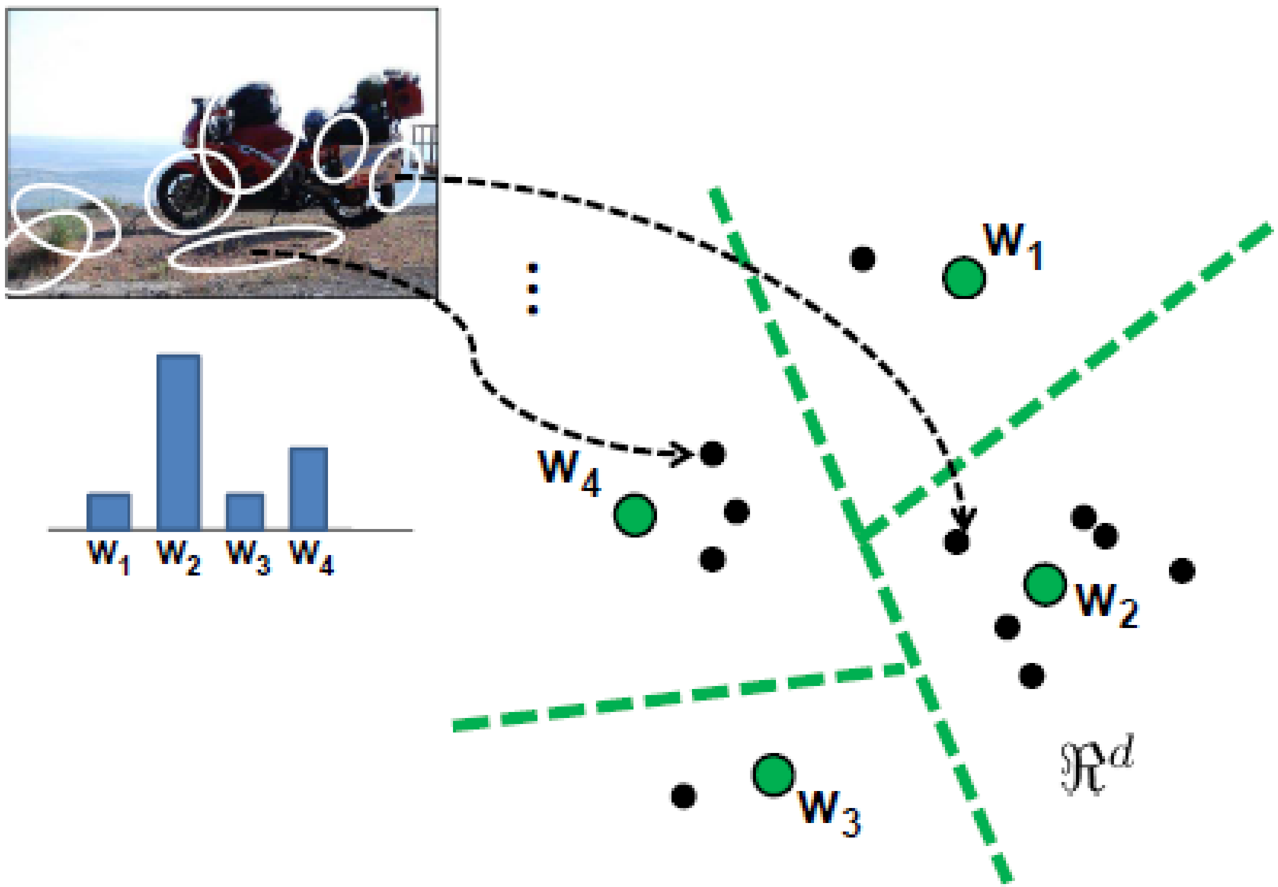
James Hays

Many slides from Derek Hoiem, Florent Perronnin, and Hervé Jégou

Why do good recognition systems go bad?

- E.g. Why isn't our Bag of Words classifier at 90% instead of 70%?
- Training Data
  - Huge issue, but not necessarily a variable you can manipulate.
- Learning method
  - Probably not such a big issue, unless you're learning the representation (e.g. deep learning).
- Representation
  - Are the local features themselves lossy? Guest lecture Nov 8th will address this.
  - What about feature quantization? That's VERY lossy.

Standard Kmeans Bag of Words



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

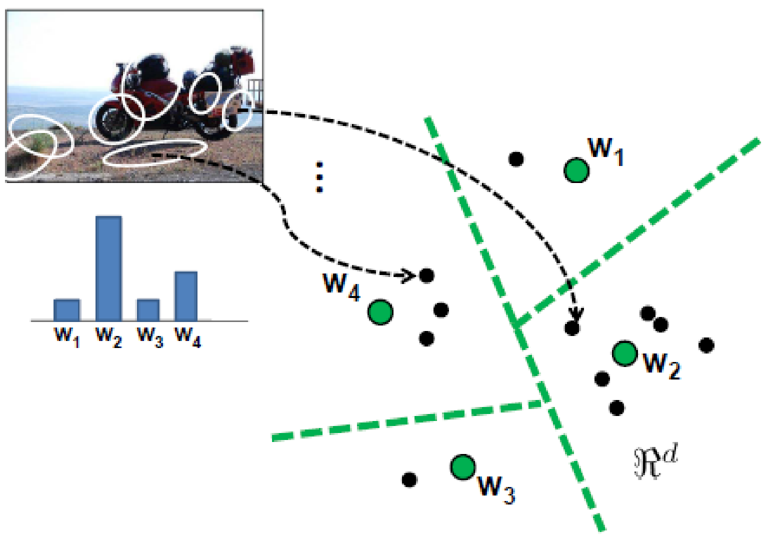
Today's Class

- More advanced quantization / encoding methods that represent the state-of-the-art in image classification and image retrieval.
  - Soft assignment (a.k.a. Kernel Codebook)
  - VLAD
  - Fisher Vector
- Mixtures of Gaussians

Motivation

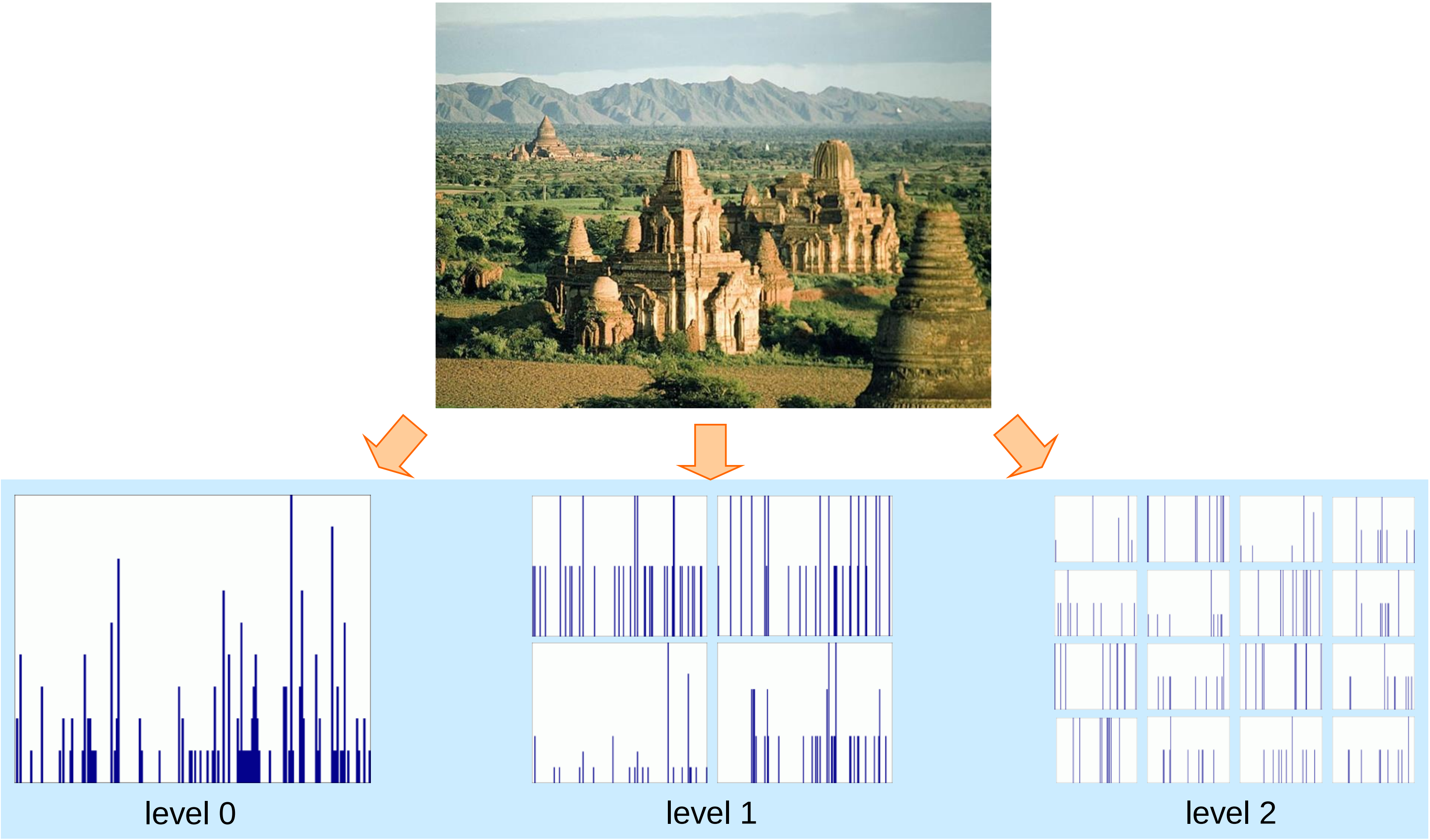
Bag of Visual Words is only about **counting** the number of local descriptors assigned to each Voronoi region

Why not including **other statistics**?



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

We already looked at the Spatial Pyramid



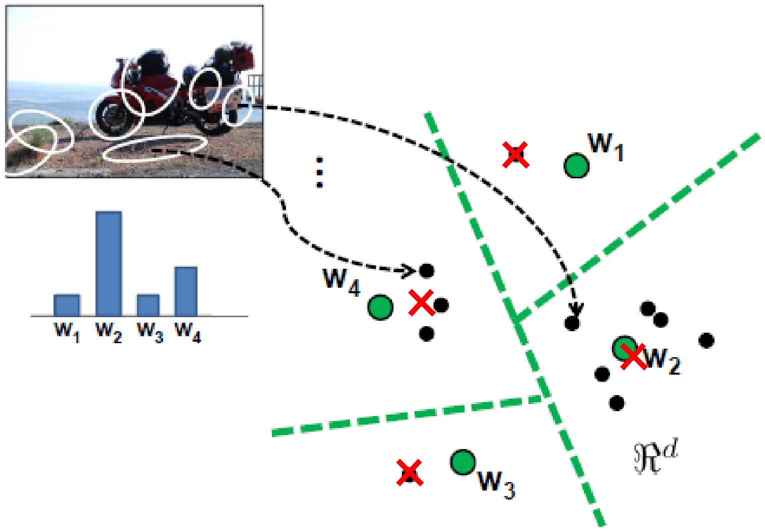
But today we’re not talking about ways to preserve *spatial* information.

Motivation

Bag of Visual Words 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 ×



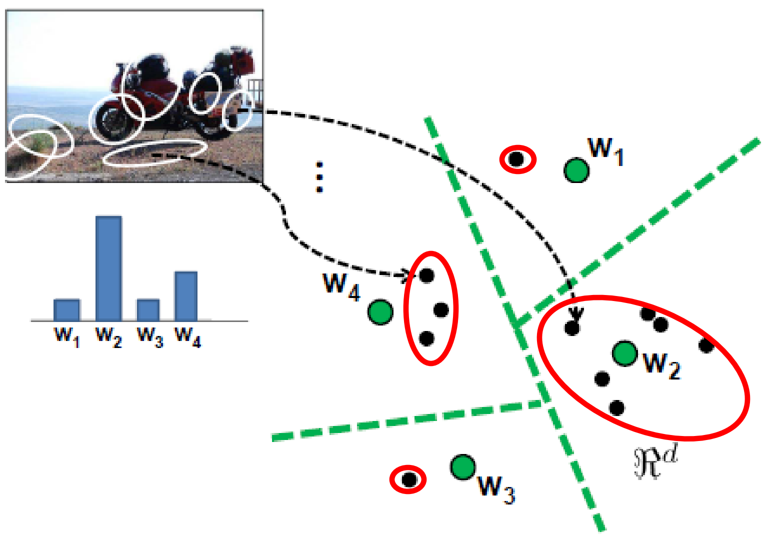
http://www.cs.utexas.edu/~grauman/courses/fall2009/papers/bag_of_visual_words.pdf

Motivation

Bag of Visual Words is only about **counting** the number of local descriptors assigned to each Voronoi region

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

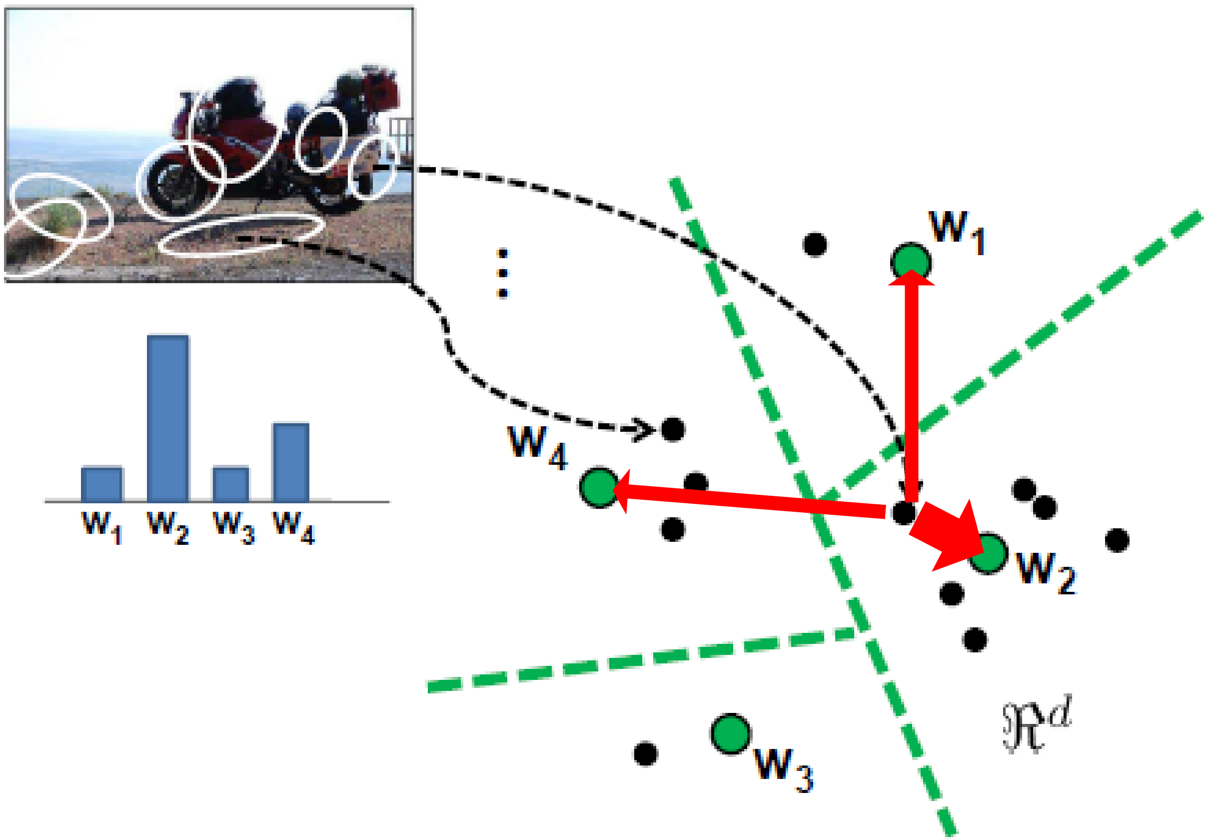
- mean of local descriptors
- (co)variance of local descriptors



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

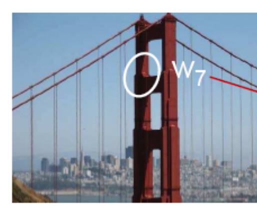
Simple case: Soft Assignment

- Called "Kernel codebook encoding" by Chatfield et al. 2011. Cast a weighted vote into the most similar clusters.



Simple case: Soft Assignment

- Called "Kernel codebook encoding" by Chatfield et al. 2011. Cast a weighted vote into the most similar clusters.
- This is fast and easy to implement (try it for Project 3!) but it does have some downsides for image retrieval – the inverted file index becomes less sparse.



New query image

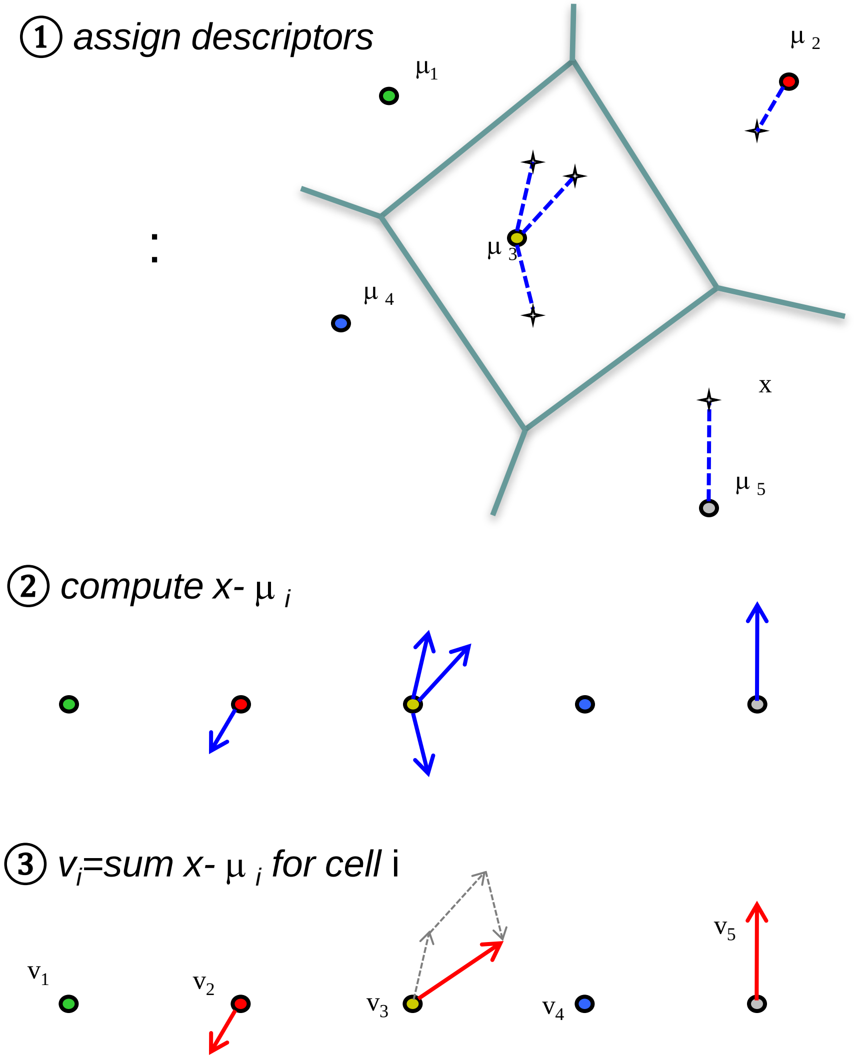
Word #	Image #
1	3
2 ...	
7	1, 2
8	3
9	
10 ...	
91	2
⋮	⋮

...

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.

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:

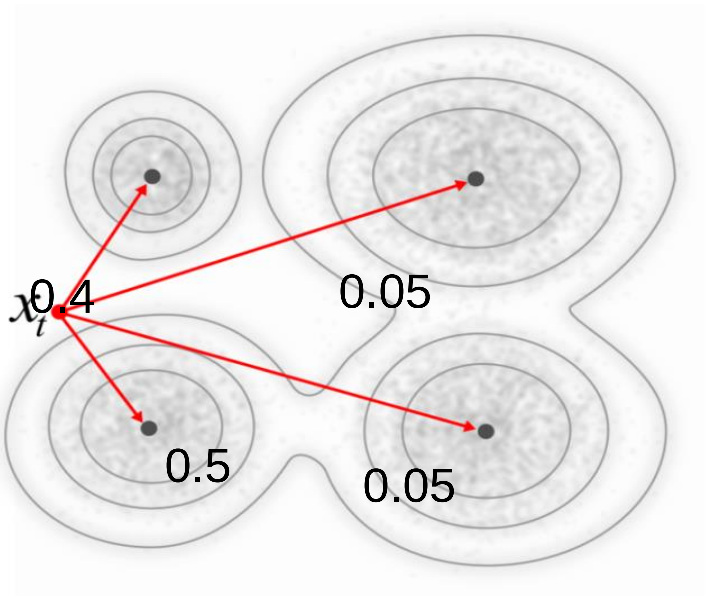
$$\boxed{G_\lambda^X = \nabla_\lambda \log u_\lambda(X)}$$

$\rightarrow$ 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.

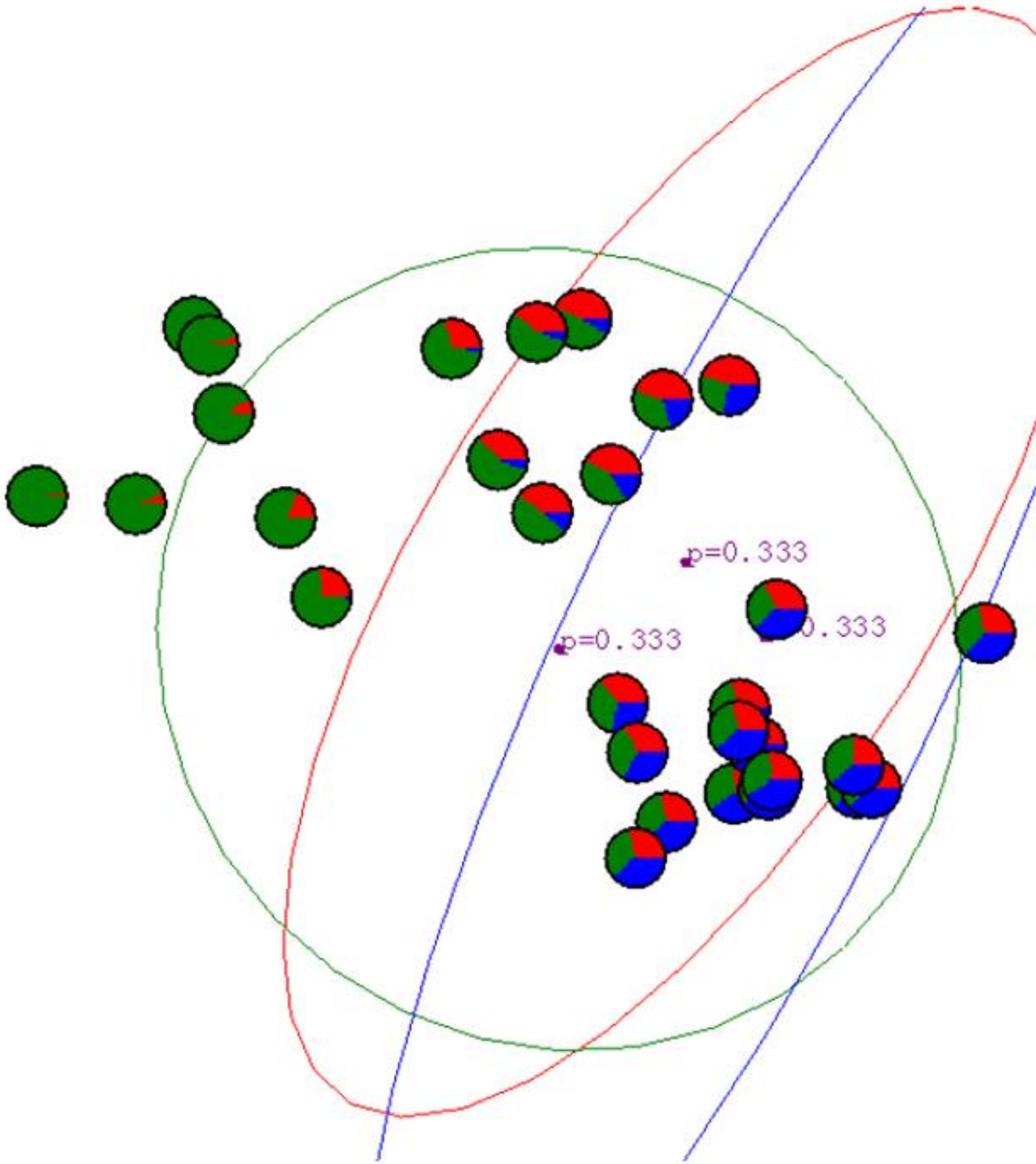
Aside: Mixture of Gaussians (GMM)

- For Fisher Vector image representations, u_λ is a GMM.
- GMM can be thought of as "soft" kmeans.



- Each component has a mean and a standard deviation along each direction (or full covariance)

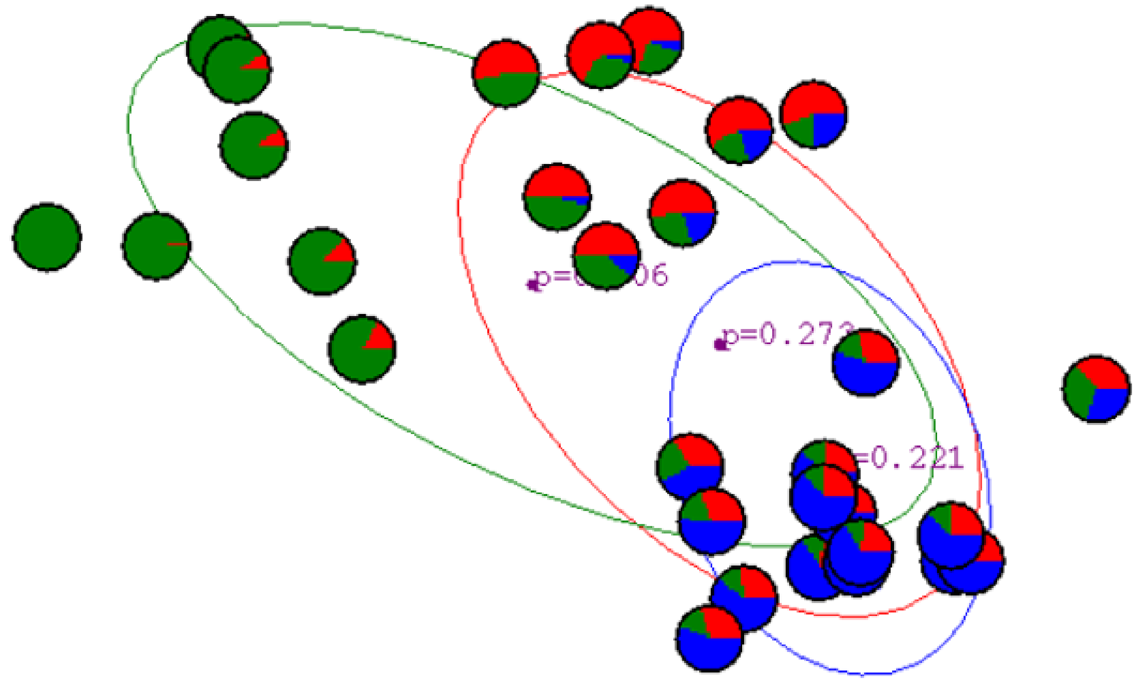
Gaussian Mixture Example: Start



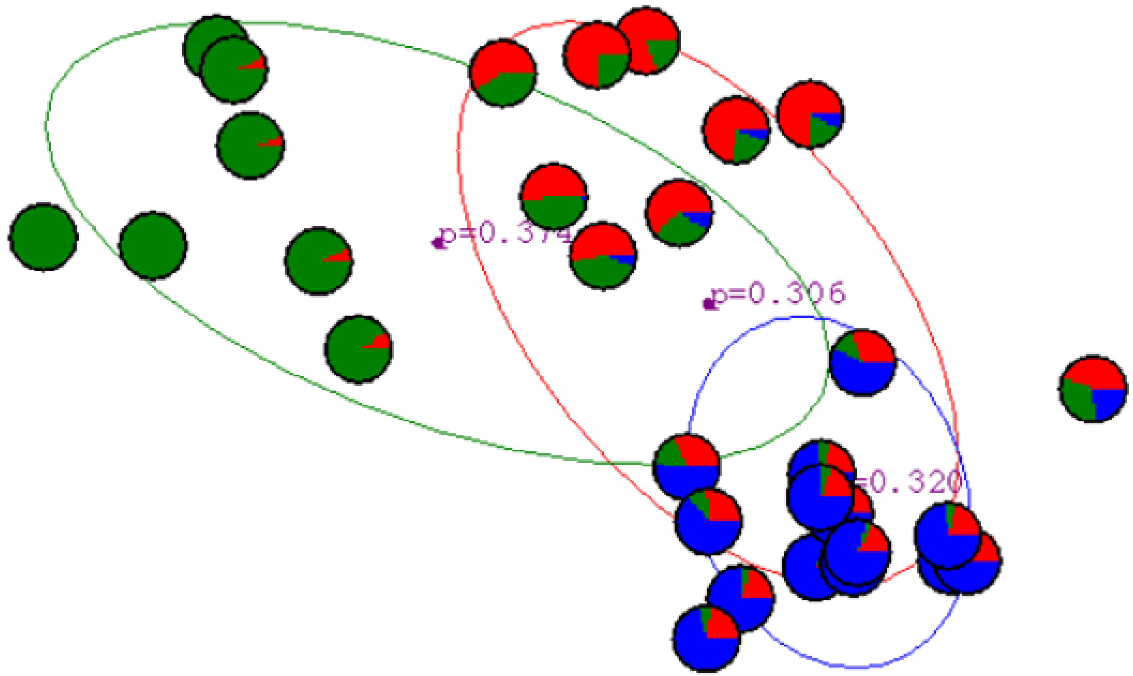
This looks like a soft version of kmeans!

Advance apologies: in Black and White this example will be incomprehensible

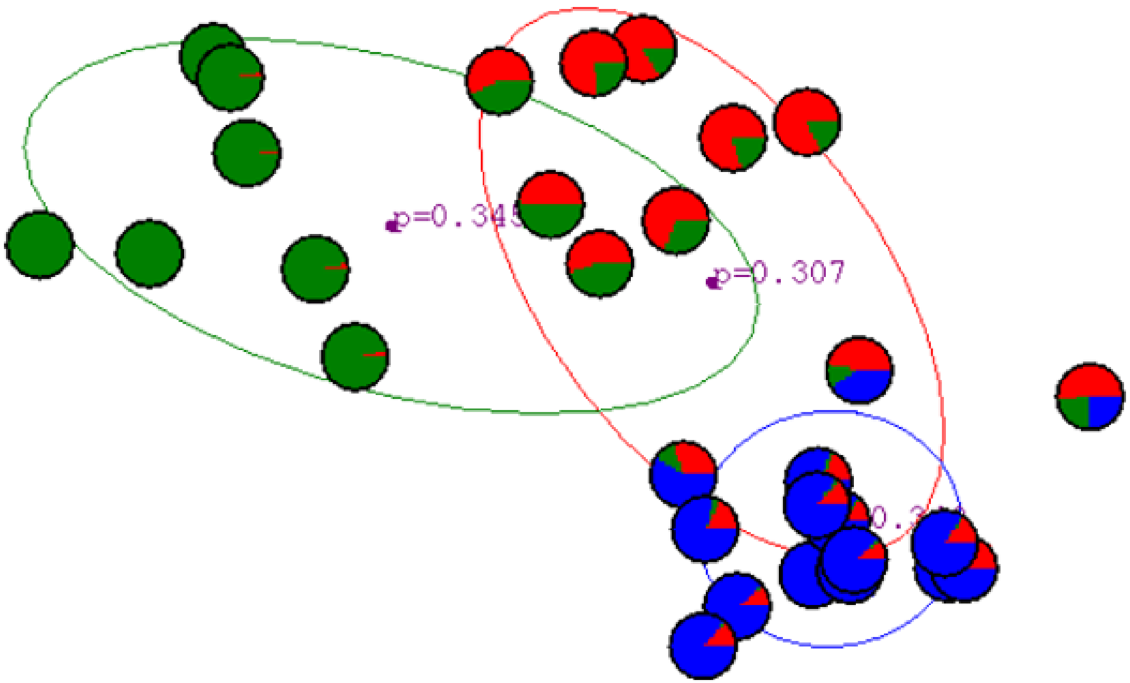
After first iteration



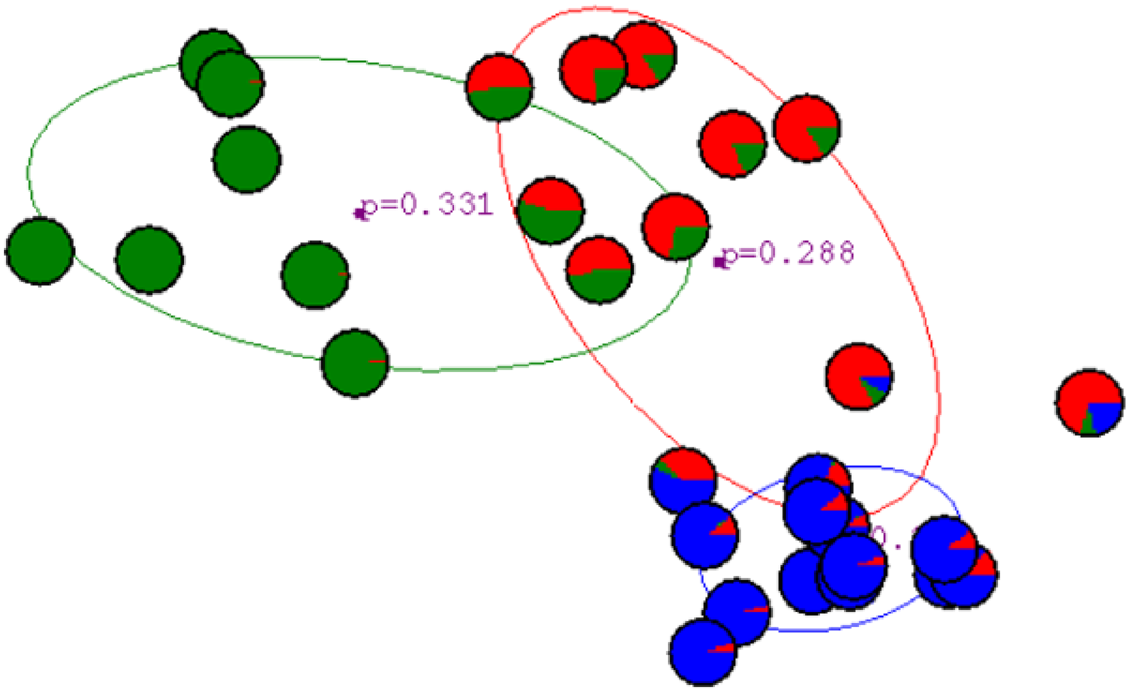
After 2nd iteration



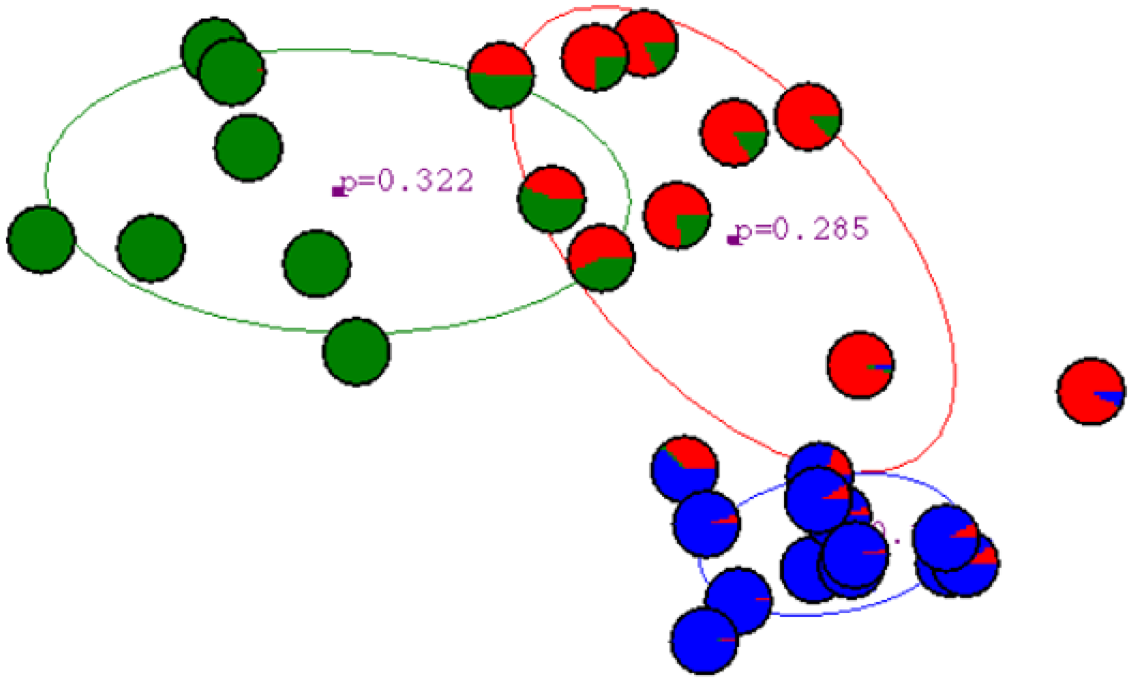
After 3rd iteration



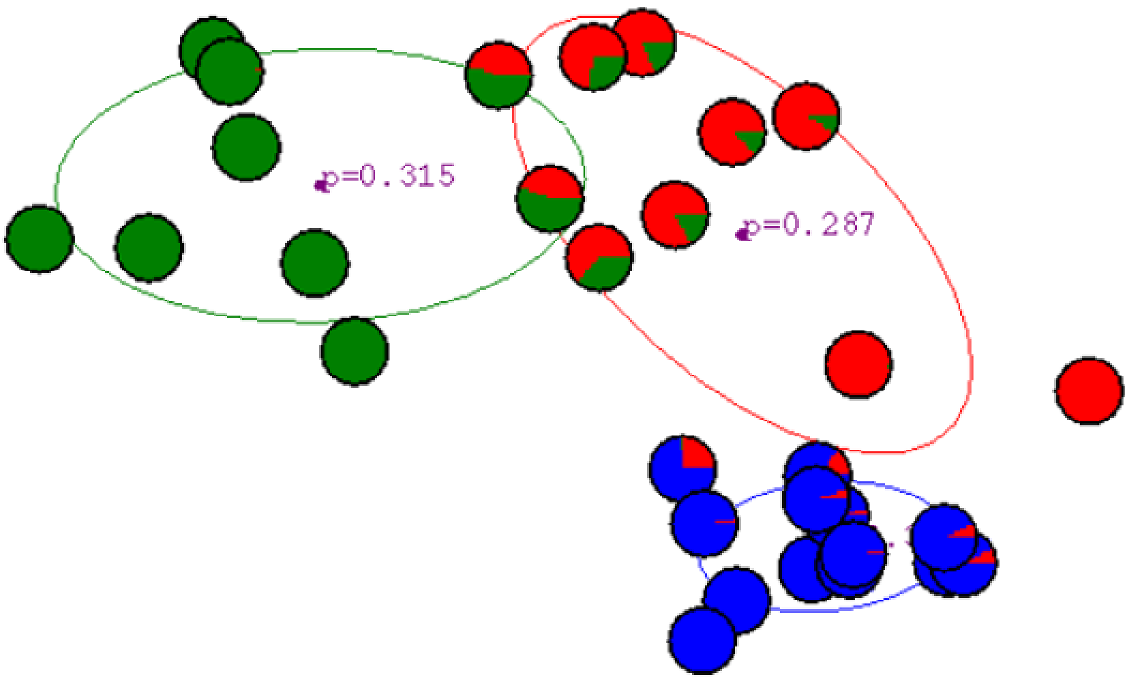
After 4th iteration



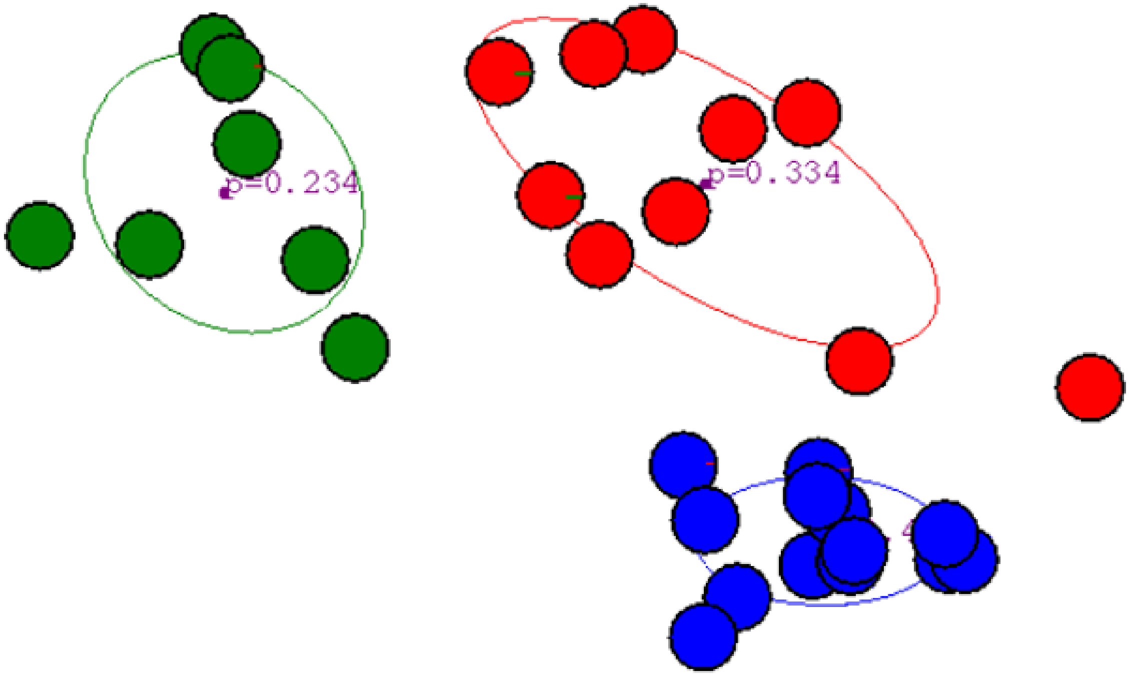
After 5th iteration



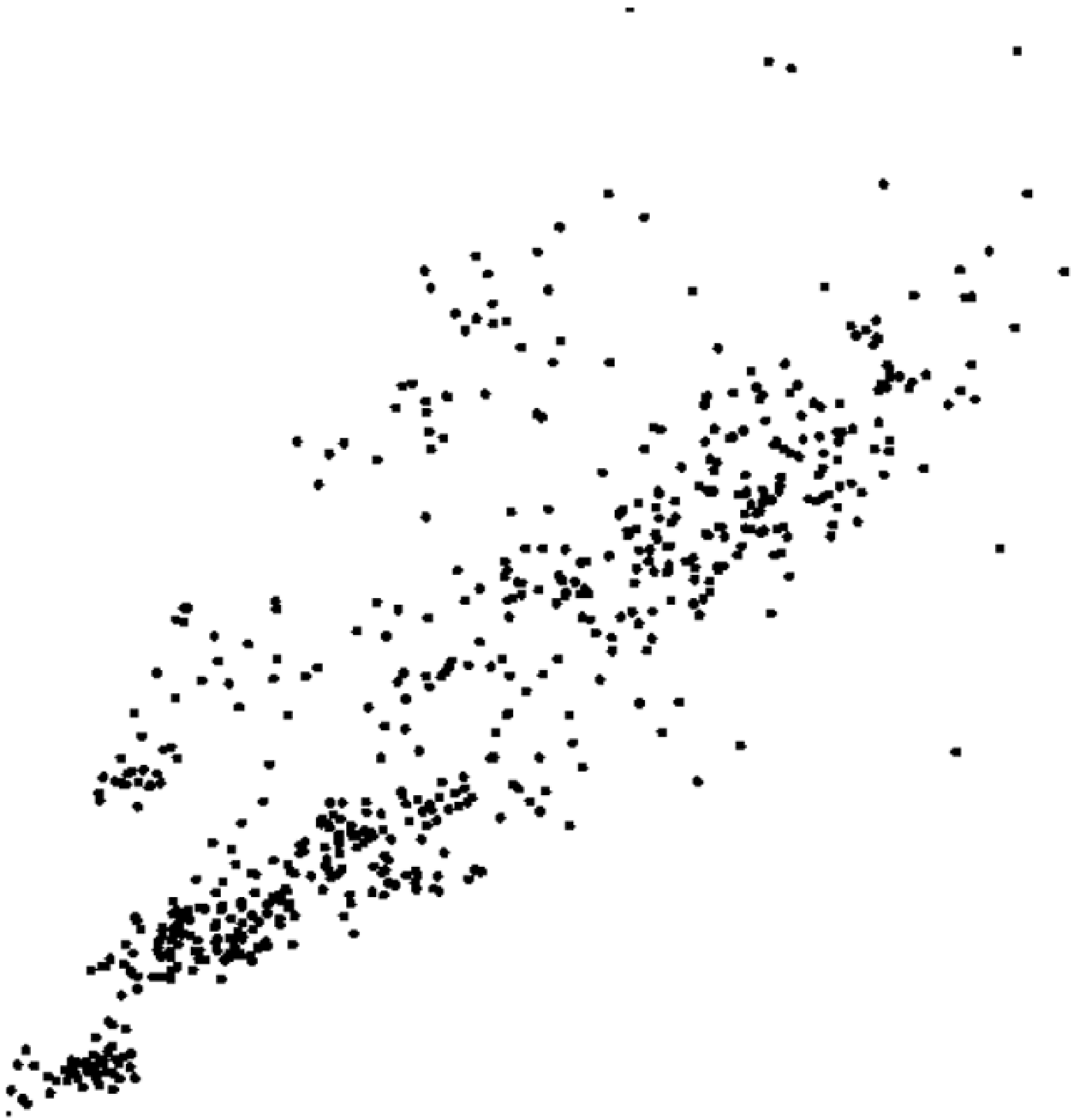
After 6th iteration



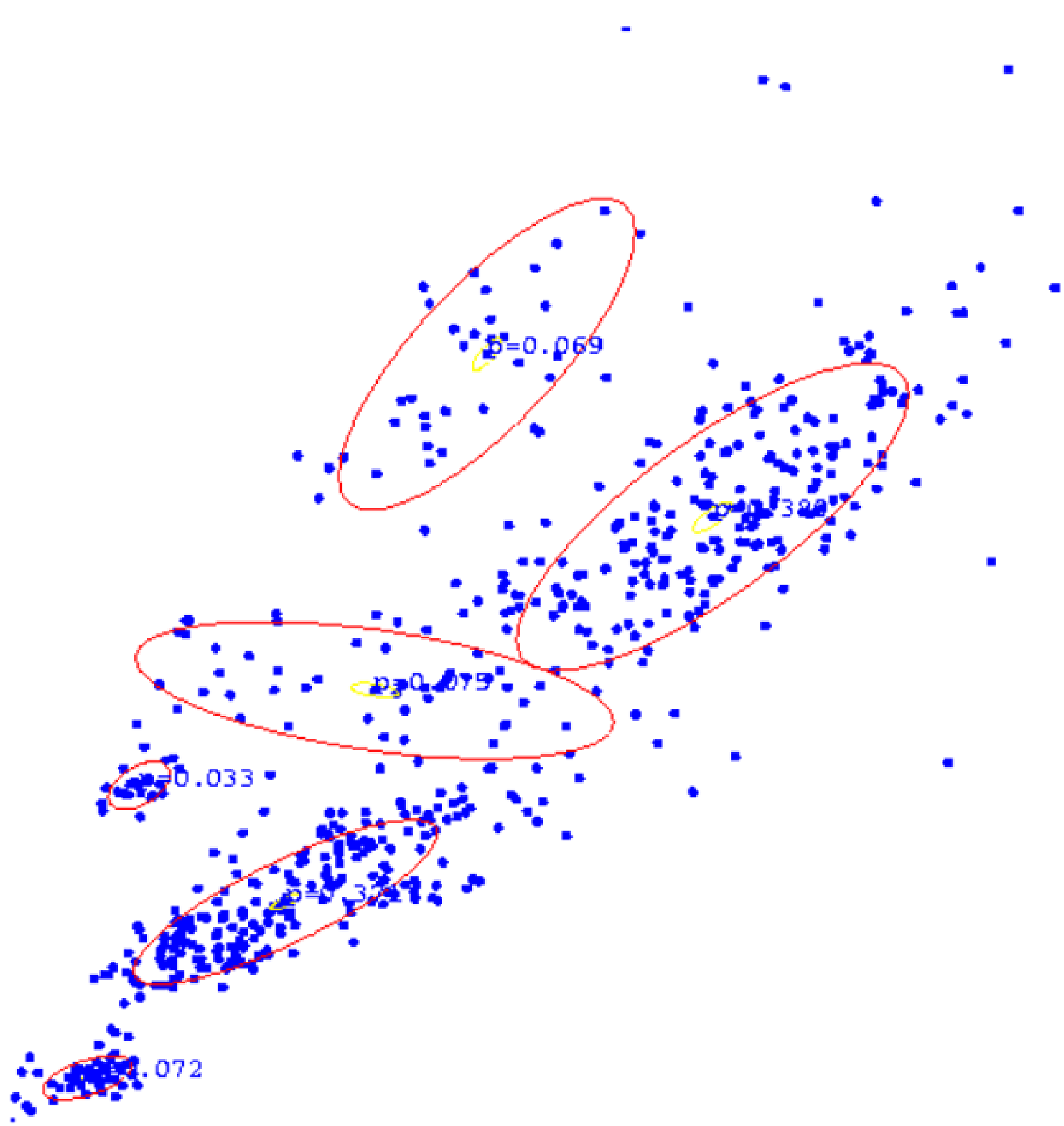
After 20th iteration



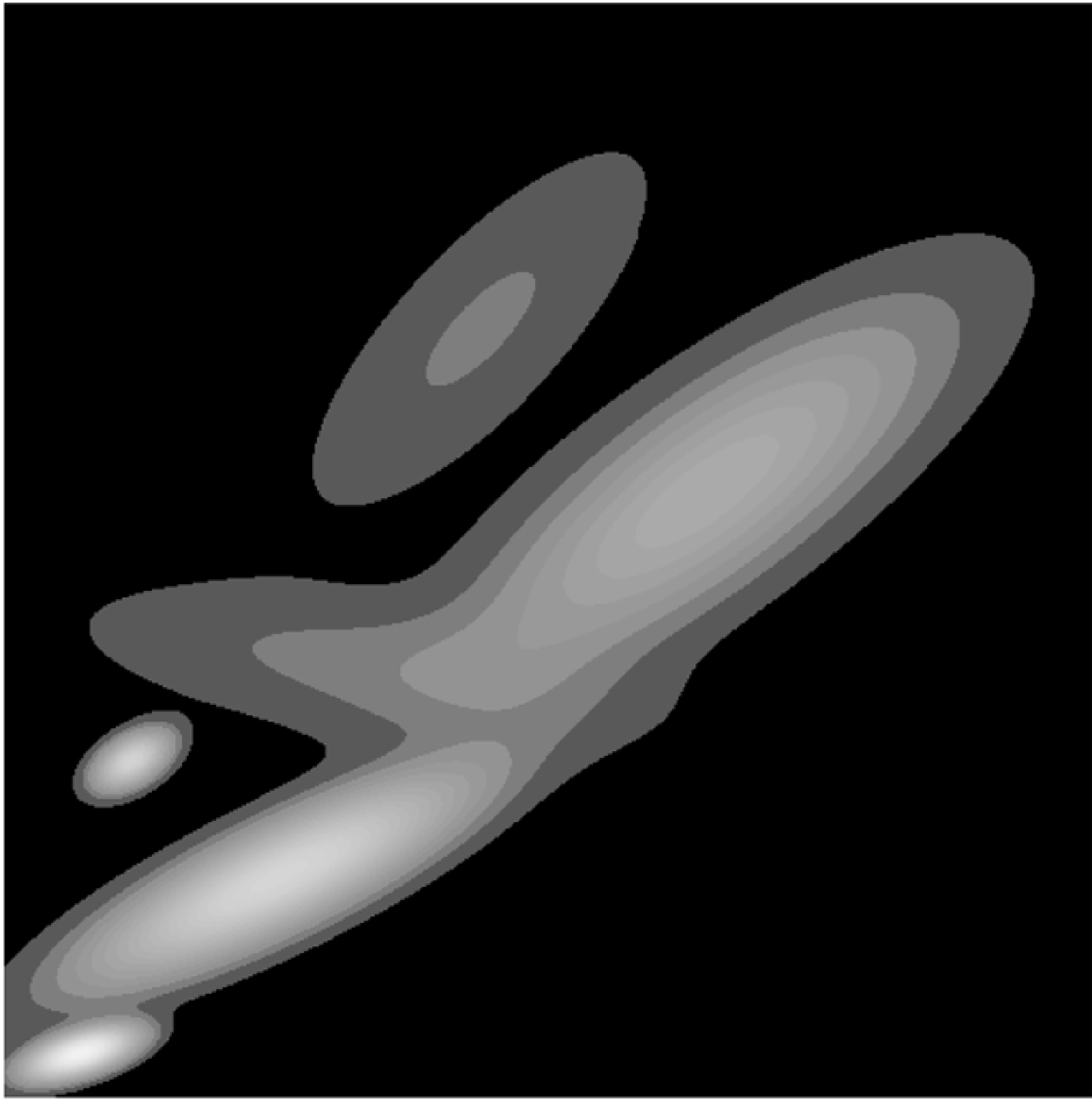
Some Bio Assay data



GMM clustering of the assay data



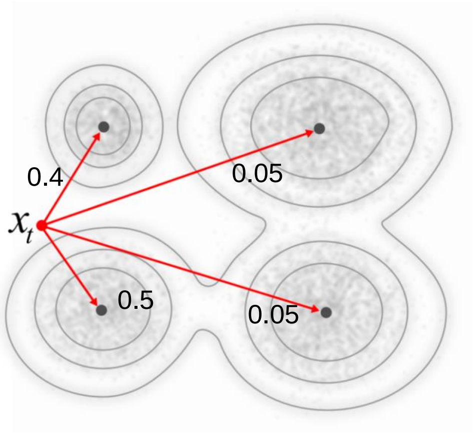
Resulting Density Estimator



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

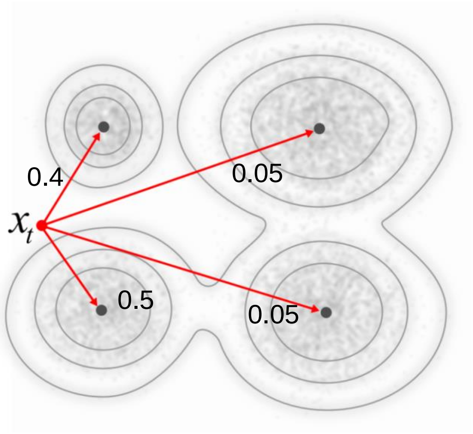
Relationship with the BOV

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.

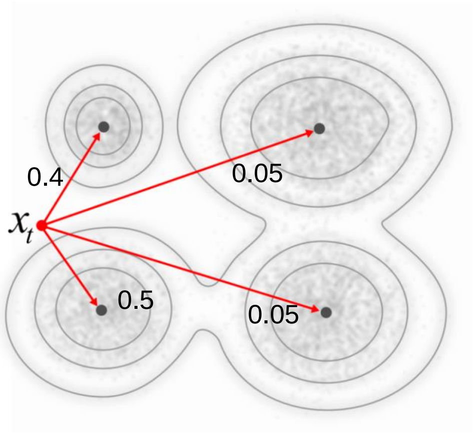
The Fisher vector
Relationship with the BOV

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

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

$\gamma_t(i)$ = 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

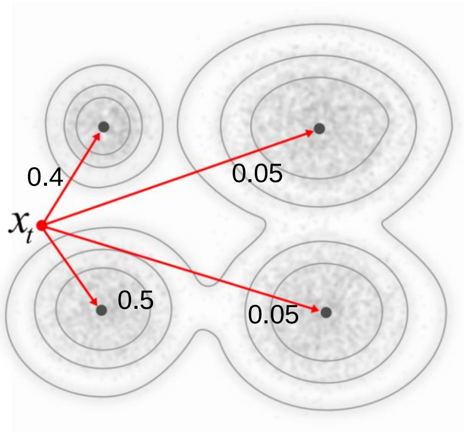
Relationship with the BOV

FV formulas:

- gradient wrt to w

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

$\rightarrow$ **soft BOV**



- gradient wrt to μ and σ

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

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

$\rightarrow$ compared to BOV, include **higher-order statistics** (up to order 2)

$\rightarrow$ FV **much higher-dim** than BOV for a **given visual vocabulary size**

$\rightarrow$ 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:

$\rightarrow$ uncorrelated features are more consistent with diagonal assumption of covariance matrices in GMM

$\rightarrow$ FK performs whitening and enhances low-energy (possibly noisy) dimensions

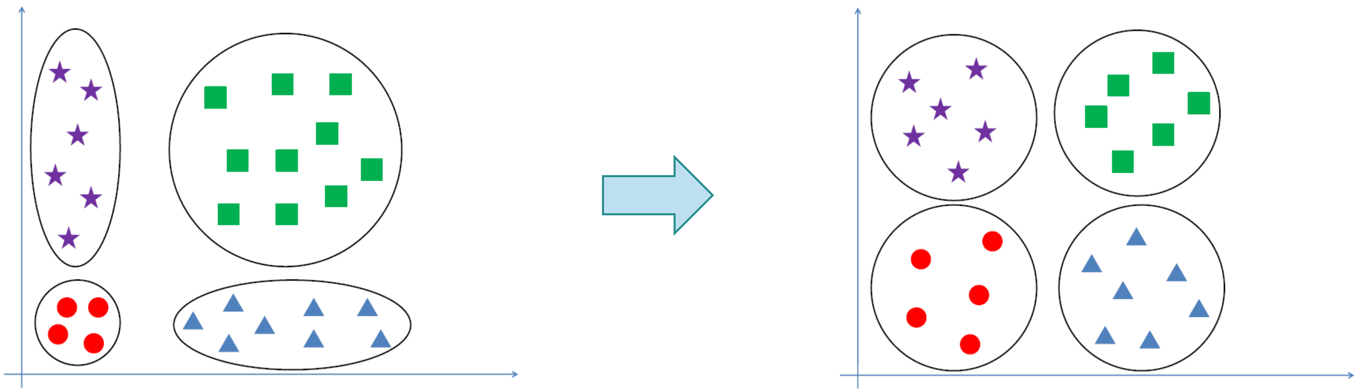
The Fisher vector

Normalization: variance stabilization

→ **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).



→ Reduce impact of bursty visual elements

Jégou, Douze, Schmid, “On the burstiness of visual elements”, ICCV’09.

Datasets for image retrieval

INRIA Holidays dataset: 1491 shots of personal Holiday snapshot

- 500 queries, each associated with a small number of results 1-11 results
- 1 million distracting images (with some “false false” positives)

Hervé Jégou, Matthijs Douze and Cordelia Schmid
Hamming Embedding and Weak Geometric consistency for large-scale image search, *ECCV'08*

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	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

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

$\rightarrow$ 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 $\uparrow$

$\rightarrow$ 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:

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

Vlfeat does it too!

http://www.vlfeat.org

Summary

- We've looked at methods to better characterize the distribution of visual words in an image:
  - Soft assignment (a.k.a. Kernel Codebook)
  - VLAD
  - Fisher Vector

- Mixtures of Gaussians is conceptually a soft form of kmeans which can better model the data distribution.