Extended Bag-of-Words Formalism for Image Classification

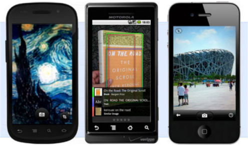
Sandra Avila[1,2] (Cotutelle PhD Candidate),
Arnaldo de A. Araújo[1] (Advisor), Matthieu Cord[2] (Advisor),
Nicolas Thome[2] (Co-Advisor), Eduardo Valle[3] (Collaborator)

[1]Federal University of Minas Gerais, NPDI Lab – UFMG, Belo Horizonte, Brazil
[2]Pierre and Marie Curie University, UPMC-Sorbonne Universities, LIP6, Paris, France
[3]State University of Campinas, RECOD Lab, FEEC – UNICAMP, Campinas, Brazil

Image Classification: Why do we care?

Google Image Search
Ask.com
altavista
bing
YAHOO! SEARCH

Web Search

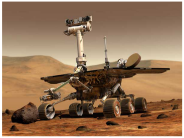


Mobile Search

Visual Search

Surveillance

Medical Diagnosis



Robot Vision

Pornography Detection

Biometric Security

Sociology Research

SmugMug flickr Picasa WOOPHY

facebook phanfare Instagram Pixable shutterfly

ImageShack online media hosting BlueMelon FotKi webshots don't blink

photobucket snapfish imgur

Panoramio streamzoo iStockphoto

Huge amount of image is available

Why image classification is a hard problem?

Many classes and concepts

Viewpoint changes

Illumination variations

Occlusion

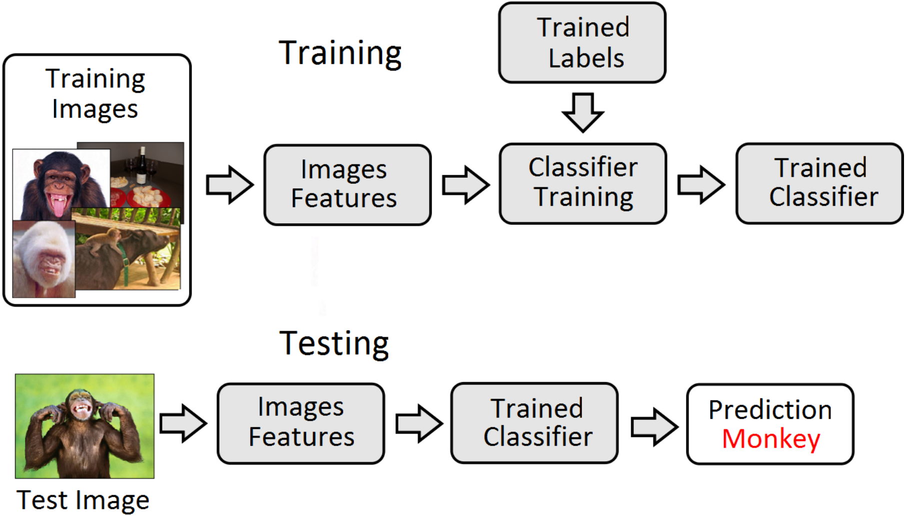
Background clutter

Inter-class similarity

Intra-class diversity

Much diversity in the data

How do we classify images?

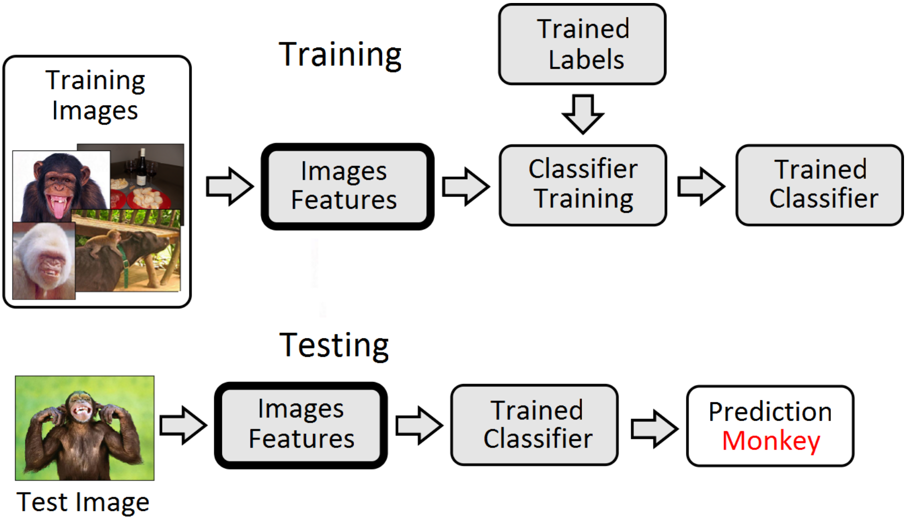

Training
Training Images
Images Features
Trained Labels
Classifier Training
Trained Classifier
Testing
Test Image
Images Features
Trained Classifier
Prediction
Monkey

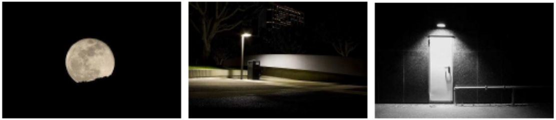
Problem Statement

Given an image dataset,

how to represent their visual content information

for a classification task?


Training
Training Images
Images Features
Trained Labels
Classifier Training
Trained Classifier
Testing
Images Features
Trained Classifier
Prediction
Monkey
Test Image



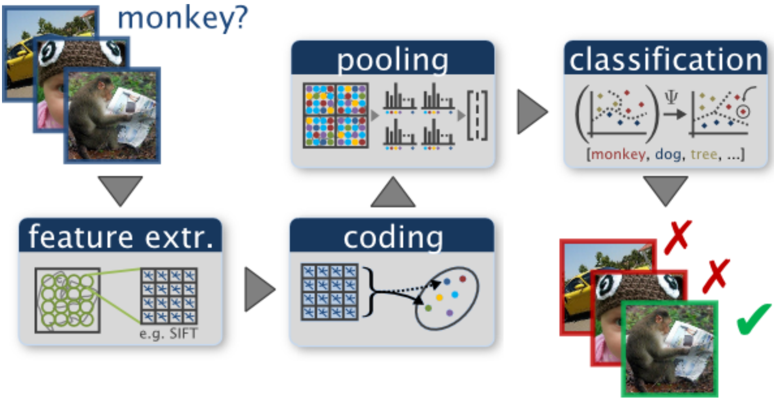
night scenes

sunset scenes

young people

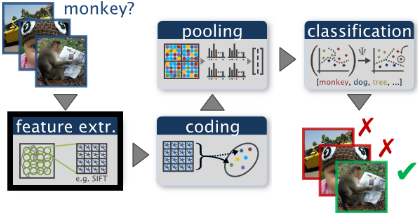
old people

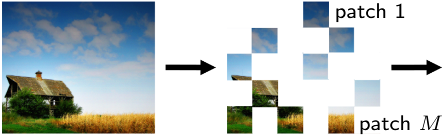
Bag-of-Visual-Words (**BoW**)
[Sivic and Zisserman, 2003; Csurka et al., 2004]



Slide credit: Ken Chatfield

Low-level Visual Feature Extraction



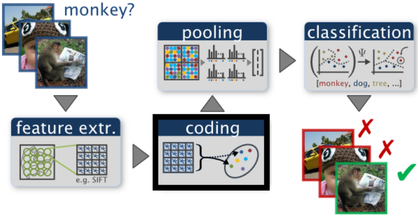


$$\begin{bmatrix} l_{1,1} & \dots & l_{1,N} \\ l_{2,1} & \dots & l_{2,N} \\ \vdots & & \vdots \\ l_{M,1} & \dots & l_{M,N} \end{bmatrix}$$

Local feature extraction

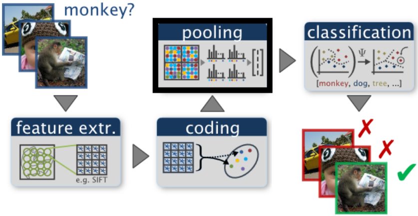
- **Patch detection**: interest points, dense sampling, ...
- **Feature extraction**: SIFT [Lowe, 2004], SURF [Bay et al., 2008], ...

Visual Codebook Coding step

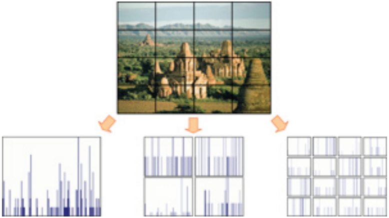


- **Visual codebook learning**: random, unsupervised (e.g., k-means, GMM), supervised [Perronnin et al., 2006; Goh et al., 2012], . . .
- **Coding**: hard-assignment, soft-assignment [van Gemert et al., 2008, 2010], sparse coding [Yang et al., 2009; Boureau et al., 2010], . . .
- **Feature coding based on the vector difference**: VLAD [Jégou et al., 2010], SVC [Zhou et al., 2010], VLAT [Picard et al., 2011], . . .

Pooling step

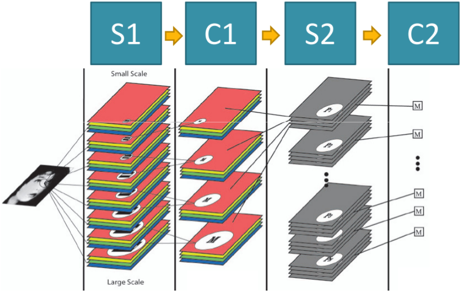


- **Pooling**: sum/average-pooling, max-pooling [Yang et al., 2009], . . .
- **Spatial pooling**: spatial pyramid matching [Lazebnik et al., 2006], [Jia et al., 2012], . . .



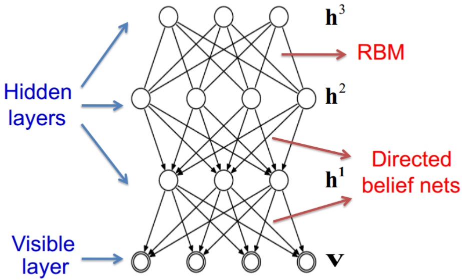
Spatial Pyramid Matching

Other Approaches



Biologically-inspired Models

[Fukushima and Miyake, 1982; LeCun et al., 1990; Riesenhuber and Poggio, 1999; Serre et al., 2007; Thériault et al., 2012]



Deep Learning Models

[Hinton and Salakhutdinov, 2006; Ranzato et al., 2007; Bengio, 2009]

BossaNova Representation

Coding & Pooling Matrix Representation

$$\mathbf{H} = \begin{array}{c} \\ \mathbf{c}_1 \\ \vdots \\ \mathbf{c}_m \\ \vdots \\ \mathbf{c}_M \end{array} \begin{array}{c} \begin{array}{ccccc} \mathbf{x}_1 & \dots & \mathbf{x}_j & \dots & \mathbf{x}_N \end{array} \\ \begin{bmatrix} \alpha_{1,1} & \dots & \alpha_{1,j} & \dots & \alpha_{1,N} \\ \vdots & & \vdots & & \vdots \\ \alpha_{m,1} & \dots & \alpha_{m,j} & \dots & \alpha_{m,N} \\ \vdots & & \vdots & & \vdots \\ \alpha_{M,1} & \dots & \alpha_{M,j} & \dots & \alpha_{M,N} \end{bmatrix} \end{array}$$

Notations:

$\mathcal{X} = \{\mathbf{x}_j\}$, $j \in \{1, \dots, N\}$: set of local descriptors (e.g., SIFT)

$\mathcal{C} = \{\mathbf{c}_m\}$, $m \in \{1, \dots, M\}$: visual codebook

Coding & Pooling Matrix Representation

$$\mathbf{H} = \begin{array}{cc} & \begin{array}{ccccc} \mathbf{x}_1 & \dots & \mathbf{x}_j & \dots & \mathbf{x}_N \end{array} \\ \begin{array}{c} \mathbf{c}_1 \\ \vdots \\ \mathbf{c}_m \\ \vdots \\ \mathbf{c}_M \end{array} & \begin{bmatrix} \alpha_{1,1} & \dots & \alpha_{1,j} & \dots & \alpha_{1,N} \\ \vdots & & \vdots & & \vdots \\ \alpha_{m,1} & \dots & \alpha_{m,j} & \dots & \alpha_{m,N} \\ \vdots & & \vdots & & \vdots \\ \alpha_{M,1} & \dots & \alpha_{M,j} & \dots & \alpha_{M,N} \end{bmatrix} \end{array}$$

$$\Downarrow$$

f : **Coding**

Coding: $\mathbf{x}_j \rightarrow f(\mathbf{x}_j) = \{\alpha_{m,j}\}, \quad \alpha_{m,j} = 1$ iff $m = \underset{k \in \{1,\dots,M\}}{\arg\min} \|\mathbf{x}_j - \mathbf{c}_k\|_2^2$

Coding & Pooling Matrix Representation

$$
\mathbf{H} = \begin{array}{c c}
 & \begin{array}{ccccc} \mathbf{x}_1 & \dots & \mathbf{x}_j & \dots & \mathbf{x}_N \end{array} \\
\begin{array}{c} \mathbf{c}_1 \\ \vdots \\ \mathbf{c}_m \\ \vdots \\ \mathbf{c}_M \end{array} &
\begin{bmatrix}
\alpha_{1,1} & \dots & \alpha_{1,j} & \dots & \alpha_{1,N} \\
\vdots & & \vdots & & \vdots \\
\alpha_{m,1} & \dots & \alpha_{m,j} & \dots & \alpha_{m,N} \\
\vdots & & \vdots & & \vdots \\
\alpha_{M,1} & \dots & \alpha_{M,j} & \dots & \alpha_{M,N}
\end{bmatrix}
\end{array}
\Rightarrow g : \textbf{Pooling}
$$

Coding: $\mathbf{x}_j \to f(\mathbf{x}_j) = \{\alpha_{m,j}\}, \quad \alpha_{m,j} = 1$ iff $m = \underset{k \in \{1,\dots,M\}}{\arg\min} \|\mathbf{x}_j - \mathbf{c}_k\|_2^2$

Pooling: $g(\{\alpha_j\}) = \mathbf{z} : \ \forall m, \ z_m = \sum_{j=1}^{N} \alpha_{m,j}$

Coding & Pooling Matrix Representation

$$\mathbf{H} = \begin{array}{cc} & \begin{array}{ccccc} \mathbf{x}_1 & \dots & \mathbf{x}_j & \dots & \mathbf{x}_N \end{array} \\ \begin{array}{c} \mathbf{c}_1 \\ \vdots \\ \mathbf{c}_m \\ \vdots \\ \mathbf{c}_M \end{array} & \begin{bmatrix} \alpha_{1,1} & \dots & \alpha_{1,j} & \dots & \alpha_{1,N} \\ \vdots & & \vdots & & \vdots \\ \alpha_{m,1} & \dots & \alpha_{m,j} & \dots & \alpha_{m,N} \\ \vdots & & \vdots & & \vdots \\ \alpha_{M,1} & \dots & \alpha_{M,j} & \dots & \alpha_{M,N} \end{bmatrix} \end{array} \qquad \mathbf{z} = \begin{bmatrix} z_1 \\ \vdots \\ z_m \\ \vdots \\ z_M \end{bmatrix}$$

Coding: $\mathbf{x}_j \rightarrow f(\mathbf{x}_j) = \{\alpha_{m,j}\}, \quad \alpha_{m,j} = 1$ iff $m = \underset{k \in \{1,\ldots,M\}}{\arg\min} \|\mathbf{x}_j - \mathbf{c}_k\|_2^2$

Pooling: $g(\{\alpha_j\}) = \mathbf{z} : \ \forall m, \ z_m = \sum_{j=1}^{N} \alpha_{m,j}$

BoW representation: $\mathbf{z} = [z_1, z_2, \cdots, z_M]^{\mathsf{T}}$

Early Ideas

- We pointed out the weakness in the standard pooling operation used in the BoW signature generation.
- Instead of averaging all the values from one row in the $\mathbf{H}$ matrix, we proposed to describe their distribution.
- BOSSA representation (**B**ag **O**f **S**tatistical **S**ampling **A**nalysis) introduces **our density function-based pooling strategy**.

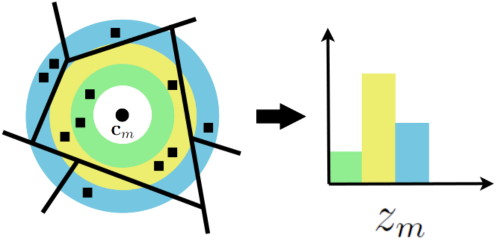
Early Ideas

- We pointed out the weakness in the standard pooling operation used in the BoW signature generation.
- Instead of averaging all the values from one row in the $\mathbf{H}$ matrix, we proposed to describe their distribution.
- BOSSA representation (**B**ag **O**f **S**tatistical **S**ampling **A**nalysis) introduces **our density function-based pooling strategy**.

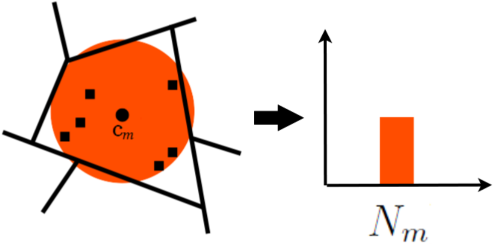
Early Ideas

- We pointed out the weakness in the standard pooling operation used in the BoW signature generation.
- Instead of averaging all the values from one row in the $\mathbf{H}$ matrix, we proposed to describe their distribution.
- BOSSA representation (**B**ag **O**f **S**tatistical **S**ampling **A**nalysis) introduces **our density function-based pooling strategy**.

Our Pooling Illustration



Our Pooling



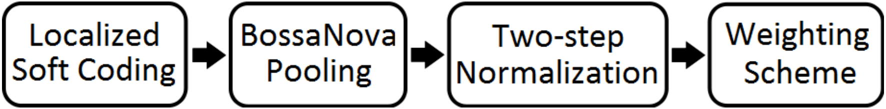
BoW Pooling

Our Pooling Formalism

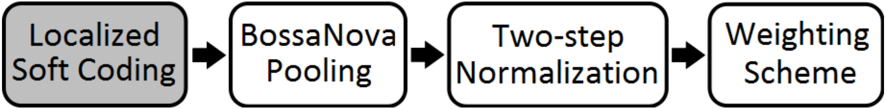
$$
\begin{aligned}
g : \mathbb{R}^N &\longrightarrow \mathbb{R}^B \\
\alpha_{\mathbf{m}} &\longrightarrow g(\alpha_m) = z_m \\
z_{m,b} &= \operatorname{card}\left(\mathbf{x}_j \mid \alpha_{m,j} \in \left[\frac{b}{B}; \frac{b+1}{B}\right]\right) \\
& \frac{b}{B} \geq \alpha_m^{min} \text{ and } \frac{b+1}{B} \leq \alpha_m^{max}
\end{aligned}
$$

B denotes the number of bins of each histogram z_m, and $[\alpha_m^{min}; \alpha_m^{max}]$ limits the range of distances

BossaNova Representation

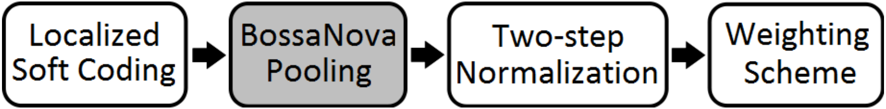


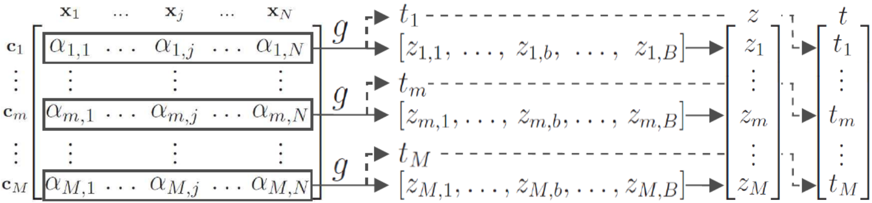
BossaNova Representation



$$
\begin{array}{c c}
 & \begin{array}{ccccc} \mathbf{x}_1 & \dots & \mathbf{x}_j & \dots & \mathbf{x}_N \end{array} \\
\begin{array}{c} \mathbf{c}_1 \\ \vdots \\ \mathbf{c}_m \\ \vdots \\ \mathbf{c}_M \end{array} &
\begin{bmatrix}
\alpha_{1,1} & \dots & \alpha_{1,j} & \dots & \alpha_{1,N} \\
\vdots & & \vdots & & \vdots \\
\alpha_{m,1} & \dots & \alpha_{m,j} & \dots & \alpha_{m,N} \\
\vdots & & \vdots & & \vdots \\
\alpha_{M,1} & \dots & \alpha_{M,j} & \dots & \alpha_{M,N}
\end{bmatrix}
\end{array}
$$

$$
\alpha_{m,j} = \frac{exp^{-\beta_m d_2(\mathbf{x}_j, \mathbf{c}_m)}}{\sum_{m'=1}^{K} exp^{-\beta_m d_2(\mathbf{x}_j, \mathbf{c}_{m'})}}
$$


Localized Soft Coding
BossaNova Pooling
Two-step Normalization
Weighting Scheme


$\mathbf{x}_1$ ... $\mathbf{x}_j$ ... $\mathbf{x}_N$
$\mathbf{c}_1$
$\mathbf{c}_m$
$\mathbf{c}_M$
$\alpha_{1,1} \dots \alpha_{1,j} \dots \alpha_{1,N}$
$\alpha_{m,1} \dots \alpha_{m,j} \dots \alpha_{m,N}$
$\alpha_{M,1} \dots \alpha_{M,j} \dots \alpha_{M,N}$
g
t_1
$[z_{1,1}, \dots, z_{1,b}, \dots, z_{1,B}]$
t_m
$[z_{m,1}, \dots, z_{m,b}, \dots, z_{m,B}]$
t_M
$[z_{M,1}, \dots, z_{M,b}, \dots, z_{M,B}]$
z
z_1 z_m z_M
t
t_1 t_m t_M

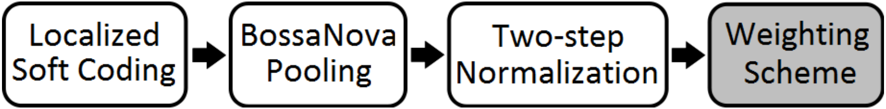
BossaNova Representation

Localized Soft Coding → BossaNova Pooling → **Two-step Normalization** → Weighting Scheme

PN $\longrightarrow$ $\boxed{\begin{array}{c} z_{m,b} = \sqrt{z_{m,b}} \\ t_m = \sqrt{t_m} \end{array}}$

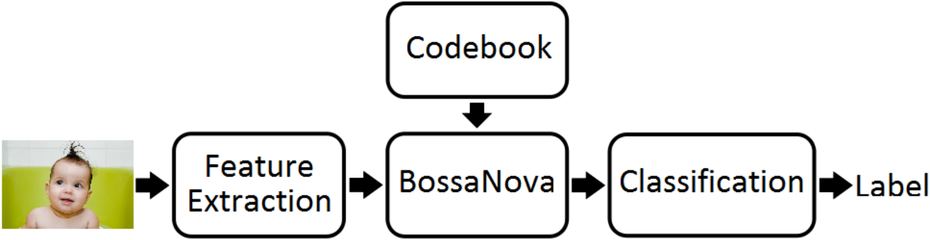
&

ℓ_2-norm $\longrightarrow$ $\boxed{\begin{array}{c} z = z/\|z\|_2 \\ t = t/\|t\|_2 \end{array}}$

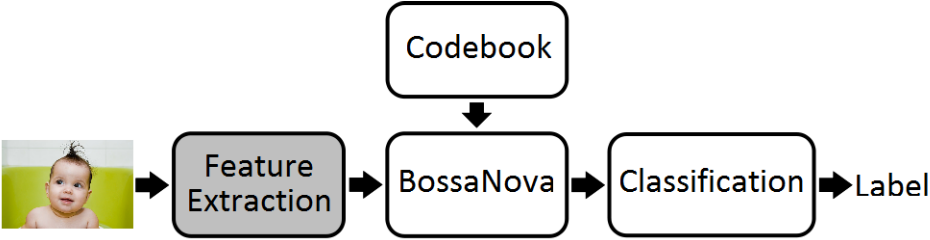


$$\begin{bmatrix} z_1, st_1 \\ \vdots \\ z_m, st_m \\ \vdots \\ z_M, st_M \end{bmatrix}$$

BossaNova Scheme



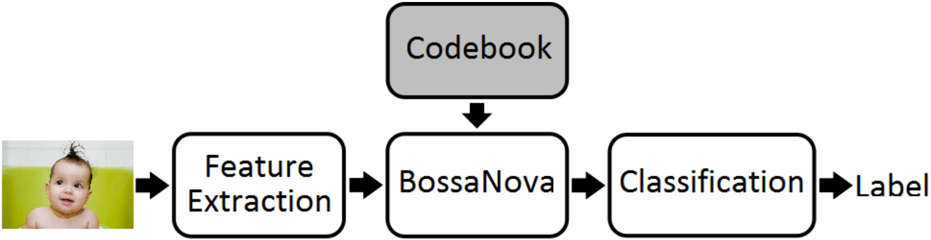
BossaNova Scheme



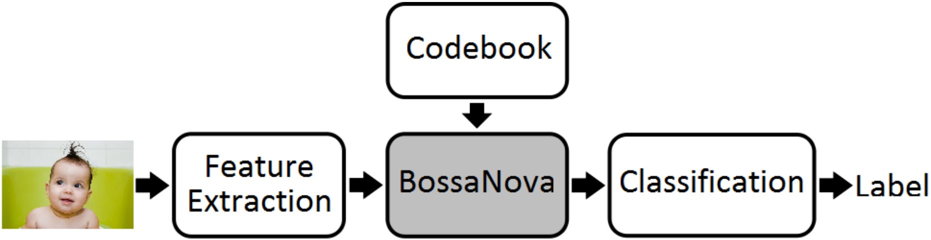
- **SIFT descriptors** on a dense spatial grid at multiple scales
- Dimensionality reduction by applying **PCA** (128 $\rightarrow$ 64)

BossaNova Scheme

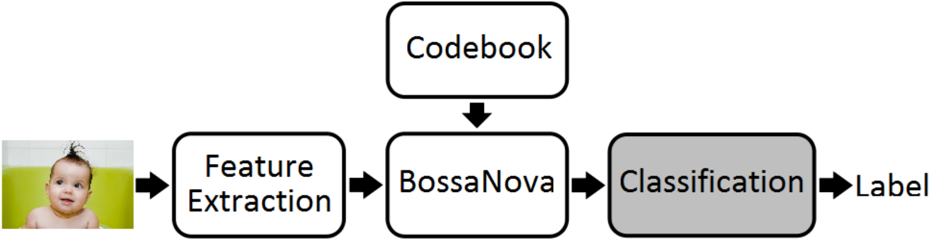
- **k-means algorithm**



BossaNova Scheme



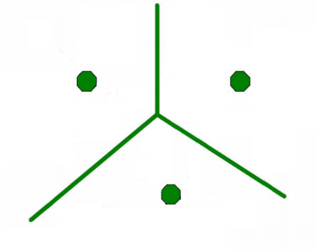
BossaNova Scheme



- **SVM classifiers** are applied by using a **nonlinear Gauss–ℓ_2 kernel**

Let us consider the underlying distribution of the local features x as a mixture of several (basic) distribution functions $p_k(x)$:

$$p(x|\theta) = p_\theta(x) = \sum_{k=1}^{K} w_k p_k(x) \qquad (1)$$



BossaNova as a Generative Formalism

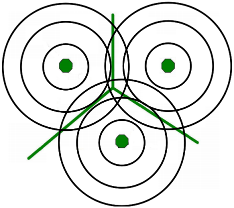
BossaNova: a mixture of B constant non overlapping radius-based functions $p_b(x|k)$ between α_k^{min} and α_k^{max} to each visual word c_k:

$$p_k(x) = \sum_{b=1}^{B} w_{(b,k)} p_b(x|k) \quad (2)$$

$$p_b(x|k) = \mathbb{1}_{\alpha_k^{min} + (b-1)\Delta_k \leq ||x - c_k|| \leq \alpha_k^{min} + b\Delta_k}$$

Combining with global mixtures, **the generative model is**:

$$p(x|\theta) = p_\theta(x) = \sum_{k=1}^{K} w_k \left(\sum_{b=1}^{B} w_{(b,k)} p_b(x|k) \right)$$



BossaNova as a Fisher Kernel Formalism

Fisher kernel from our generative model:

Fisher Representation [Jaakkola and Haussler, 1998; Perronnin and Dance, 2007]: log-likelihood of $p(x|\theta)$.

The resulting scores are:

$$g(\alpha_k, X) = \frac{1}{T}\sum_{t=1}^{T} \gamma_k(x_t) - w_k \quad (3)$$

$$g(\beta_{(b,k)}, X) = \frac{1}{T}\sum_{t=1}^{T} \left(\gamma_{(b,k)}(x_t) - w_{(b,k)}\right) \gamma_k(x_t) \quad (4)$$

The Fisher score is easy to compute for the (Fisher) BossaNova model.

Experimental Results

Experimental Results

1. BOSSA to BossaNova Improvements Analysis
2. BossaNova Parameter Evaluation
3. Comparison of State-of-the-Art Methods
4. BossaNova in the ImageCLEF 2012 Challenge

Experimental Results – Datasets

- **MIRFLICKR**: 25,000 images, 38 class

- **ImageCLEF 2011 Photo Annotation**: 18,000 images, 99 class

- **PASCAL VOC 2007**: 9,963 images, 20 class

- **15-Scenes**: 4,485 images, 15 class

1. BOSSA to BossaNova Improvements Analysis
2. BossaNova Parameter Evaluation
3. Comparison of State-of-the-Art Methods
4. BossaNova in the ImageCLEF 2012 Challenge

- **ANOVA**: to measure the relative impact of each improvement

  Weight: 3% of the BossaNova performance

  Soft: 48% of the BossaNova performance

  Norm: 31% of the BossaNova performance

  Weight-Soft: 9% of the BossaNova performance

- **t-test**: to evaluate the relevance of the three modifications

  Weight: No = no cross-validation, Yes = cross-validation

  Soft: No = hard assignment, Yes = localized soft assignment

  Norm: No = ℓ_1 block norm, Yes = power normalization + ℓ_2-norm

Table: Impact of the proposed improvements to the BossaNova on VOC 2007.

	Weight	Soft	Norm	mAP	CI (95%)
1	No	No	No	54.9 ± 0.5	
2	Yes	No	No	55.2 ± 0.4	$2 \leftrightarrow 1$ ✓
3	No	Yes	No	55.8 ± 0.5	$3 \leftrightarrow 1$ ✓
4	No	No	Yes	55.6 ± 0.4	$4 \leftrightarrow 1$ ✓
5	Yes	No	Yes	55.9 ± 0.4	$5 \leftrightarrow 1$ ✓, $5 \leftrightarrow 4$ ✓
6	Yes	Yes	No	56.4 ± 0.4	$6 \leftrightarrow 1$ ✓, $6 \leftrightarrow 4$ ✓
7	No	Yes	Yes	58.1 ± 0.4	$7 \leftrightarrow 1$ ✓, $7 \leftrightarrow 4$ ✓
8	Yes	Yes	Yes	58.8 ± 0.4	$8 \leftrightarrow 1$ ✓, $8 \leftrightarrow 7$ ✓

Experimental Results – BOSSA to BossaNova

- **t-test**: to evaluate the relevance of the three modifications

  Weight: No = no cross-validation, Yes = cross-validation

  Soft: No = hard assignment, Yes = localized soft assignment

  Norm: No = ℓ_1 block norm, Yes = power normalization + ℓ_2-norm

Table: Impact of the proposed improvements to the BossaNova on VOC 2007.

		Weight	Soft	Norm	mAP	CI (95%)
BOSSA	1	**No**	**No**	**No**	54.9 ± 0.5	
	2	Yes	No	No	55.2 ± 0.4	$2 \leftrightarrow 1$ ✓
	3	No	Yes	No	55.8 ± 0.5	$3 \leftrightarrow 1$ ✓
	4	No	No	Yes	55.6 ± 0.4	$4 \leftrightarrow 1$ ✓
	5	Yes	No	Yes	55.9 ± 0.4	$5 \leftrightarrow 1$ ✓, $5 \leftrightarrow 4$ ✓
	6	Yes	Yes	No	56.4 ± 0.4	$6 \leftrightarrow 1$ ✓, $6 \leftrightarrow 4$ ✓
	7	No	Yes	Yes	58.1 ± 0.4	$7 \leftrightarrow 1$ ✓, $7 \leftrightarrow 4$ ✓
BossaNova	8	**Yes**	**Yes**	**Yes**	58.8 ± 0.4	$8 \leftrightarrow 1$ ✓, $8 \leftrightarrow 7$ ✓

Experimental Results

1. BOSSA to BossaNova Improvements Analysis
2. BossaNova Parameter Evaluation
3. Comparison of State-of-the-Art Methods
4. BossaNova in the ImageCLEF 2012 Challenge

The key parameters in BossaNova representation are:

- the number of codewords M
- the number of bins B in each local histogram z_m
- the range of distances $[\alpha_m^{min}, \alpha_m^{max}]$

The key parameters in BossaNova representation are:

- **the number of codewords** M
- the number of bins B in each local histogram z_m
- **the range of distances** $[\alpha_m^{min}, \alpha_m^{max}]$

Number of codewords M (using $B = 2$)

- BossaNova $vs.$ BoW

	Codebook size			
	1024	2048	4096	8192
BossaNova [Avila et al., 2013]	51.8	52.9	54.4	**55.2**
BoW [Sivic and Zisserman, 2003]	50.3	51.3	**51.5**	51.1

- BossaNova $vs.$ Hierarchical BoW

	Codebook size		
	1024	2048	4096
BossaNova [Avila et al., 2013]	51.8	52.9	**54.4**
Hierarchical BoW	50.6	51.3	**51.4**

Number of codewords M (using $B = 2$)

- BossaNova $vs.$ BoW

	Codebook size			
	1024	2048	4096	8192
BossaNova [Avila et al., 2013]	51.8	52.9	54.4	**55.2**
BoW [Sivic and Zisserman, 2003]	50.3	51.3	**51.5**	51.1

- BossaNova $vs.$ Hierarchical BoW

	Codebook size		
	1024	2048	4096
BossaNova [Avila et al., 2013]	51.8	52.9	**54.4**
Hierarchical BoW	50.6	51.3	**51.4**

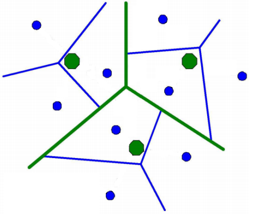
Number of codewords M (using $B = 2$)

- BossaNova *vs.* BoW

	Codebook size			
	1024	2048	4096	8192
BossaNova [Avila et al., 2013]	51.8	52.9	54.4	**55.2**
BoW [Sivic and Zisserman, 2003]	50.3	51.3	**51.5**	51.1

- BossaNova *vs.* Hierarchical BoW

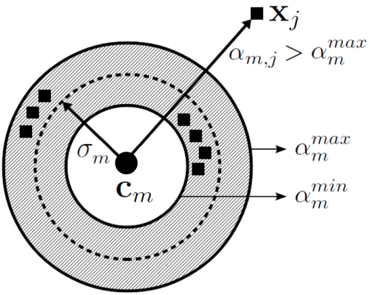
	Codebook size		
	1024	2048	4096
BossaNova [Avila et al., 2013]	51.8	52.9	**54.4**
Hierarchical BoW	50.6	51.3	**51.4**



Minimum Distance α_m^{min} (using $M = 4096$, $B = 2$)

Range of distances	mAP
$\lambda_{min} = 0.0$, $\lambda_{max} = 2.0$	54.4
$\lambda_{min} = 0.4$, $\lambda_{max} = 2.0$	**54.9**

$$\alpha_m^{min} = \lambda_{min} \cdot \sigma_m$$
$$\alpha_m^{max} = \lambda_{max} \cdot \sigma_m$$



1. BOSSA to BossaNova Improvements Analysis
2. BossaNova Parameter Evaluation
3. Comparison of State-of-the-Art Methods
4. BossaNova in the ImageCLEF 2012 Challenge

Experimental Results – Comparison of State-of-the-Art

- Datasets:

  MIRFLICKR, ImageCLEF 2011, PASCAL VOC 2007, 15-Scenes

- Implemented methods:

  Bag-of-Words (BoW), Fisher Vector (FV),

  BOSSA, BossaNova (BN)

- Datasets:

  MIRFLICKR, ImageCLEF 2011, **PASCAL VOC 2007**, 15-Scenes

- Implemented methods:

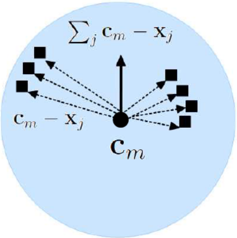
  Bag-of-Words (BoW), Fisher Vector (FV),

  BOSSA, BossaNova (BN)

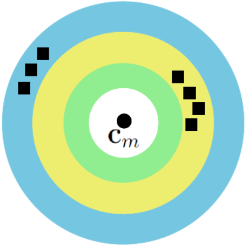
Experimental Results – MIRFLICKR

		mAP (%)
Our methods	BOSSA [Avila et al., 2011]	52.7
	BN [Avila et al., 2013]	**54.4**
Implemented methods	BoW [Sivic and Zisserman, 2003]	51.5
	FV [Perronnin et al., 2010]	54.3
Published results	[Huiskes et al., 2010]	37.5
	[Guillaumin et al., 2010]	53.0

BossaNova & Fisher Vector: Pooling Complementarity



Fisher Vector
(average-pooling)



BossaNova
(our pooling)

Combination: Linear kernel combination or Late fusion

$$K_{BN+FV} = \varphi \cdot K_{BN} + (1 - \varphi) \cdot K_{FV}$$

Experimental Results – MIRFLICKR

		mAP (%)
Our methods	BOSSA [Avila et al., 2011]	52.7
	BN [Avila et al., 2013]	54.4
	BN + FV [Avila et al., 2013]	**56.0**
Implemented methods	BoW [Sivic and Zisserman, 2003]	51.5
	FV [Perronnin et al., 2010]	54.3
Published results	[Huiskes et al., 2010]	37.5
	[Guillaumin et al., 2010]	53.0

Experimental Results – PASCAL VOC 2007

		mAP (%)
Our methods	BOSSA [Avila et al., 2011]	54.4
	BN [Avila et al., 2013]	58.5
	BN + FV [Avila et al., 2013]	61.6
	Late Fusion (BN + FV)	**62.4**
Implemented methods	BoW [Sivic and Zisserman, 2003]	53.2
	FV [Perronnin et al., 2010]	59.5
Published results	[Krapac et al., 2011]	56.7
	[Chatfield et al., 2011]	61.7
	[Sánchez et al., 2012]	66.3

Experimental Results

1. BOSSA to BossaNova Improvements Analysis
2. BossaNova Parameter Evaluation
3. Comparison of State-of-the-Art Methods
4. BossaNova in the ImageCLEF 2012 Challenge

Experimental Results – ImageCLEF 2012

- ImageCLEF 2012 Photo Annotation: 25,000 images and 94 class
- 13 teams (Brazil, France, Germany, Italy, Japan, Spain, ...)
- 28 visual submissions

	Rank	mAP (%)
[Liu et al., 2012]	1	34.8
BN + FV [Avila et al., 2012]	2	34.4
BN [Avila et al., 2012]	3	33.6
Paper not available	6	33.2
[Ushiku et al., 2012]	10	32.4
[Xioufis et al., 2012]	11	31.8

Experimental Results – ImageCLEF 2012

- ImageCLEF 2012 Photo Annotation: 25,000 images and 94 class
- 13 teams (Brazil, France, Germany, Italy, Japan, Spain, ...)
- 28 visual submissions

	Rank	mAP (%)
[Liu et al., 2012]	1	34.8
BN + FV [Avila et al., 2012]	2	34.4
BN [Avila et al., 2012]	3	33.6
Paper not available	6	33.2
[Ushiku et al., 2012]	10	32.4
[Xioufis et al., 2012]	11	31.8

Application: Pornography Detection

The importance of pornography detection is attested by the **large** literature on the subject.

[Fleck et al., 1996]
[Forsyth and Fleck, 1996]
[Forsyth and Fleck, 1997]
[Forsyth and Fleck, 1999]
[Jones and Rehg, 2002]
[Rowley et al., 2006]
[Lee et al., 2007]
[Zuo et al., 2010]

[Hu et al., 2011]
[Ries and Lienhart, 2012]
[Deselaers et al., 2008]
[Lopes et al., 2009a]
[Lopes et al., 2009b]
[Avila et al., 2011]
[Avila et al., 2013]
[Ulges and Stahl, 2011]

[Steel, 2012]
[Tong et al., 2005]
[Endeshaw et al., 2008]
[Jansohn et al., 2009]
[Valle et al., 2012]
[Rea et al., 2006]
[Liu et al., 2011]
[Ulges et al., 2012]

The importance of pornography detection is attested by the **large** literature on the subject.

Skin Detection

[Fleck et al., 1996]
[Forsyth and Fleck, 1996]
[Forsyth and Fleck, 1997]
[F[illegible] and Fleck, [illegible]]
[Jones and Rehg, 2002]
[Rowley et al., 2006]
[Lee et al., 2007]
[Zuo et al., 2010]

BoW-based Approaches

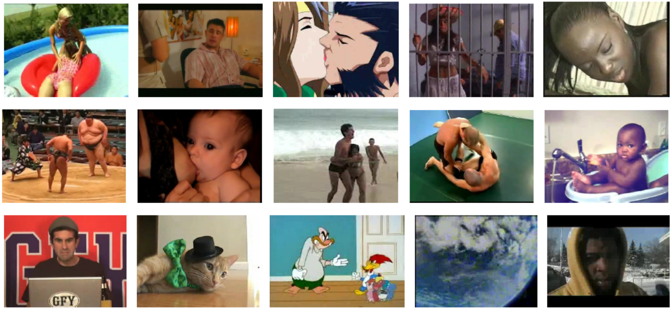
[Hu et al., 2011]
[Ries and Lienhart, 2012]
[Deselaers et al., 2008]
[Lopes et al., 2009a]
[Lo[illegible]]
[Av[illegible]]
[Avila et al., 2013]
[Ulges and Stahl, 2011]

Spatiotemporal Features

Audio Features

[Steel, 2012]
[Tong et al., 2005]
[[illegible]]
[[illegible], 2009]
[Valle et al., 2012]
[Rea et al., 2006]
[[illegible]]
[Ulges et al., 2012]

Pornography Database: nearly 80 hours, 800 videos: 400 porn, 200 non-porn easy and 200 non-porn difficulty.



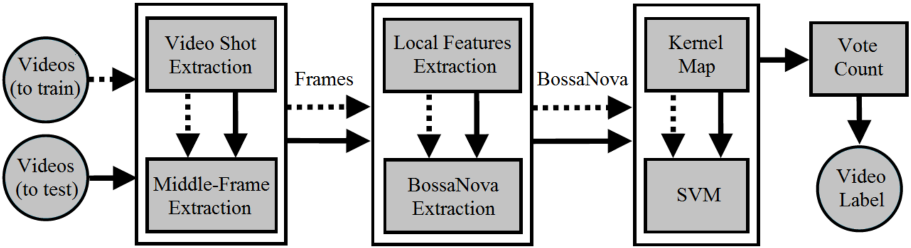
porn

non-porn diff.

non-porn easy

http://www.npdi.dcc.ufmg.br/pornography

Our Scheme



Videos (to train)
Videos (to test)
Video Shot Extraction
Middle-Frame Extraction
Frames
Local Features Extraction
BossaNova Extraction
BossaNova
Kernel Map
SVM
Vote Count
Video Label

Application: Pornography Detection

- BossaNova $vs.$ BOSSA $vs.$ BoW

	mAP (frames)	Accuracy (videos)
Our methods		
BossaNova [Avila et al., 2013]	**96.4 ± 1**	**89.5 ± 1**
BOSSA [Avila et al., 2011]	94.6 ± 1	87.1 ± 2
Implemented methods		
BoW [Sivic and Zisserman, 2003]	91.4 ± 1	83.0 ± 3

- BossaNova $vs.$ PornSeer

		Video was labeled porn	nonporn
Video was	porn	88.2%	11.8%
	nonporn	9.2%	90.8%

		Video was labeled porn	nonporn
Video was	porn	65.1%	34.9%
	nonporn	12.5%	87.5%

Application: Pornography Detection

- BossaNova *vs.* BOSSA *vs.* BoW

	mAP (frames)	Accuracy (videos)
Our methods		
BossaNova [Avila et al., 2013]	**96.4 ± 1**	**89.5 ± 1**
BOSSA [Avila et al., 2011]	94.6 ± 1	87.1 ± 2
Implemented methods		
BoW [Sivic and Zisserman, 2003]	91.4 ± 1	83.0 ± 3

- BossaNova *vs.* PornSeer

		Video was labeled	
		porn	nonporn
Video was	porn	88.2%	11.8%
	nonporn	9.2%	90.8%

		Video was labeled	
		porn	nonporn
Video was	porn	65.1%	34.9%
	nonporn	12.5%	87.5%

Application: Pornography Detection

- BossaNova *vs.* BOSSA *vs.* BoW

	mAP (frames)	Accuracy (videos)
Our methods		
BossaNova [Avila et al., 2013]	**96.4 ± 1**	**89.5 ± 1**
BOSSA [Avila et al., 2011]	94.6 ± 1	87.1 ± 2
Implemented methods		
BoW [Sivic and Zisserman, 2003]	91.4 ± 1	83.0 ± 3

- BossaNova *vs.* PornSeer

		Video was labeled	
		porn	nonporn
Video was	porn	88.2%	11.8%
	nonporn	9.2%	90.8%

		Video was labeled	
		porn	nonporn
Video was	porn	65.1%	34.9%
	nonporn	12.5%	87.5%

Conclusion and Future Work

Contributions

- BossaNova representation
- BossaNova and Fisher Vector's complementarity
- Experimental evaluation
- BossaNova in Pornography detection
- Publication of the BossaNova source code
  www.npdi.dcc.ufmg.br/bossanova

Contributions

- BossaNova representation
- BossaNova and Fisher Vector's complementarity
- Experimental evaluation
- BossaNova in Pornography detection
- Publication of the BossaNova source code
  www.npdi.dcc.ufmg.br/bossanova

Contributions

- BossaNova representation
- BossaNova and Fisher Vector's complementarity
- Experimental evaluation
- BossaNova in Pornography detection
- Publication of the BossaNova source code
  www.npdi.dcc.ufmg.br/bossanova

Contributions

- BossaNova representation
- BossaNova and Fisher Vector's complementarity
- Experimental evaluation
- BossaNova in Pornography detection
- Publication of the BossaNova source code

  www.npdi.dcc.ufmg.br/bossanova

Contributions

- BossaNova representation
- BossaNova and Fisher Vector's complementarity
- Experimental evaluation
- BossaNova in Pornography detection
- Publication of the BossaNova source code

  www.npdi.dcc.ufmg.br/bossanova

Java

Future Work

- BossaNova parameters study
  - Number of bins B
  - Range of distances $[\alpha_m^{min}; \alpha_m^{max}]$
- Large-scale experiments
  - ImageNet LSVR 2010 dataset
    (1000 categories and 1.2 million training images)
- Further exploring the (Fisher) BossaNova model
- Exploit the hierarchical structure

Publications

Journal

- **Avila, S.**, Thome, N., Cord, M., Valle, E., Araújo, A.. Pooling in Image Representation: the Visual Codeword Point of View. *CVIU*, 2013.

International Conferences

- **Avila, S.**, Thome, N., Cord, M., Valle, E., Araújo, A.. BossaNova at ImageCLEF 2012 Flickr Photo Annotation Task. In: *Working Notes of the CLEF*, Rome, 2012.
- **Avila, S.**, Thome, N., Cord, M., Valle, E., Araújo, A.. BOSSA: Extended BoW Formalism for Image Classification. In: *ICIP*, Brussels, 2011.
- Lopes, A., **Avila, S.**, Peixoto, A., Oliveira, R., Araújo, A.. A Bag-of-Features Approach based on Hue-SIFT Descriptor for Nude Detection. *In: EUSIPCO*, Glasgow, 2009.
- Durand, T., Thome, N., Cord, M., **Avila, S.**. Image Classification using Object Detectors (accepted). In: *ICIP*, 2013.

Brazilian Conferences

- **Avila, S.**, Thome, N., Cord, M., Valle, E., Araújo, A.. Extended Bag-of-Words Formalism for Image Classification (accepted). In: *SIBGRAPI*, WTD, 2013.
- Valle, E., **Avila, S.**, Souza, F., Coelho, M., Araújo, A.. Content-Based Filtering for Video Sharing Social Networks. In: *SBSeg*, Curitiba, 2012.
- Lopes, A., **Avila, S.**, Peixoto, A., Oliveira, R., Coelho, M., Araújo, A.. Nude Detection in Video using Bag-of-Visual-Features. In: *SIBGRAPI*, Rio de Janeiro, 2009.

Others

Summer School

- **EMC Summer School on Big Data**. Rio de Janeiro, RJ, Brazil, 04–07 February 2013.
- **ENS/INRIA Visual Recognition and Machine Learning Summer School**. Paris, France, 25–29 July 2011. Poster presentation — BOSSA: extended BoW formalism for image classification.

Workshop

- **Workshop for Women in Machine Learning (WiML)**: Theory, Applications, Experiences. Granada, Spain, December 2011. Poster presentation — BOSSA: extended BoW formalism for image classification.

Thanks! Obrigada! Merci!