Fisher Kernel Representation of images and some of its successful applications

Gabriela Csurka

Xerox Research Center Europe

6 chemin de Maupertuis

38240 Meylan, France

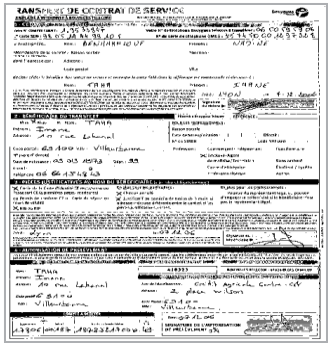
- *Joint-work with Florent Perronnin, Stephane Clinchant, Julien Ah Pine, Luca Marchesotti and others*

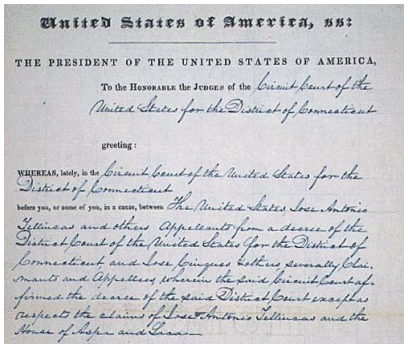
xerox

Xerox Generic Visual Toolbox (GVT)

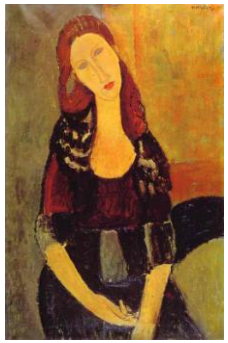
A single flexible framework for the analysis, representation and processing of a wide range of visual content types

photographs





document images



paintings
drawings

maps, charts, tables

Generic Visual Categorization
- to assign one or more labels to a given image, based on its semantic content

Generic Visual Similarity
- a metric between images for content-based retrieval and duplicate detection

Generic Visual Clustering
- to group images by similarity

Generic Visual Segmentation
- to assign one or more labels to each image pixel, based on its semantic content

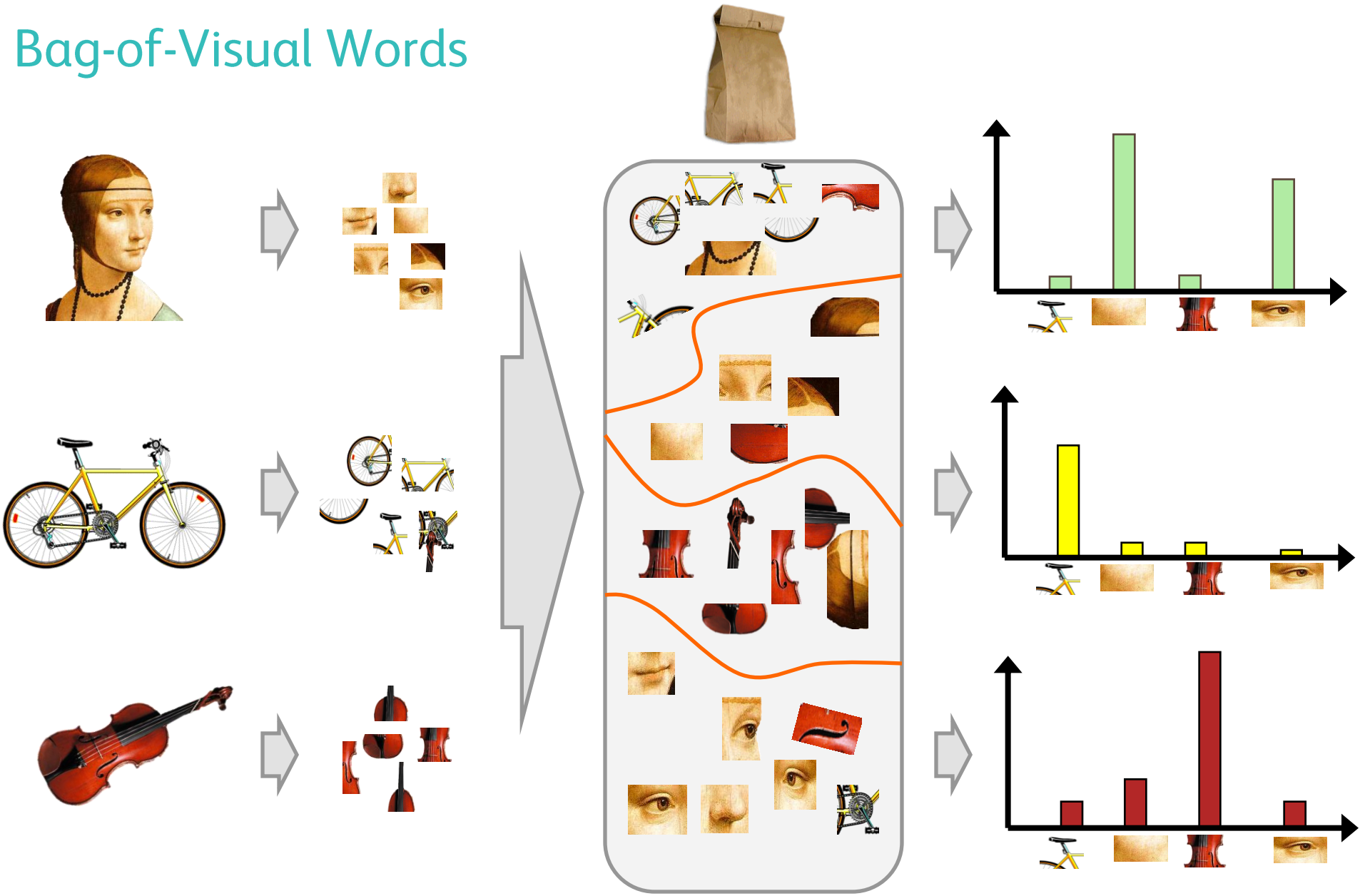
Hybrid approaches
- to include multiple modalities in the analysis

xerox

Outline

- The bag-of-visual word (BOV) and Fisher Kernel image representation
- Generic Visual Categorization
- Large Scale Image Retrieval
- Semantic Image Segmentation
- Intelligent Auto-thumbnailing
- Image Retrieval and Hybrid Content Generation

Bag-of-Visual Words



@ Slide Credit: Marco Bressan and Li Fei-Fei

BoW: Motivation from Text Mining

Order less document representation: frequencies of words from a dictionary Salton & McGill (1983)

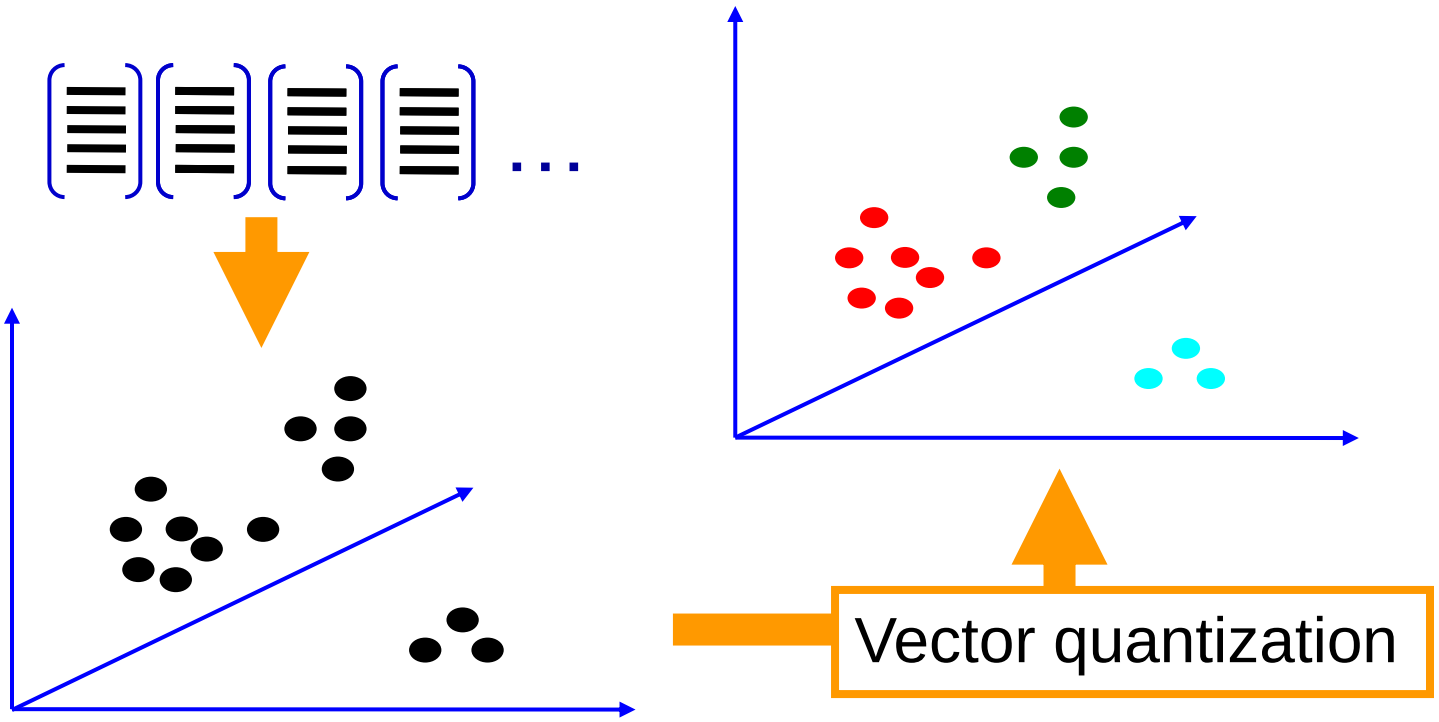
@ Slide Credit: Svetlana Lazebnik

xerox

Visual Vocabulary

Originally = Vector quantization (Kmeans) in some feature space

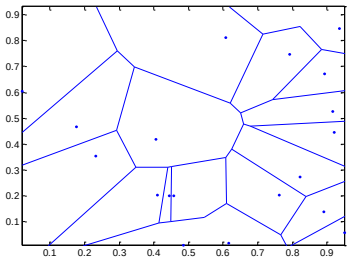
- Julesz, 1981; Cula & Dana, 2001; Leung & Malik 2001 (texture codebook)
- Sivic & Zisserman, ICCV, 2003 (object retrieval)
- Csurka et al., SLCV 2004 (object class categorization)



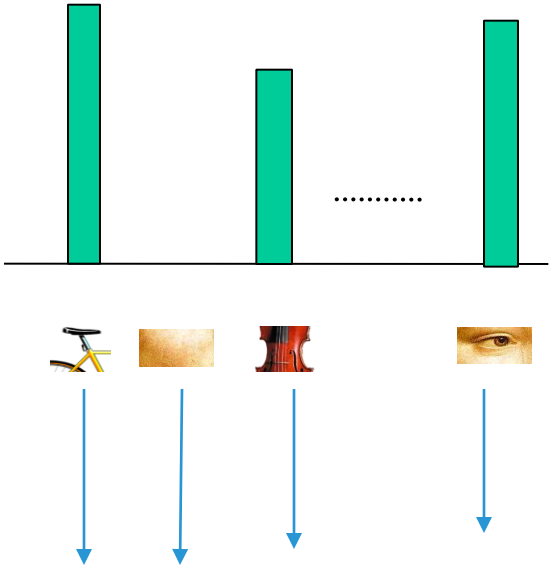
@ Slide Credit: Josef Sivic

BOV: Visual Vocabulary (K-means and GMM)

K-means



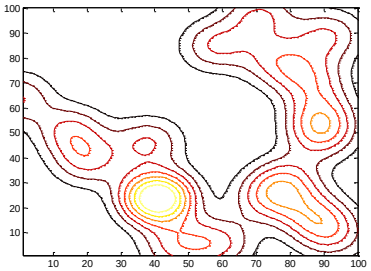
Hard assignment



$$\sum_t [0,0,\cdots,1,\cdots,0]$$

$$B_t = 1_{\{k|x_t \in C_k\}}$$

GMM



$$\sum_t [\gamma_1(x_t), \gamma_2(x_t), \ldots, \gamma_N(x_t)]$$

$$B_t$$

Soft assignment

Visual Vocabulary with a GMM

- Modeling the visual vocabulary in the feature space with a GMM:

$$p(x_t \mid \lambda) = \sum_{i=1}^{N} w_i \, p_i(x_t \mid \lambda) \qquad \text{with} \qquad p_i(x_t \mid \lambda) = \mathcal{N}\left(x_t \mid \mu_i, \Sigma_i\right)$$

- Occupancy probability:

$$\gamma_i(x_t) = p(i \mid x_t, \lambda) = \frac{w_i \, p_i(x_t \mid \lambda)}{\sum_{j=1}^{N} w_j \, p_j(x_t \mid \lambda)}$$

- The parameters λ of the GMM are estimated by EM algorithm maximizing the log-likelihood on the training data*:

$$\log p(X \mid \lambda) = \sum_t \log p(x_t \mid \lambda)$$

- Soft BOV:

$$B_I = \sum_t \left[\gamma_1(x_t), \gamma_2(x_t), \ldots, \gamma_N(x_t)\right]$$

* ***Adapted Vocabularies for Generic Visual Categorization***, F. Perronnin, C. Dance, G. Csurka and M. Bressan, ECCV 2006.

The Fisher Vector -1

- Given a generative model (GMM) with parameters λ and image $I=\{x_t, t=1..T\}$
  - the gradient vector

$$\nabla_\lambda \log p(I \mid \lambda) = \sum_t \nabla_\lambda \log p(x_t \mid \lambda)$$

  - normalized by the Fisher information matrix

$$F_\lambda = E\left[\nabla_\lambda \log p(I \mid \lambda) \cdot (\nabla_\lambda \log p(I \mid \lambda))^{\mathrm{T}}\right]$$

  - leads to a unique "*model-dependent*" but "*class-independent*" representation of the image, called ***Fisher Vector****

$$V_I = F_\lambda^{-1/2} \nabla_\lambda \log p(I \mid \lambda) = F_\lambda^{-1/2} \sum_t \nabla_\lambda \log p(x_t \mid \lambda)$$

* ***Fisher Kernels on Visual Vocabularies for Image Categorization***, F. Perronnin and C. Dance, CVPR 2007.

The Fisher Vector -formulas

- We can deduce the following formulas for the partial derivatives*:

$$\frac{\partial \log p(x_t \mid \lambda)}{\partial w_i} = \left[\frac{\gamma_i(x_t)}{w_i} - \frac{\gamma_1(x_t)}{w_1} \right] \quad , \quad \frac{\partial \log p(x_t \mid \lambda)}{\partial \mu_i^d} = \gamma_i(x_t) \left[\frac{x_t^d - \mu_i^d}{(\sigma_i^d)^2} \right]$$

$$\frac{\partial \log p(x_t \mid \lambda)}{\partial \sigma_i^d} = \gamma_i(x_t) \left[\frac{(x_t^d - \mu_i^d)^2}{(\sigma_i^d)^3} - \frac{1}{\sigma_i^d} \right]$$

- and a diagonal approximation for the Fisher Information matrix*:

$$F_{w_i} = T \left[\frac{1}{w_i} + \frac{1}{w_1} \right] \; , \; F_{\mu_i^d} = \frac{T w_i}{(\sigma_i^d)^2} \; , \; F_{\sigma_i^d} = \frac{2 T w_i}{(\sigma_i^d)^2}$$

- leading to:

$$f_I = (F_\lambda^{-1/2})^{\mathrm{T}} \sum_t \underbrace{\left[\ldots, \frac{\partial \log p(x_t \mid \lambda)}{\partial \mu_i^d}, \ldots, \frac{\partial \log p(x_t \mid \lambda)}{\partial \sigma_i^d}, \ldots \right]}_{f_{x_t}}$$

* ***Fisher Kernels on Visual Vocabularies for Image Categorization***, F. Perronnin and C. Dance, CVPR 2007.

The Fisher Vector -3

Notes:

- the Fisher Vector describes in which direction the parameters of the model λ should be modified to best fit the data (image I)
- the gradient with respect to the mixture weights does not contain significant extra information (we ignore them)
- dimension L= 2 x D x N, where D is the dimension of low level features and N is the number of Gaussians
- sparse, as only a few number of components i (typically < 5) have a non-negligible "occurrence probability " $\gamma_i(x_t)$ for a given t.

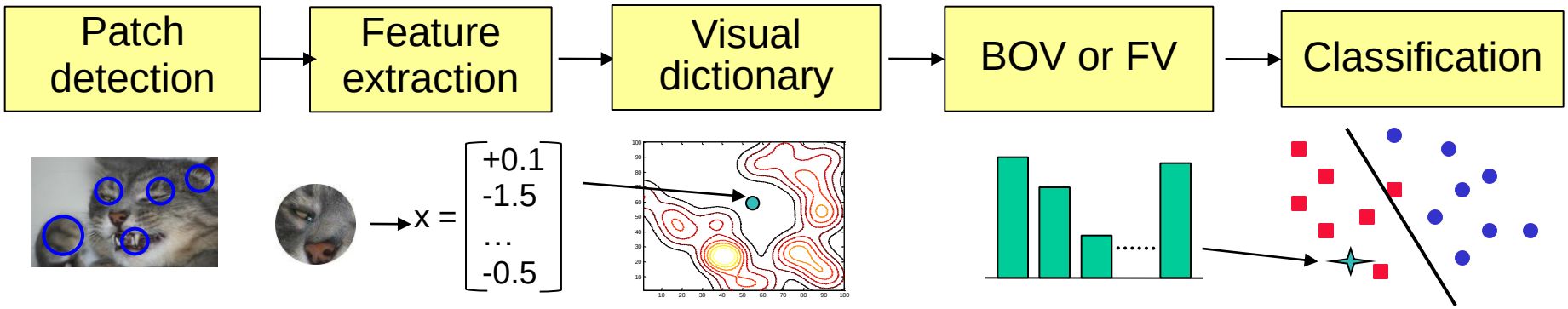
Advantages:

- a linear decision function on the Fisher vectors corresponds to a piecewise-quadratic decision function in the original space
- this "*model-dependent*" but "*class-independent*" representation allows its usage both in supervised (categorization, semantic image segmentation) and unsupervised tasks (clustering, retrieval).

Outline

- The bag-of-visual word (BOV) and Fisher Kernel image representation
- Generic Visual Categorization
- Large Scale Image Retrieval
- Semantic Image Segmentation
- Intelligent Auto-thumbnailing
- Cross-modal Image Retrieval and Hybrid Content Generation

Generic Visual Categorization (GVC)

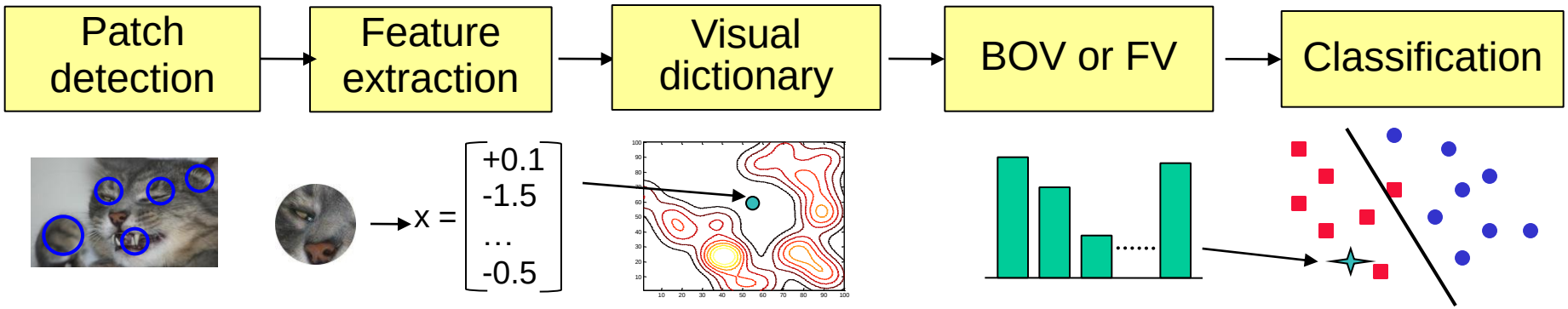


Main components:

- **Patch detection:** identify where to extract information in the image
- **Feature extraction:** compute local features
- **Visual dictionary:** map local features to visual words
- **Histogram computation:** build a high level image representation
- **Learning and Classification:**
  - learn a generative model or a discriminative classifier

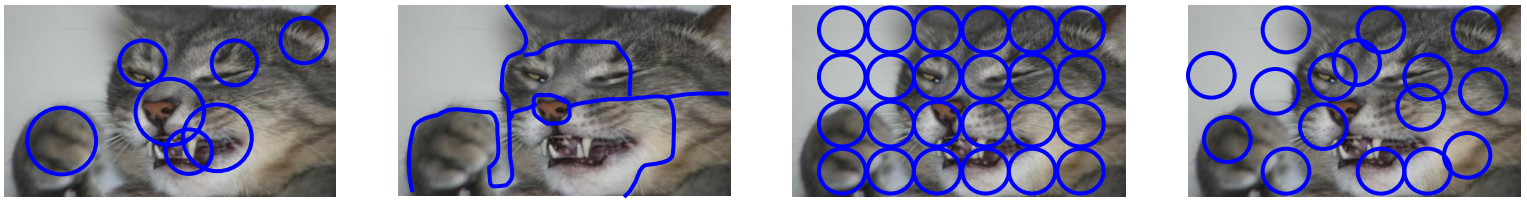
*Visual Categorization with Bags of Keypoints, Csurka et al, SLCV (ECCV Workshop) 2004

Categorization: Pipeline



Patch Detection:

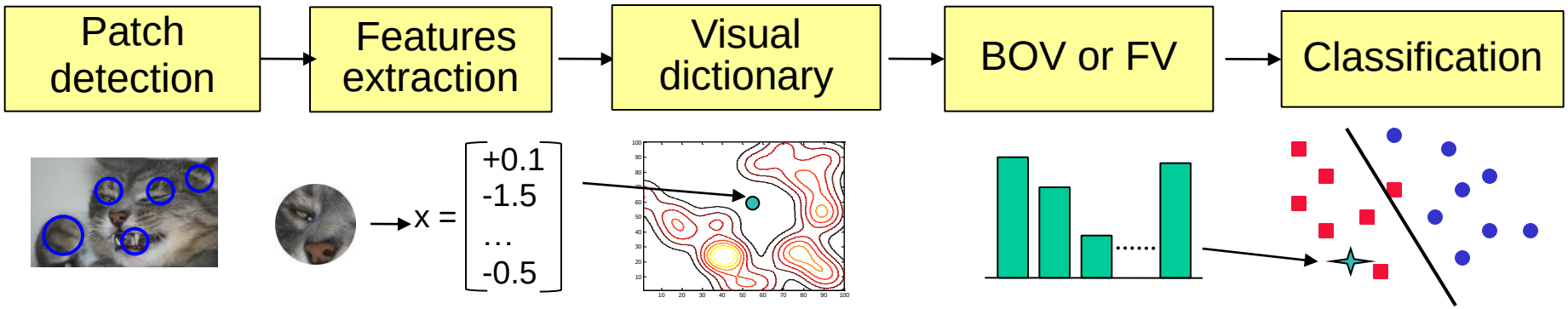
- **Interest point***: detector: Harris Laplace, MSR, …
- **Blob:** low level image segmentation
- **Regular Grid**: single or multiple scales
- **Random:** randomly selected local regions.



* *A comparison of affine region detectors* , Mikolajczyk et al, IJCV (65)1/2, 2005.

xerox

Categorization: Pipeline



Local descriptors:

- **Color**: histograms, local statistics, color moments, color correlograms, …
- **Texture***:, SIFT, cross-correlation, Gabor filters, steerable filters, differential invariants, spin images, …
- **Shape:** convexity, moment invariants , shape context
- and their combination, e.g Color SIFT…

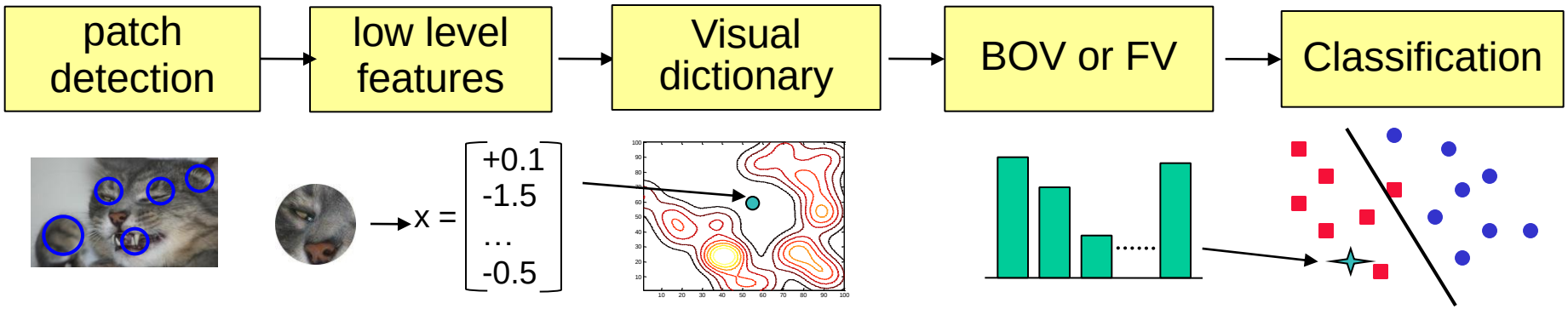
Dimension reduction with PCA to reduce computational cost and noise

---

* *A performance evaluation of local descriptors*, Mikolajczyk and Schmid, PAMI (27)10, 2005.

xerox

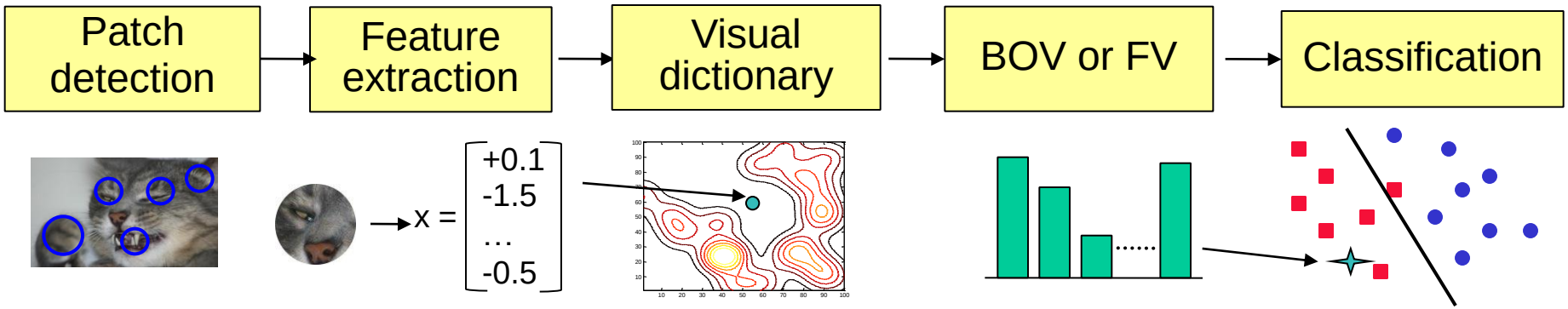
Categorization: Pipeline



Visual Vocabulary:

- **Unsupervised**
  - Kmeans, Mean Shift, GMM
- **Supervised:**
  - Concatenation of class dependent visual vocabularies, Randomized Forest, Adapted Vocabularies, …

Categorization: Pipeline



Classification:

- **generative**
  - Naïve Bayes, pLSA, LDA , KDA
- **discriminative:**
  - SVM, (K)SLR, Pyramid Match kernels, AdaBoost

Multiple features:

- **early fusion:** weighted feature level, concatenation of pyramid level
- **late fusion:** combining classifier outputs

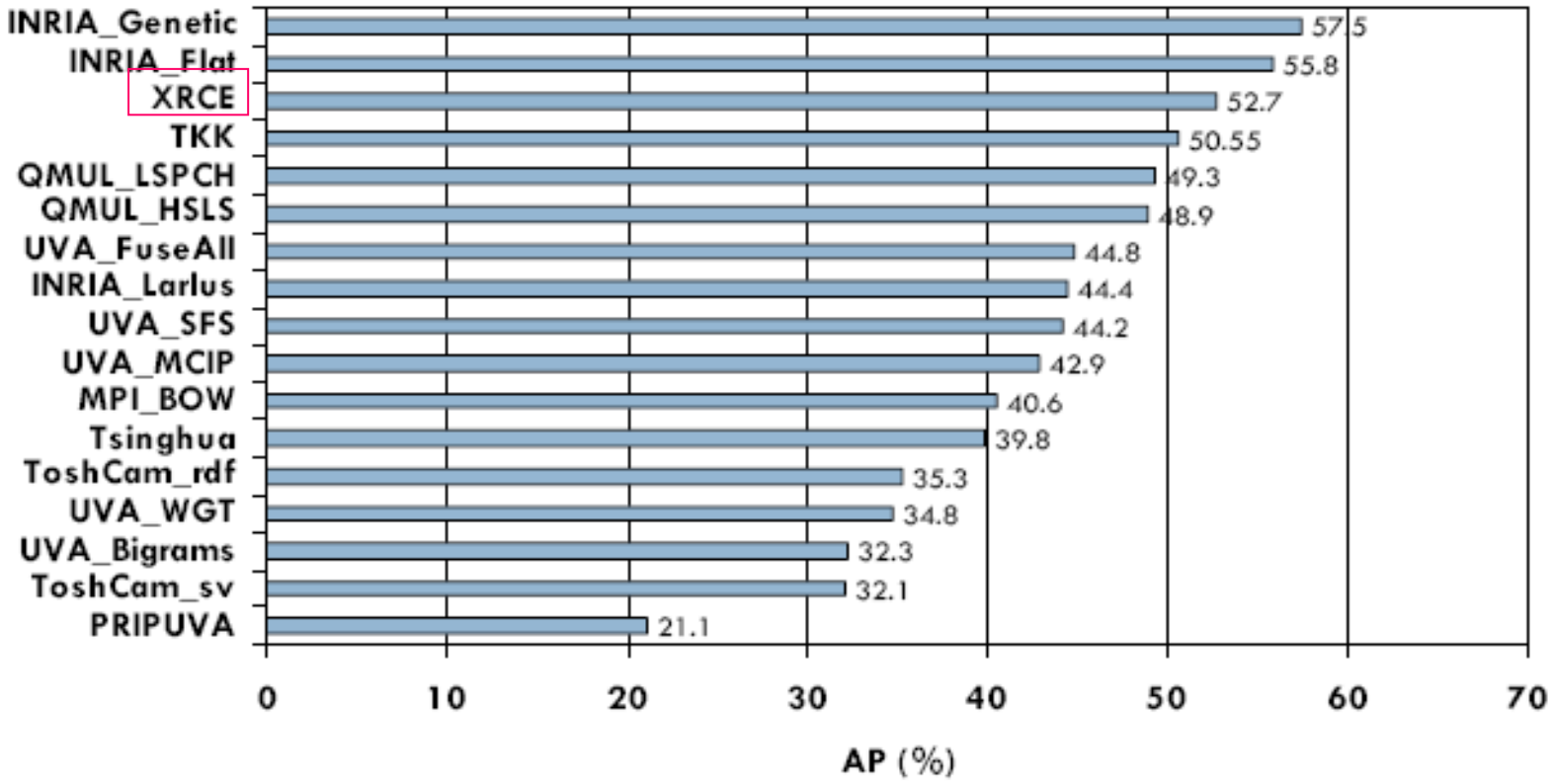
Generic Visual Categorization - experiments

- BOV versus FK:
  - Train on VOC 2008 tested on VOC 2007
  - BOV (with 2048 Gaussians) and FV with (128 Gaussians)

	BOV		FV	
	Linear	Non-linear	Linear	Non-linear
SIFT	0.35	0.44	0.40	0.47
COLOR	0.27	0.35	0.30	0.39
BOTH	0.38	0.46	0.43	0.49

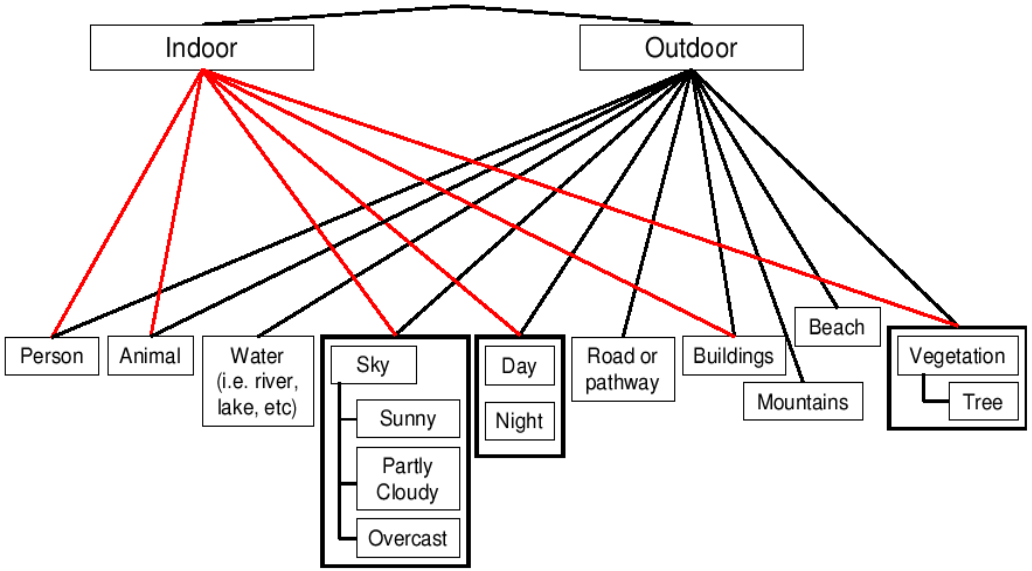
Generic Visual Categorization - experiments

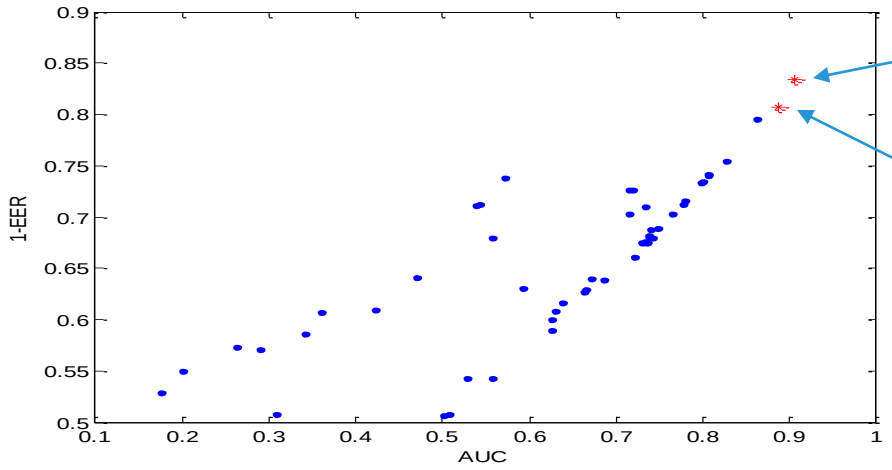
- Pascal VOC 2007 Challenge (non-linear FV with SIFT+Color)



xerox

Visual Concept Detection Task ImageClef 2008





Non-linear system

Linear system

xerox

Visual Concept Detection Task ImageClef 2009

The Main Task – annotate images with a set of visual concepts:

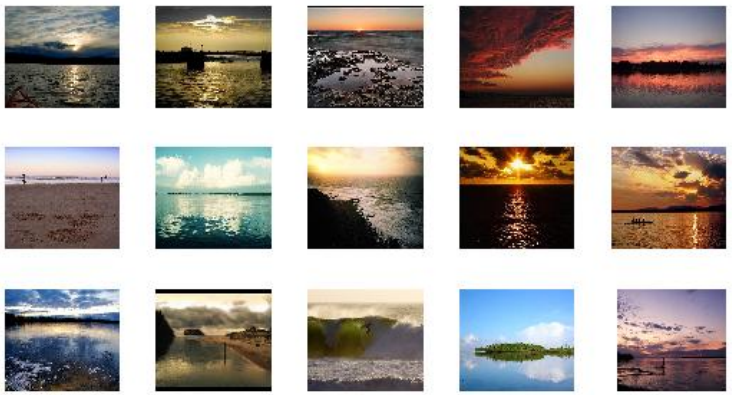
- Abstract Categories (Landscape, Family&Friends, Partylife …)
- Seasons (Summer, Spring, …)
- Time of Day (Day, Night, Sunset, …)
- Persons (no, single, big groups)
- Quality (blurred, underexposed …)
- Representation (portrait, macro image, canvas …)

- Database: 5000 training images and 13000 test images
- 53 concepts organized in an ontology (hierarchy, disjoint concepts, implications)

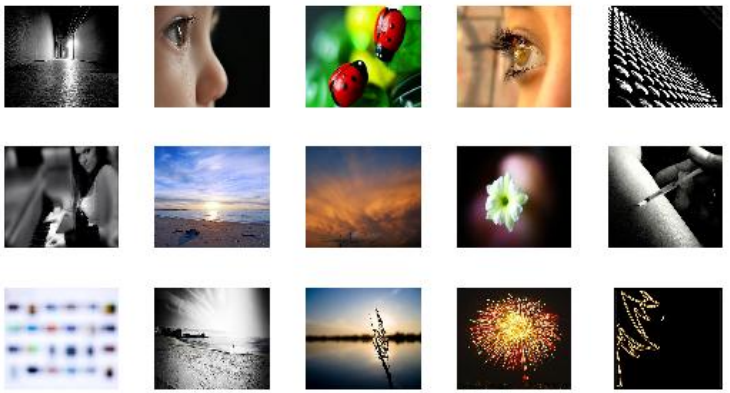
- **Evaluation is based on two measures:**
  - Evaluation per concept: EER and AUC
  - Evaluation per image : a hierarchical measure (HM) that determines the annotation performance for each image, penalizing ontology inconsistencies

xerox

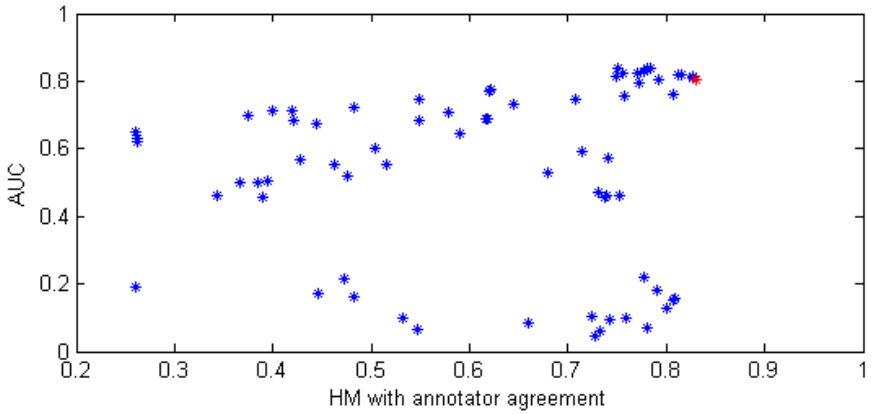
Visual Concept Detection Task ImageClef 2009

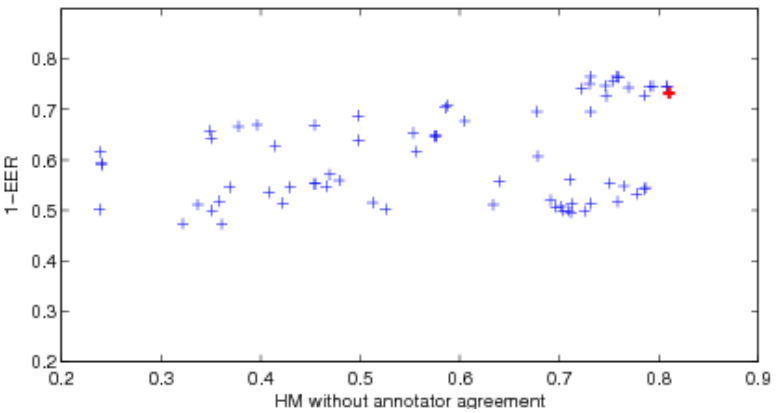


Water (Lake, River, See)



Aesthetic impression

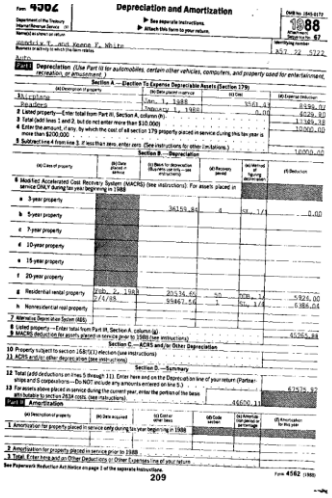


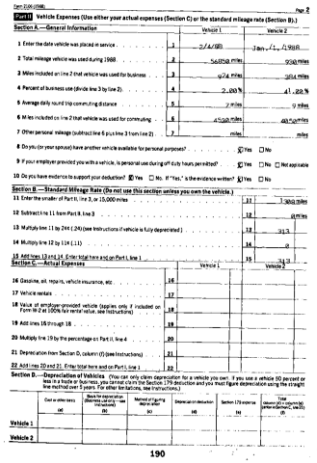


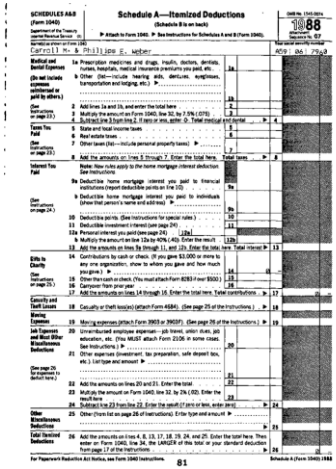
xerox

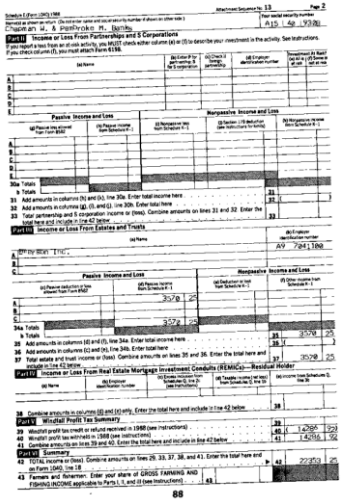
Document Image Categorization

- 20 categories.
- 5590 images.
- simulated USA tax forms.
- Validation:
  - training set of 10 images per category,
  - testing set of all the other images.
- Performance 100% (best SOA 99,82 %)









Discussion

- Its strengths:
  - Low computational cost (both at training and testing)
  - The system works pretty well with default parameter settings
  - Gives higher classification performance than BOV
  - In the case of linear classifier even lower training cost (less GMMs)
  - Its generic (we applied to different type of images).
  - While slightly below top winners in e.g. Pascal challenge, the accuracy could be further improved with more features considered (we use only 2)

- Its limitations
  - Still remains at the “bag-of-visual word” level (local information)
  - No geometry, no or limited spatial knowledge, shape or global object model

Outline

- The bag-of-visual word (BOV) and Fisher Kernel image representation
- Generic Visual Categorization
- Large Scale Image Retrieval
- Semantic Image Segmentation
- Intelligent Auto-thumbnailing
- Cross-modal Image Retrieval and Hybrid Content Generation

Large Scale Image Retrieval

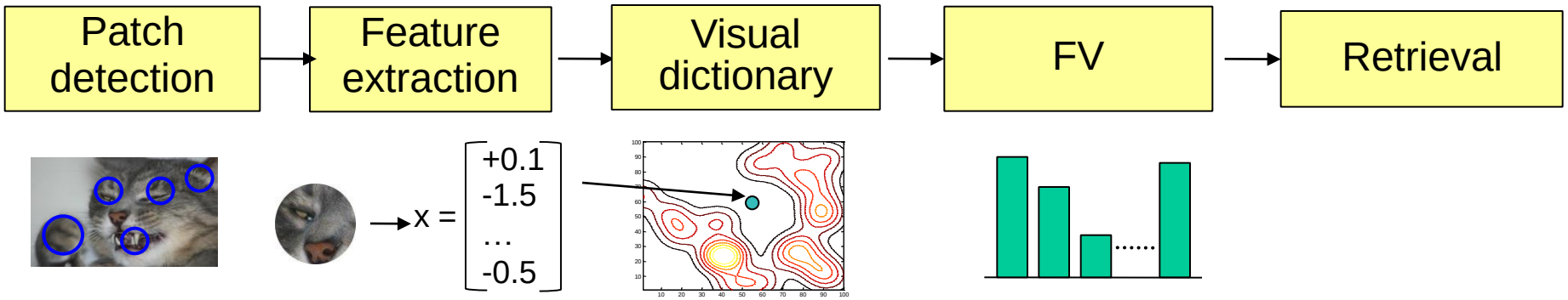


Image similarity:

- replace dot-product by a kernel which is more robust to sparse vectors

⇨ use L1 distance (Laplacian kernel)

- "unsparsify" the vector so that we can keep the dot-product

⇨ use power transformation

$$\hat{f} = \mathrm{sign}(f)\,|f|^{\alpha}$$

⇨ we use α=0.5 (Bhattacharyya kernel)

*** Large-Scale Image Retrieval with Compressed Fisher Vectors, F. Perronnin et al, CVPR 2010.**

xerox

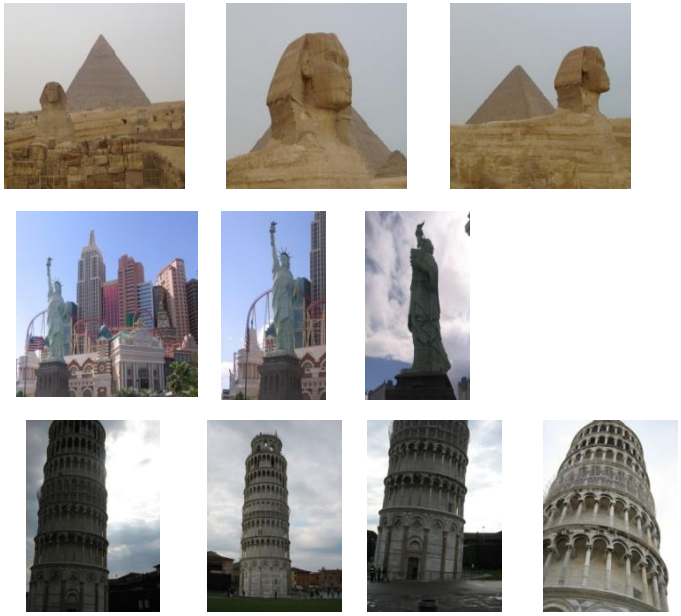
Large Scale Image Retrieval

The databases:

(a) The INRIA Holiday dataset : 1,491 images of 500 scenes / objects

(b) The UKbench dataset: 2,550 objects x 4 occurrences = 10,200 images

(a)



Challenge:

- Variations in scaling, view point, lighting, occlusion, etc.

Evaluation:

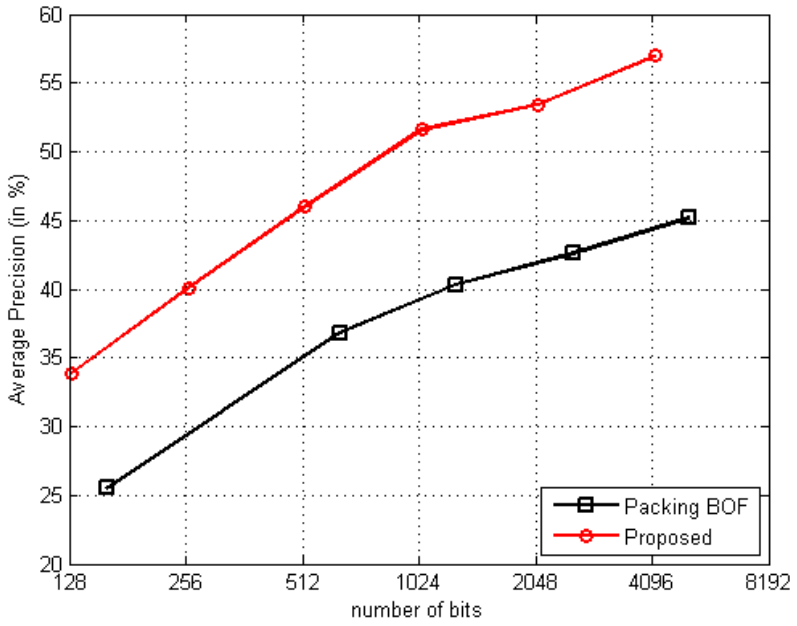
- a query with one image, retrieve images of the same scene
- Accuracy measured with Average Precision (AP)

(b)

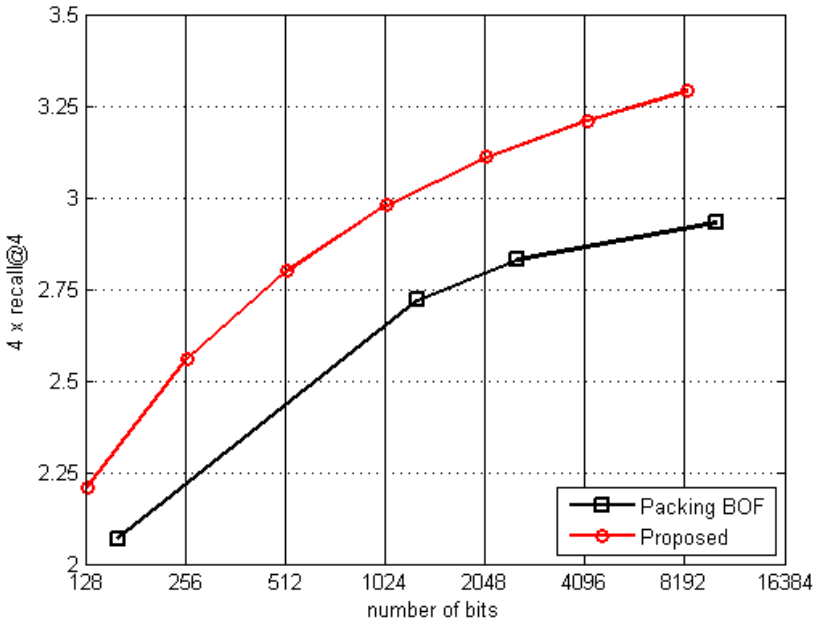
*** Large-Scale Image Retrieval with Compressed Fisher Vectors, F. Perronnin et al, CVPR 2010.**

Compressed Fisher vs compressed BOV (large scale)

Holliday dataset



UKbench dataset



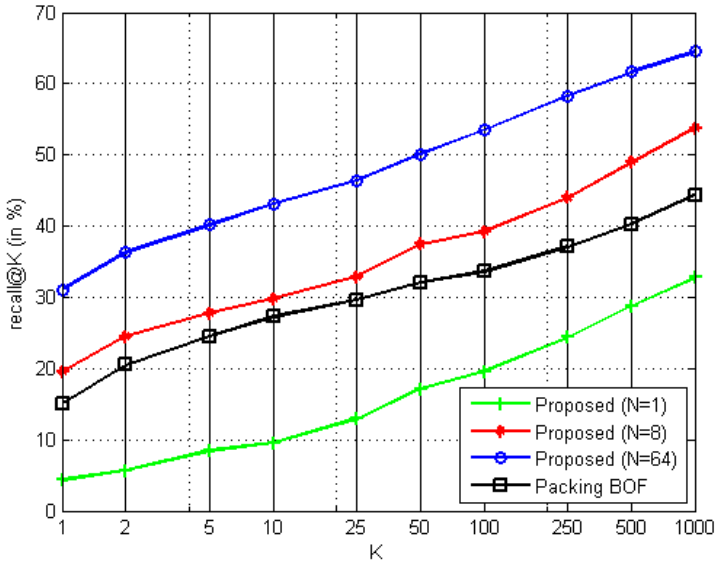
- State-of-the-art “packing BOF” : Jegou et al, CVPR 2010
- For Fisher Vector: using a simplistic binarization strategy*

*** Large-Scale Image Retrieval with Compressed Fisher Vectors, F. Perronnin et al, CVPR 2010.**

xerox

Large scale experiment

Embed Holiday dataset in
1M random Flickr images



4,160 bits
520 bits
5,120 bits
65 bits

*** Large-Scale Image Retrieval with Compressed Fisher Vectors, F. Perronnin et al, CVPR 2010.**

xerox

Outline

- The bag-of-visual word (BOV) and Fisher Kernel image representation
- Generic Visual Categorization
- Large Scale Image Retrieval
- Semantic Image Segmentation
- Intelligent Auto-thumbnailing
- Cross-modal Image Retrieval and Hybrid Content Generation

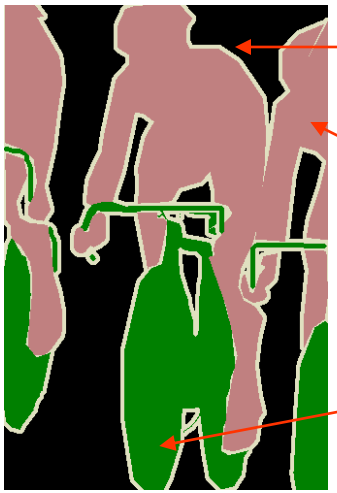
Low Level versus High Level image Segmentation

Low level segmentation (LLS)

- i.e. clustering
- mostly used as pre-processing in divers application
- ill-posed (no ground-truth)

High level (Semantic) segmentation (HLS)

- i.e. classification
- useful directly in divers application
- defined by the targeted semantic class set



Object Segmentation versus Image partition

Object (class) segmentation

- two-class problem, (e.g. horse vs. not-horse)
- easier, equivalent to mono-labeled data
- in many application no need to be of “pixel precision”

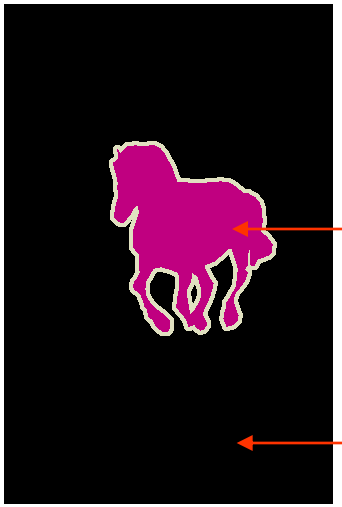
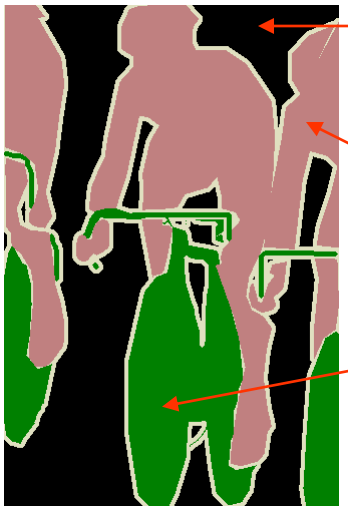


Image “partition”

- multi-class problem,
- much harder,
- confusing classes,
- pixel precision more important



xerox

High Level image Segmentation -SOA

- **Random field model (MRF,CRF) integrating low and high level cues:**
  - LOCUS (Winn and Jojic, ICCV05)
  - shape and context within CRF (He et al CVPR04),
  - OBJ CUT (Kumar et al CVPR05),
  - Textonboost (Shotton et al ECCV06, CVPR06),
  - CRF dealing with partially labeled images (Verbeek and Triggs, NIPS07),
  - Object BOV integrated with random fields (Larlus et al CVPR08, IJCV09),
  - Latent Topic Random Fields (He and Zemel, CVPR08),
  - LLS based Higher Order Potentials within MRF (Kohli et al, CVPR 08)
  - Hierarchical CRF (Kumar and Hebert, ICCV05, Gonfaus et al CVPR10),

High Level image Segmentation –SOA

- **Combination of LL segmentation with HL representations**
  - fragment-based approaches (Borenstein et al CVPR04)
  - BOF of mean-shift segments (Yang, et al CVPR07)
  - Region level LDA with topic enforcement (Cao and Fei-Fei, ICCV07)
  - multiple figure-ground segmentations with segmentation quality ranking (Carreira and Sminchisescu CVPR10)

Random Forest

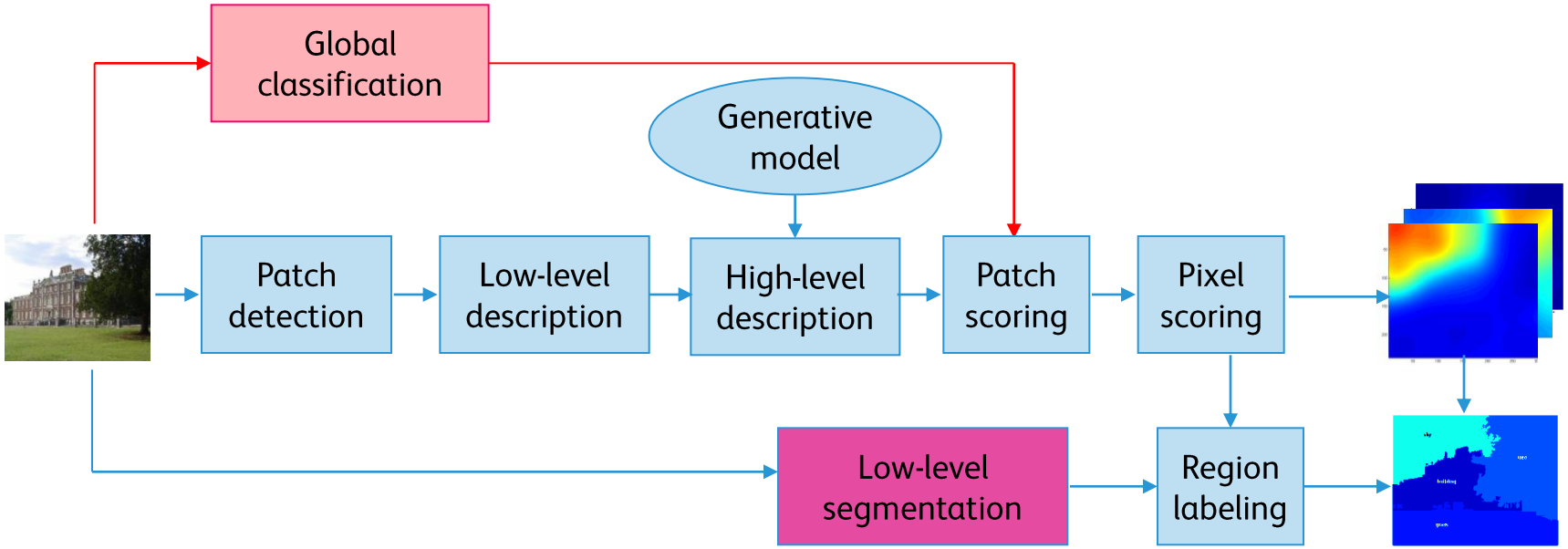
- semantic Texton Forests (Shotton et al CVPR08)
- local features random Forest and NN models (Schroff et al BMVC08)
- multiple output randomized trees (Dumont et al, VISAPP09)

Other:

- contextual emprirical Bayes(Lazebnik and Raginsky, CVPR09)

The main idea*

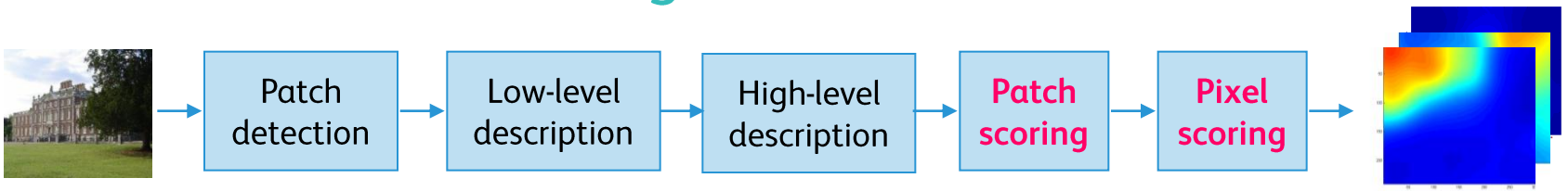
- Fisher kernel representation
- Patch level classification
- Class probability maps combined with LL segmentation
- Image level prior as fast rejection



* *A Simple High Performance Approach to Semantic Segmentation*, G. Csurka and F. Perronnin, BMVC 2008.

xerox

Patch and Pixel Scoring



- Patch classifiers (PC) were:
  - trained on labeled Fisher Vectors (using masks and bounding boxes)
  - Linear Sparse Logistic Regression scores transformed in probabilities:

$$p(c \mid f_t) = \frac{1}{1+\exp(-\alpha^T f_t + \beta)}$$

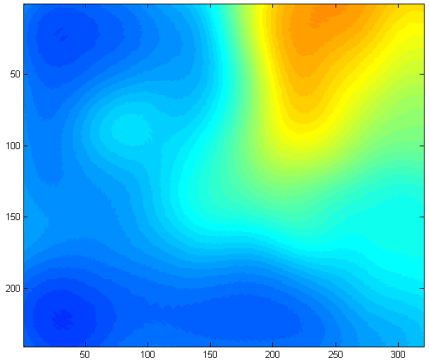
- The class posterior at pixel level is the weighted average of the class posteriors of patches containing the pixel.

$$p(c \mid z) = \frac{\sum_t p(c \mid f_t)\mathcal{N}(z \mid m_t, C_t)}{\sum_t \mathcal{N}(z \mid m_t, C_t)}$$

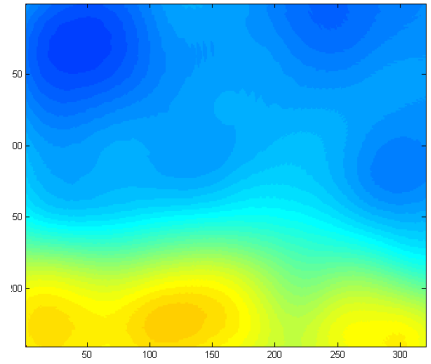
- This leads to one class probability map (P_c) per class.

*** *A Simple High Performance Approach to Semantic Segmentation*, G. Csurka and F. Perronnin, BMVC 2008.**

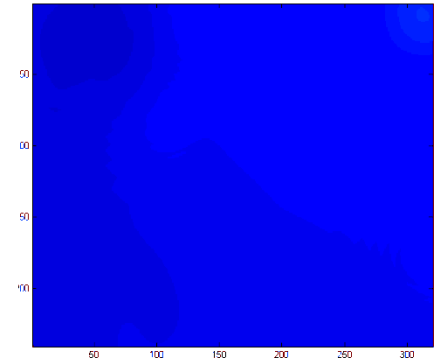
Examples of class probability maps



Tree Map

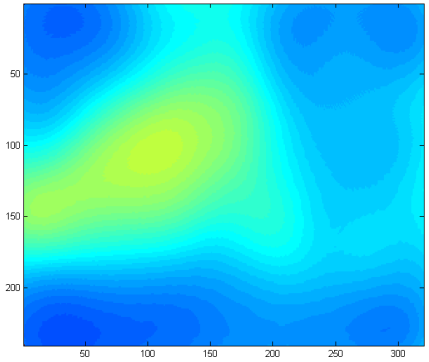


Grass Map

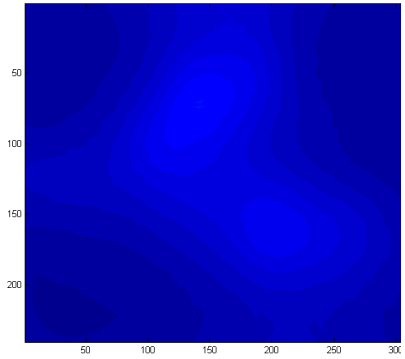


Dog Map

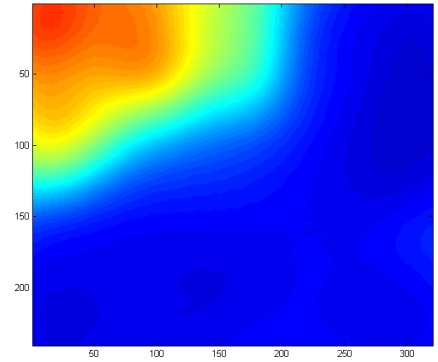
Pixel labeling



Building Map



Boat Map

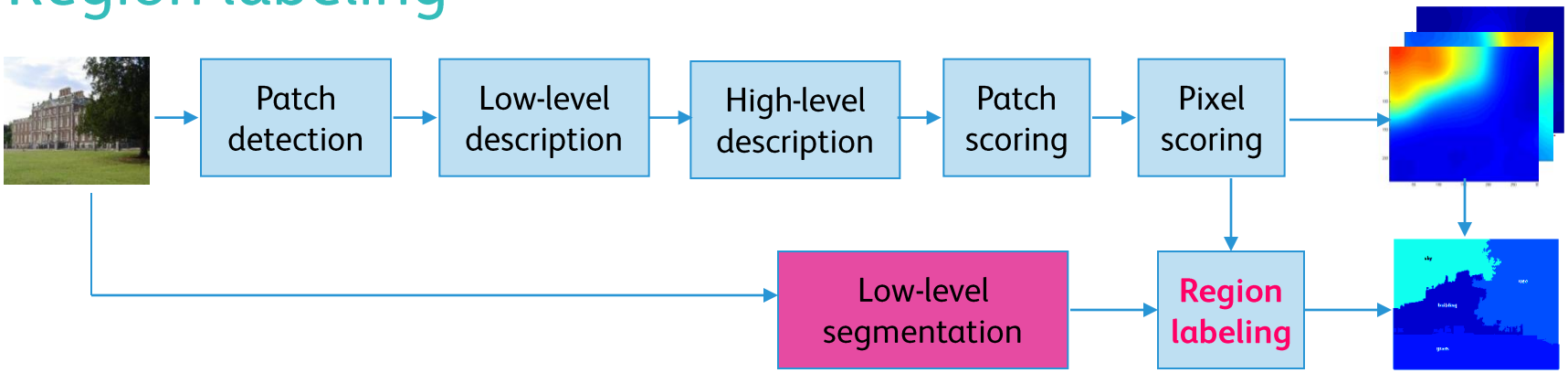


Sky Map

* *A Simple High Performance Approach to Semantic Segmentation*, G. Csurka and F. Perronnin, BMVC 2008.

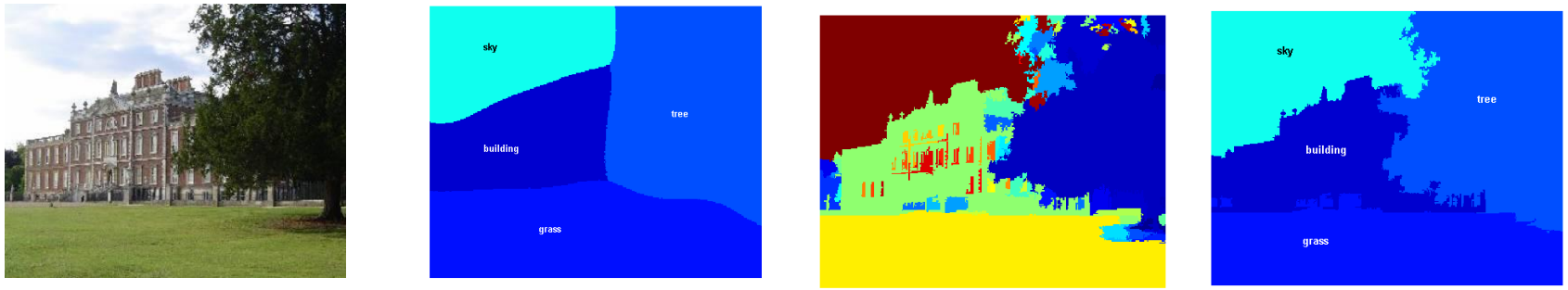
xerox

Region labeling



- Class probabilities are averaged over low level (Mean Shift) images segments and each segment R is labeled with:

$$c^* = \begin{cases} \arg\max_c \left(P_c(R)\right) & \text{if} \quad P_c(R) > \text{Thr} \\ \text{background} \end{cases}$$

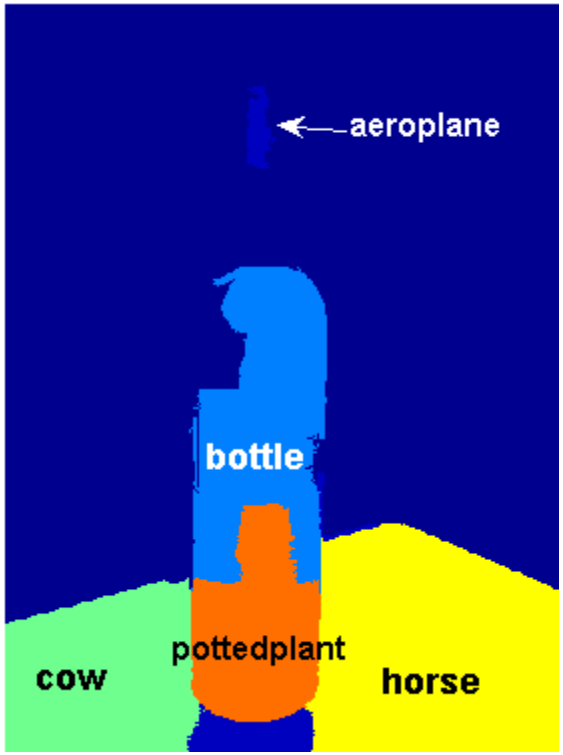


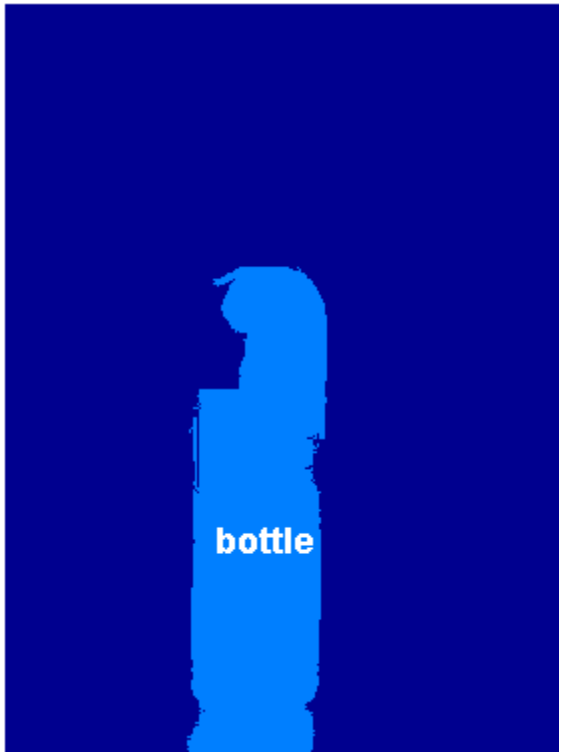
* *A Simple High Performance Approach to Semantic Segmentation*, G. Csurka and F. Perronnin, BMVC 2008.

xerox

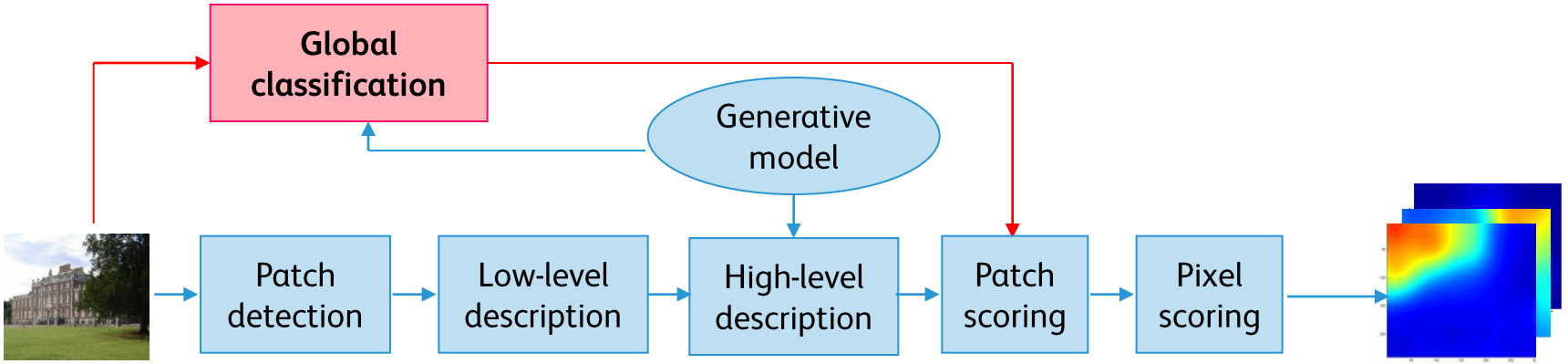
However

- Using all probability masks might introduce many local False Positives !!





Fast Rejection with Global Classification



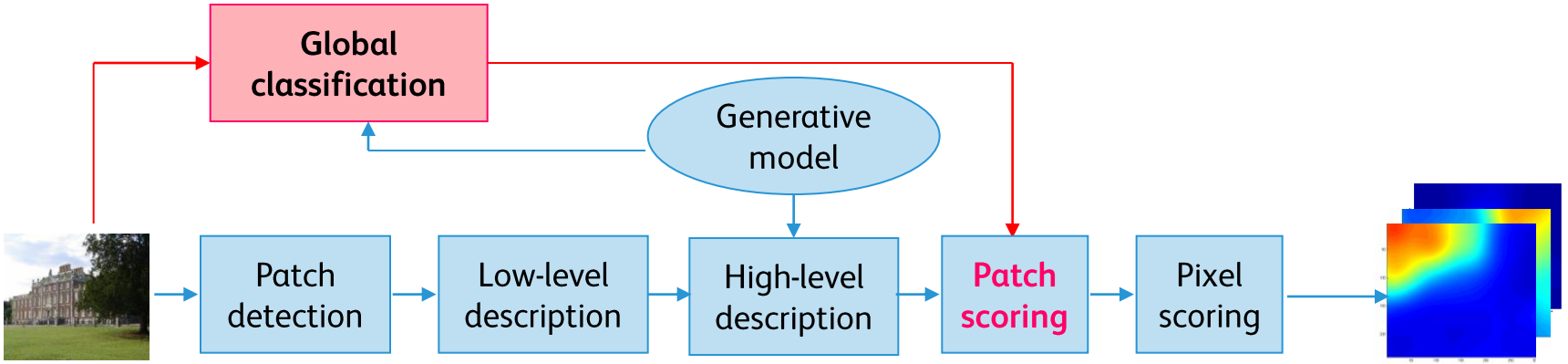
- A visual categorizer is trained on weakly labeled data to detect visual concepts/objects (any classifier can be used) and transform scores in probabilities (image level prior).
- Then image level prior (ILP) is used to fast reject "non relevant" probability maps :
- ☺ Reduce computational cost.
- ☺ Decrease false positive regions.
- ☹ Prevent the discovery of objects rejected by the global classifier.

---

* *A Simple High Performance Approach to Semantic Segmentation*, G. Csurka and F. Perronnin, BMVC 2008.

xerox

Modified Patch Classifier (MPC)

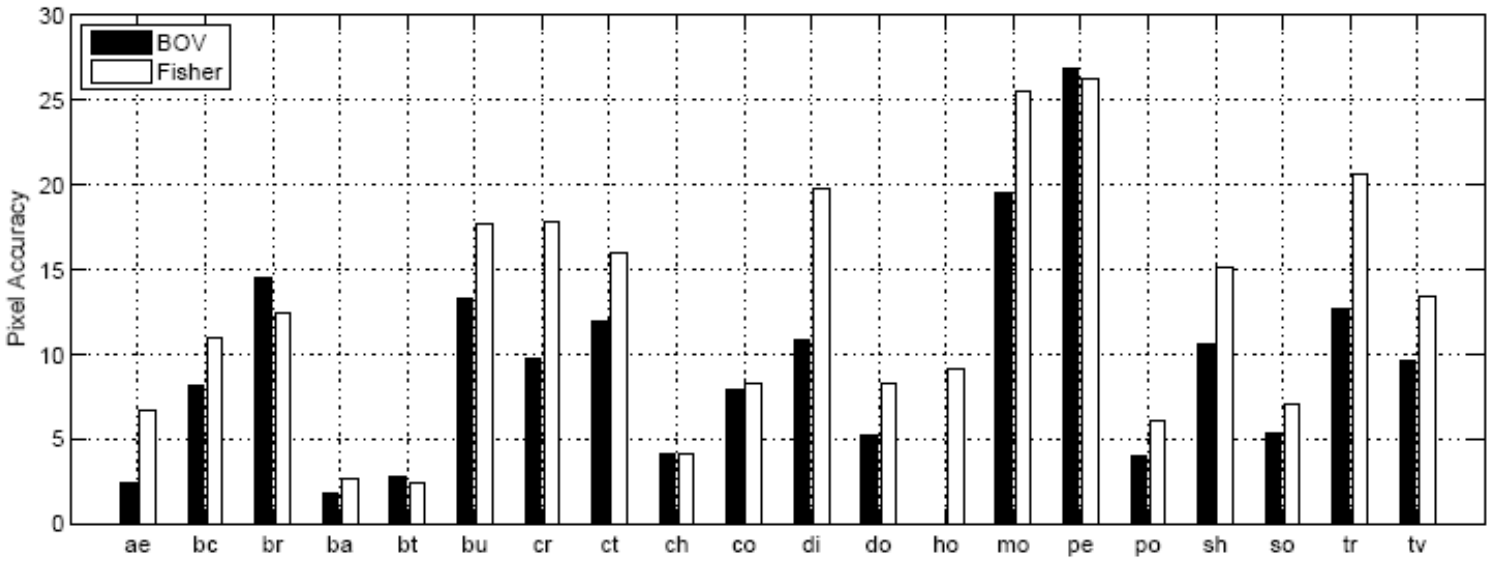


- Main idea:
  - Global image classifier rejects the improbable context/background.
  - Patch classifier separates the "object" from its usual context.
- How:
  - Train the patch classifier only with images containing the object:
    - positive patches from object masks (segments and bounding boxes)
    - negative patches from the inversed masks

* *A Simple High Performance Approach to Semantic Segmentation*, G. Csurka and F. Perronnin, BMVC 2008.

xerox

Results – BOV vs FV (on Pascal 2007)

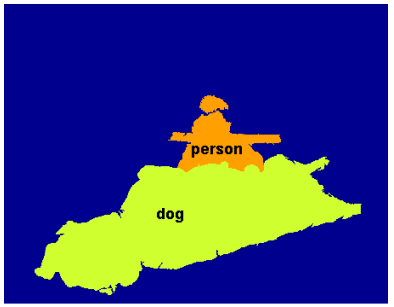


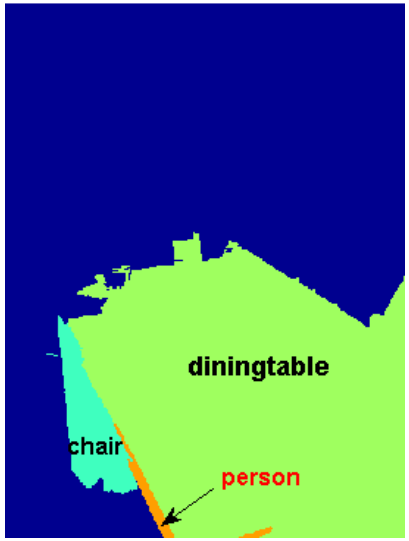
	BOV		FV	
	Patch level	Mask level	Patch level	Mask level
No Gr	11.6	10.3	15.0	15.9
GR 1	19.3	11.9	21.0	18.8
GR 2	24.2	15.9	25.8	24.0

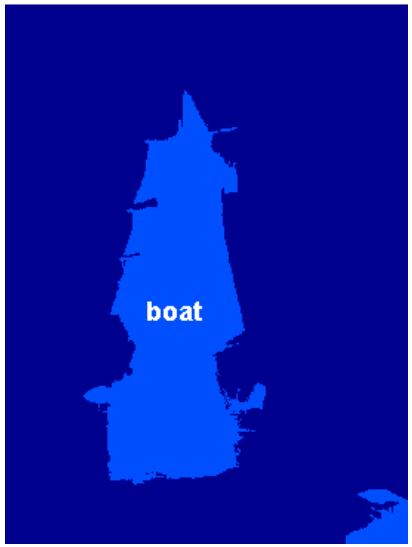
* ***A Simple High Performance Approach to Semantic Segmentation***, **G. Csurka and F. Perronnin, BMVC 2008.**

xerox

Results – winner of Pascal 2008

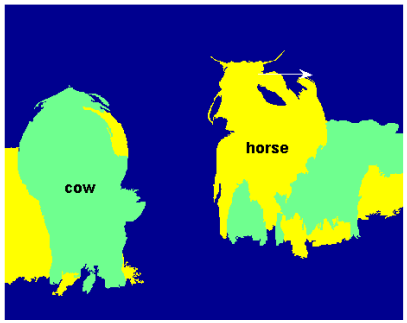


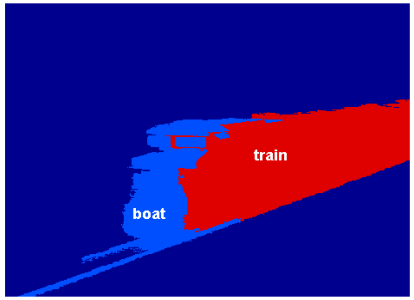




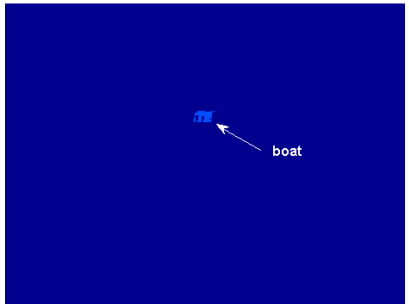
Examples where it “had difficulties”

- Confused classes



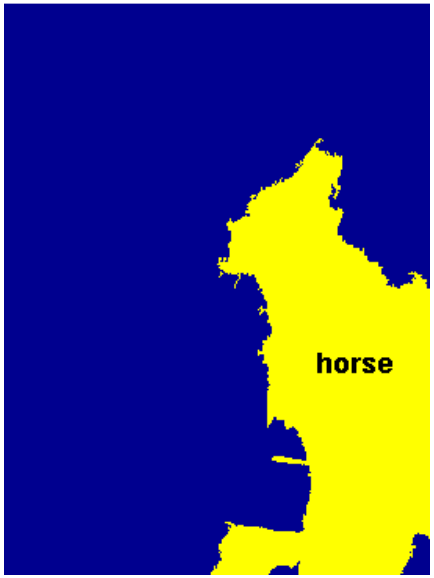


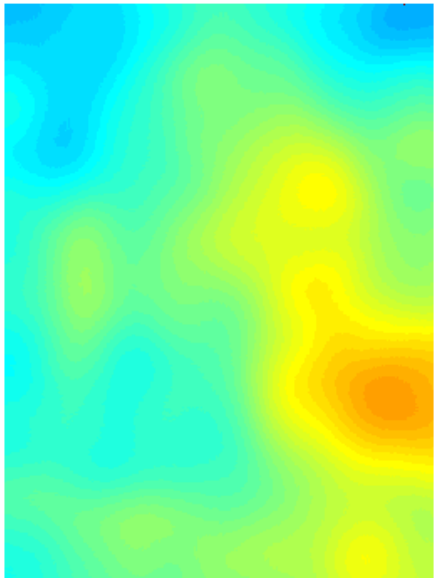
- Under and over estimation (too low or too high probability value in P_c)



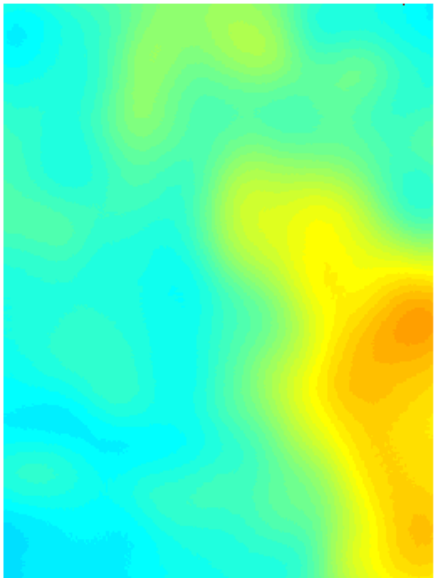
xerox

Examples where it failed (due to fast rejection ???)

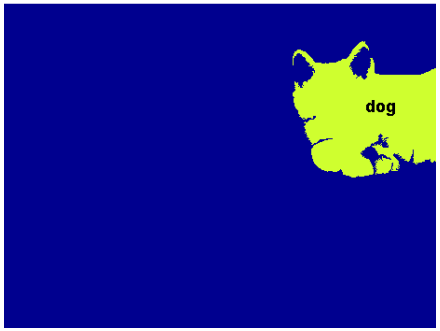


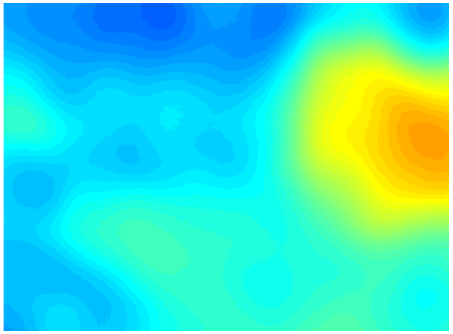


Horse Map

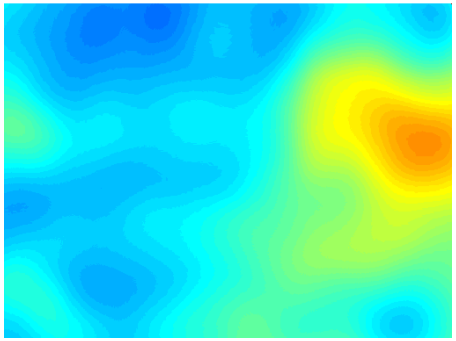


Cat Map – Not considered





Dog Map



Cat Map – Not considered

xerox

Discussion

- Its strengths:
  - Simplicity
    - Simple patch classification with high level descriptors
    - Combined with Low level segmentation and ILP
  - Low computational cost:
    - The most costly bit (Mean Shift segmentation 30 s vs 1-2 s for the rest) can be avoided for many applications (where no need for accurate object boundaries).
  - Can be a good starting point for further processing or integration in a more complex system

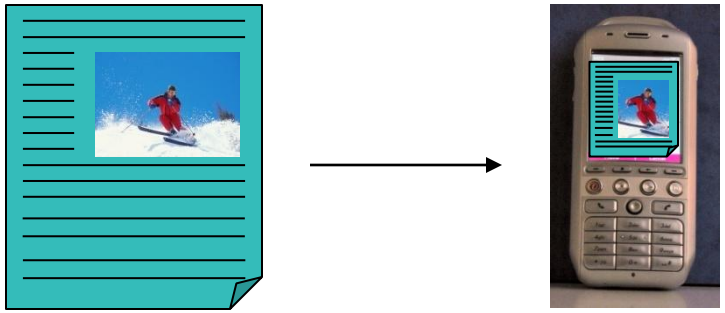
- Its limitations
  - The method is maybe too simple to give excellent results:
    - Still remains at the “bag-of-visual word” level.
    - No geometry or spatial knowledge, no shape or global object model.
    - Not suitable for object detection

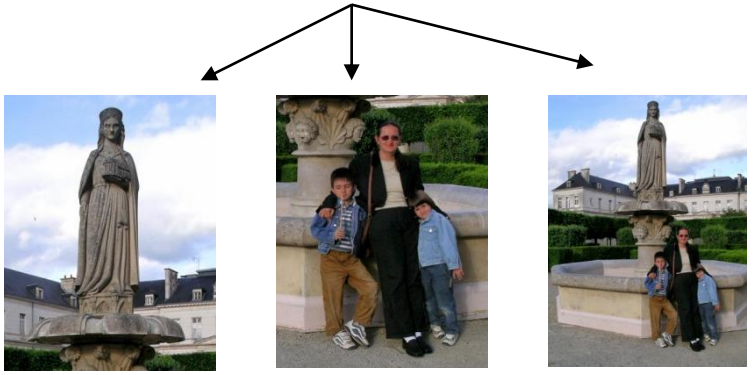
Outline

- The bag-of-visual word (BOV) and Fisher Kernel image representation
- Generic Visual Categorization
- Large Scale Image Retrieval
- Semantic Image Segmentation
- Intelligent Auto-thumbnailing
- Cross-modal Image Retrieval and Hybrid Content Generation

Intelligent Thumbnail: Problem statement

- Identification of the region(s) of interest
- It might be context dependent
- It might be content related (according to some tasks such as retrieval)
- Auto-crop or create thumbnail





State-of-the art (Saliency detection)

Bottom up:

- center-surround operation (Itti et al, PAMI 1998, Gao et al NIPS07)
- graph based activation maps (Harel et al NIPS07)
- residual of images in the spectral domain (Hou and Zhang , CVPR 07)

Top down:

- object detection (list too long)
- subject content-based intelligent cropping (Luo ICME07)
- self-adaptive image cropping (Ciocca et al ICCE07)

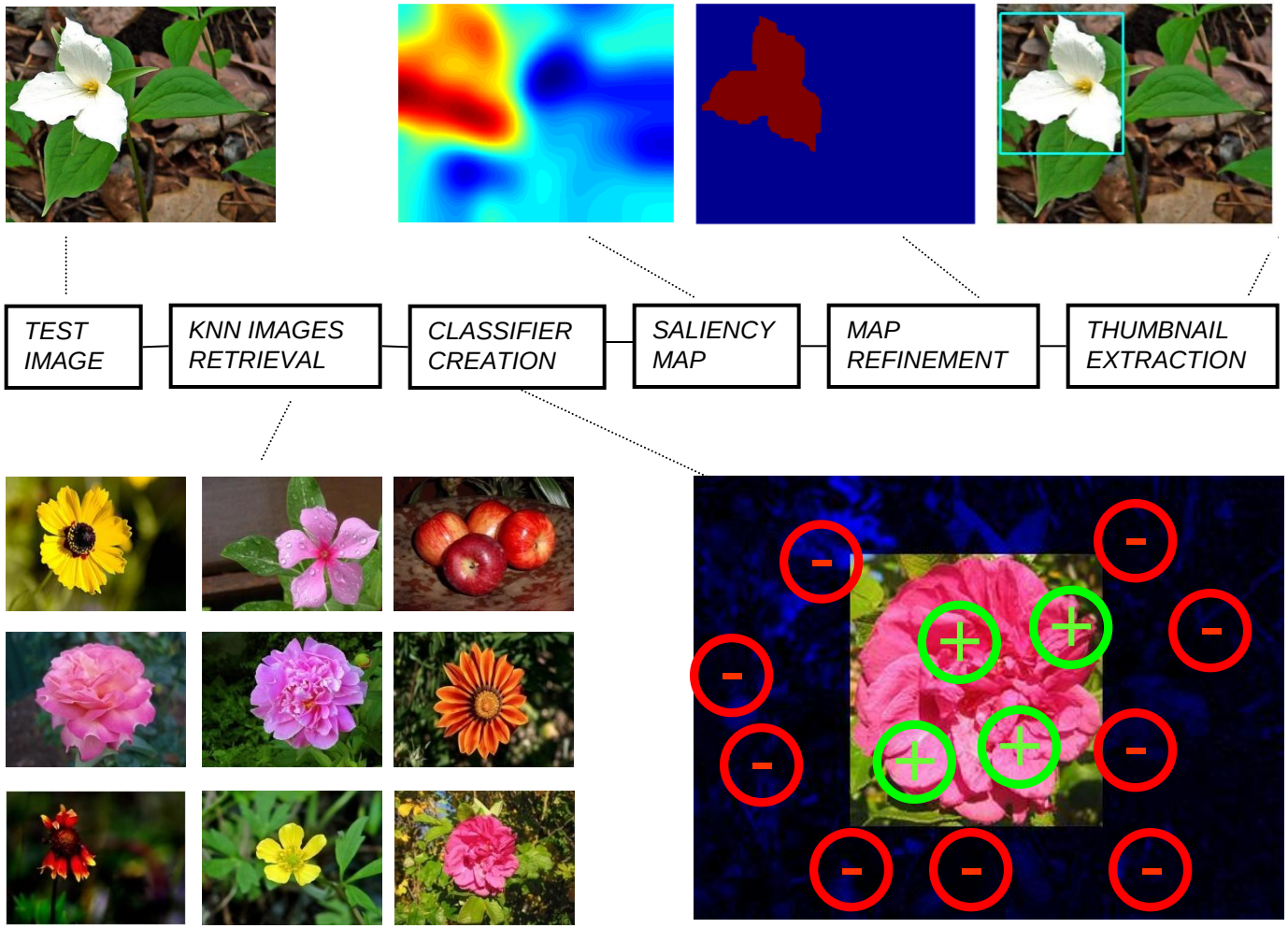
Hybrid:

- bottom-up combined with face detection (Itti and Koch 2001, Suh et al 2003, Chen et al 2003)
- learning from labeled examples (Liu et al *CVPR*, 2007)

Intelligent rescaling and re-targeting

- Setlur et al 2005, Avidan and Shamir, 2007, Simakov et al CVPR 2008

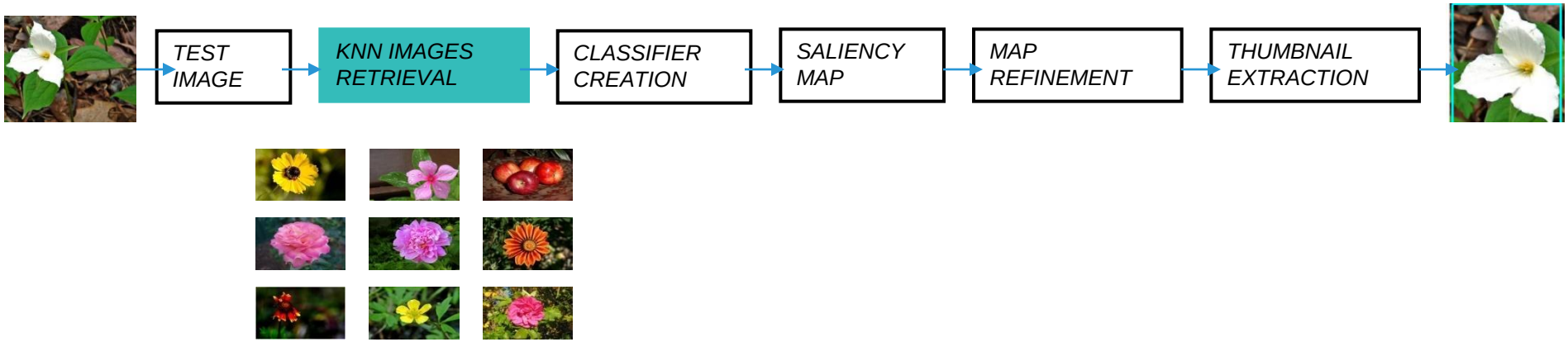
Our Method*



* A framework for visual saliency detection with applications to image Thumbnailing, Marchesotti et al ICCV 2009.

xerox

Retrieve nearest neighbors



- As similarity between images, we used the L1-norm between the normalized Fisher Vectors :

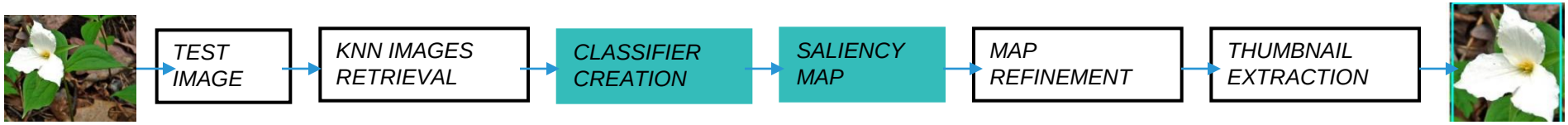
$$sim(I,J) = sim(f_I, f_J) = norm_{\max} - \| f'_I - f'_J \|_{L_1} = norm_{\max} - \sum_l | f'_I(l) - f'_J(l) |$$

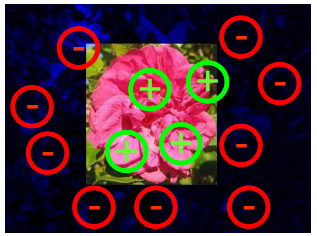
where **f'** is obtained from **f** by normalizing it to L1-norm 1.

Note: The Fisher vectors obtained for color and texture features are first concatenated to obtain $\mathbf{f}_I$.

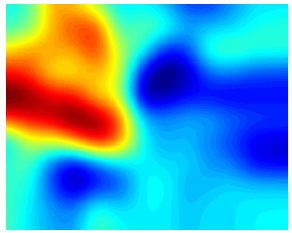
*** A framework for visual saliency detection with applications to image Thumbnailing, Marchesotti et al ICCV 2009.**

Learn saliency from the K images





- **Learning Patch classifier***
  - as for segmentation building Probability Maps
    - Saliency map if no labels using the k nearest neighbors as training data
    - Class Probability Maps from the k nearest neighbors labeled with the given class
  - drawback:
    - online learning of a classifier trained for each test image



- **Proposed alternative**

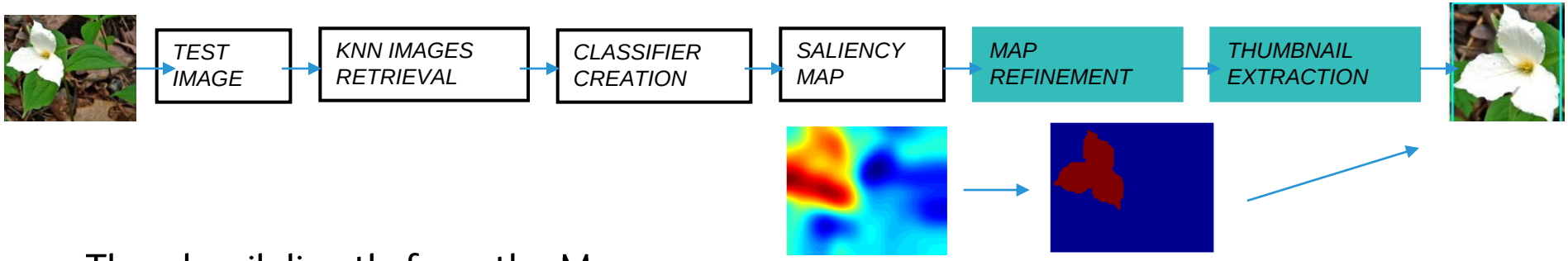
$$patch\_score(x_t) = \| f_{x_t} - \frac{1}{K}\sum_k f^+_{I_k} \|_{L_1} - \| f_{x_t} - \frac{1}{K}\sum_k f^-_{I_k} \|_{L_1}$$

- where f_x is the Fisher Vector of a patch and f_I^+ and f_I^- are Fisher Vectors of the positive and negative image regions.

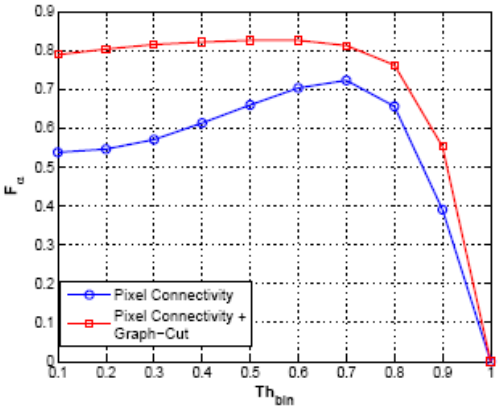
---

* ***A Simple High Performance Approach to Semantic Segmentation*, G. Csurka and F. Perronnin, BMVC 2008.**

Refine map and build thumbnail



- Thumbnail directly from the Map:
  - threshold the Map and consider the bounding box of the biggest connected component
- Refinement:
  - use Grabcut* initialized with the Probability Map to refine the object's borders

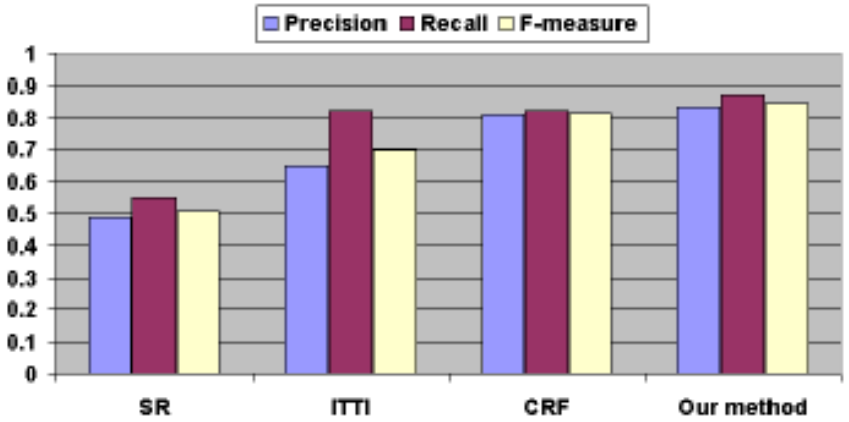


*Grabcut: Interactive foreground extraction using iterated graph cuts. C. Rother et al *SIGRAPH*04,

xerox

Results

- MRSA dataset:
  - 5000 labeled images
  - aggregated annotations from 9 annotators
- Leave-one-out strategy
- Compared with bottom-up methods:
  - Spectral residual (SR) of Hou and Zhang , CVPR07
  - Itti and Kock (VR 2000)
- Learning based method
  - CRF (Liu et al *CVPR*, 2007)

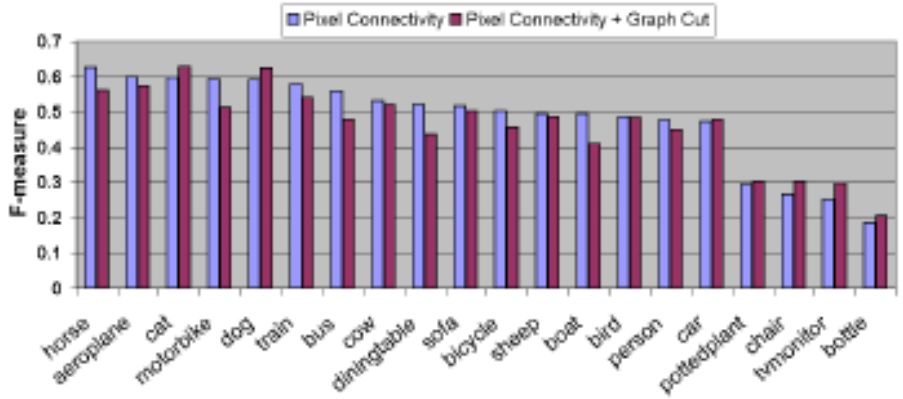


*** A framework for visual saliency detection with applications to image Thumbnailing, Marchesotti et al ICCV 2009.**

xerox

Results – labeled –target-driven

- Pascal dataset used as illustration
  - one class each time
- Not compared with detection or segmentation results as :
  - not the same goal, here easier
    - we assume that the presence is known
    - thumbnail containing one or several instances of the class is an acceptable result
    - in the experiments we measured the overlap between the biggest connected components



*** A framework for visual saliency detection with applications to image Thumbnailing, Marchesotti et al ICCV 2009.**

xerox

Potential application

Display on small devices

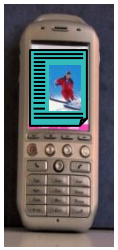


Image asset visualisation

Variable data printing

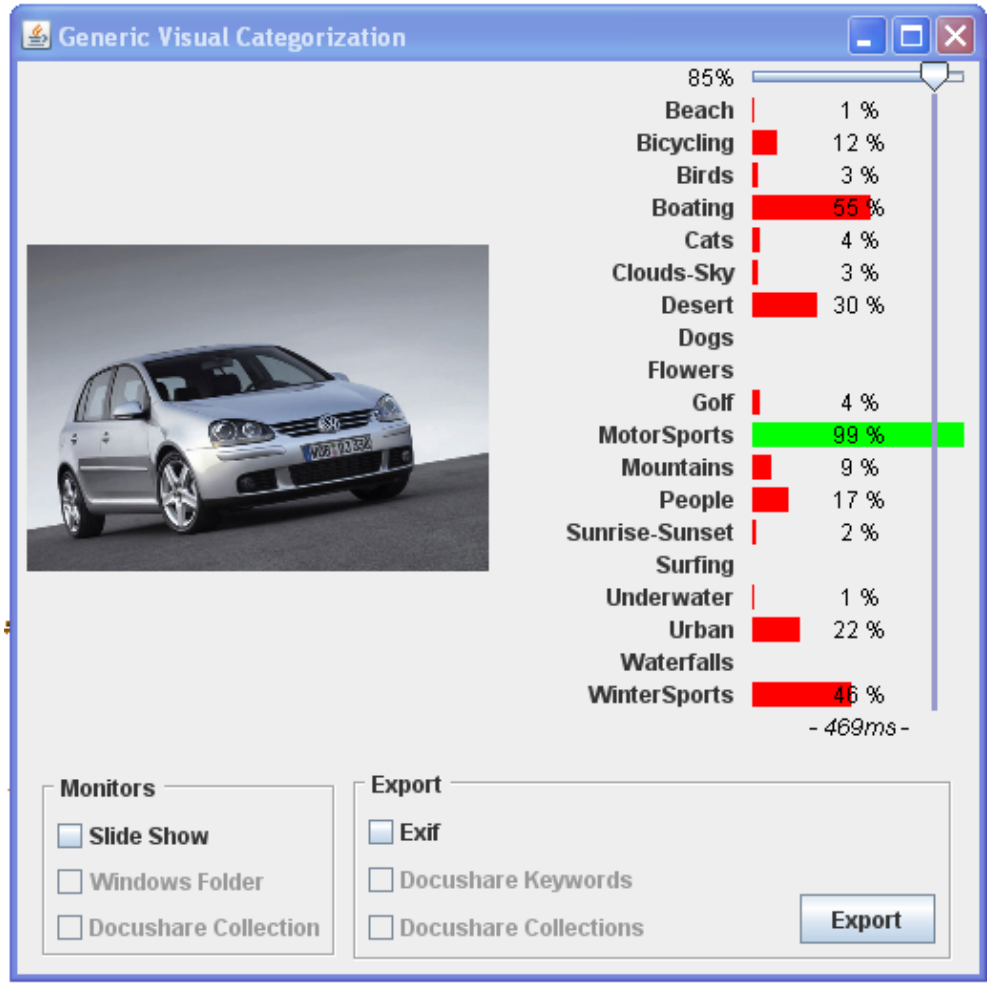
Album/video summarization

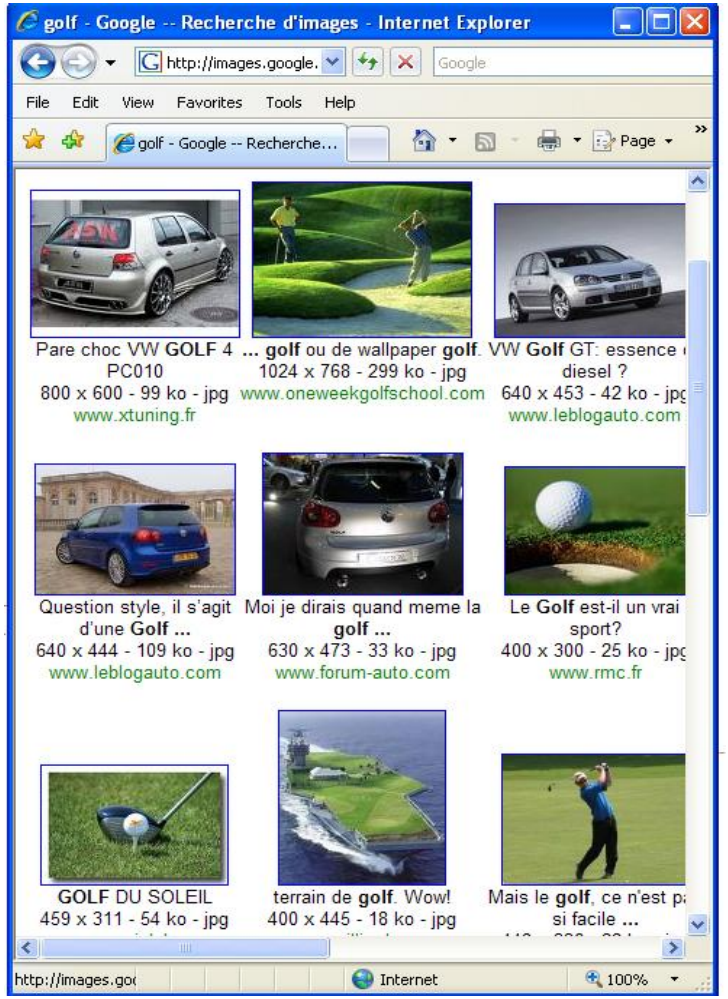
Outline

- The bag-of-visual word (BOV) and Fisher Kernel image representation
- Generic Visual Categorization
- Large Scale Image Retrieval
- Semantic Image Segmentation
- Intelligent Auto-thumbnailing
- Cross-modal Image Retrieval and Hybrid Content Generation

Motivation:

- Limitations of mono-modal systems





Motivation

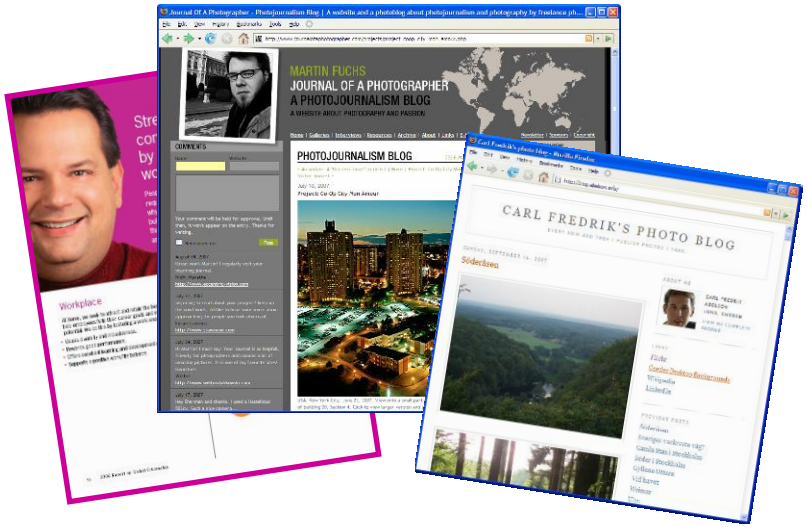
- Text and image content are often complementary

Paço Imperial, 18th century palace that served as seat for the colonial government, King John IV of Portugal and the two Emperors of Brazil.

The Third of May 1808: The Execution of the Defenders of Madrid. 5,000 Spanish civilians were executed by Napoleon's troops in the days following a popular uprising in Madrid.

The two camps have intensified campaigning with days to go

Hybrid Information Access: The Scientific Challenge



Hybrid Content analyses enables

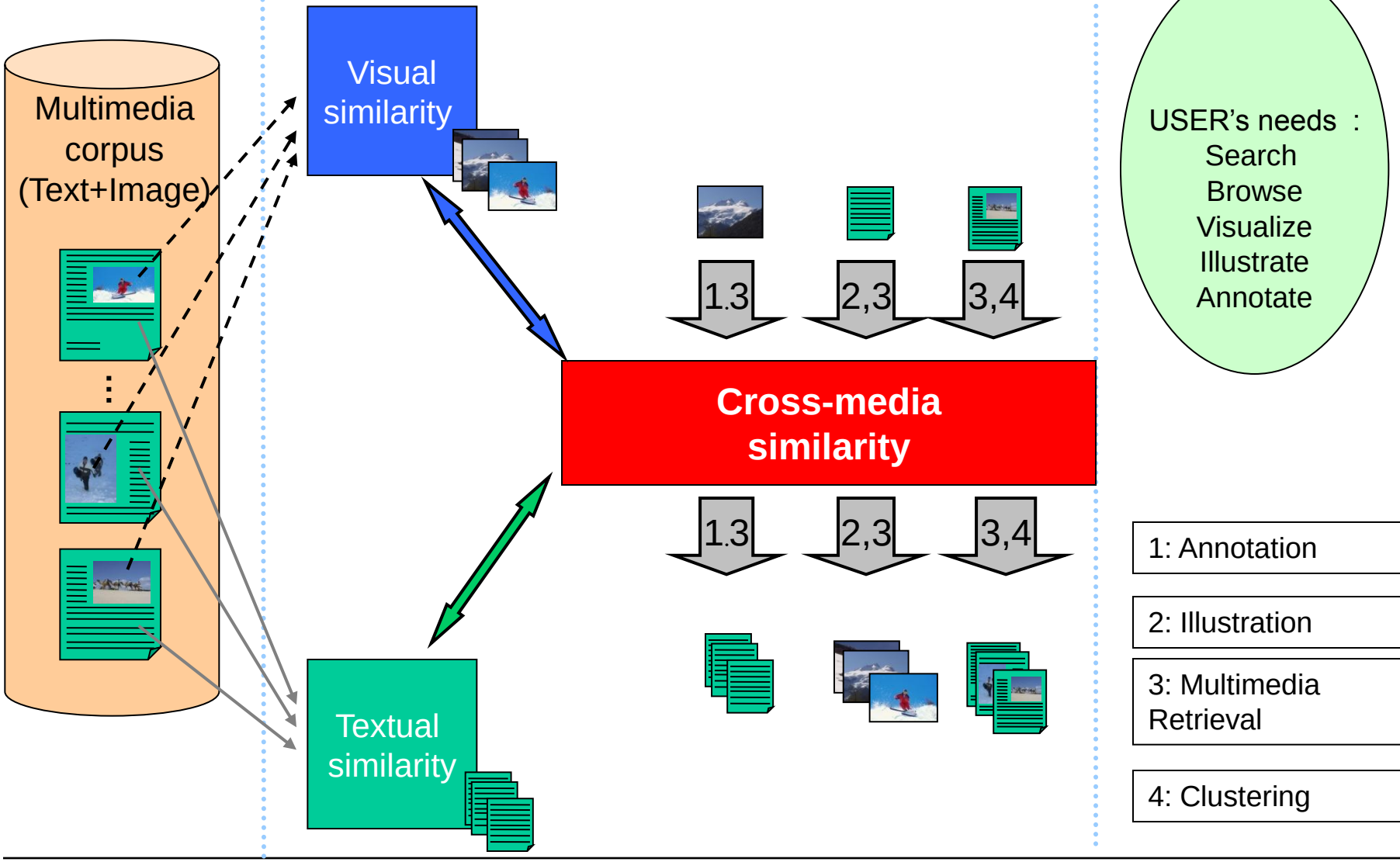
Knowledge base

Internet
Wikipedia
Social network
Proprietary

Labelling
Retrieval
Routing
Enrichment

xerox

Cross-media similarity



*Crossing textual and visual content in different application scenarios, Ah Pin et al, MTPA (42)1, March 2009

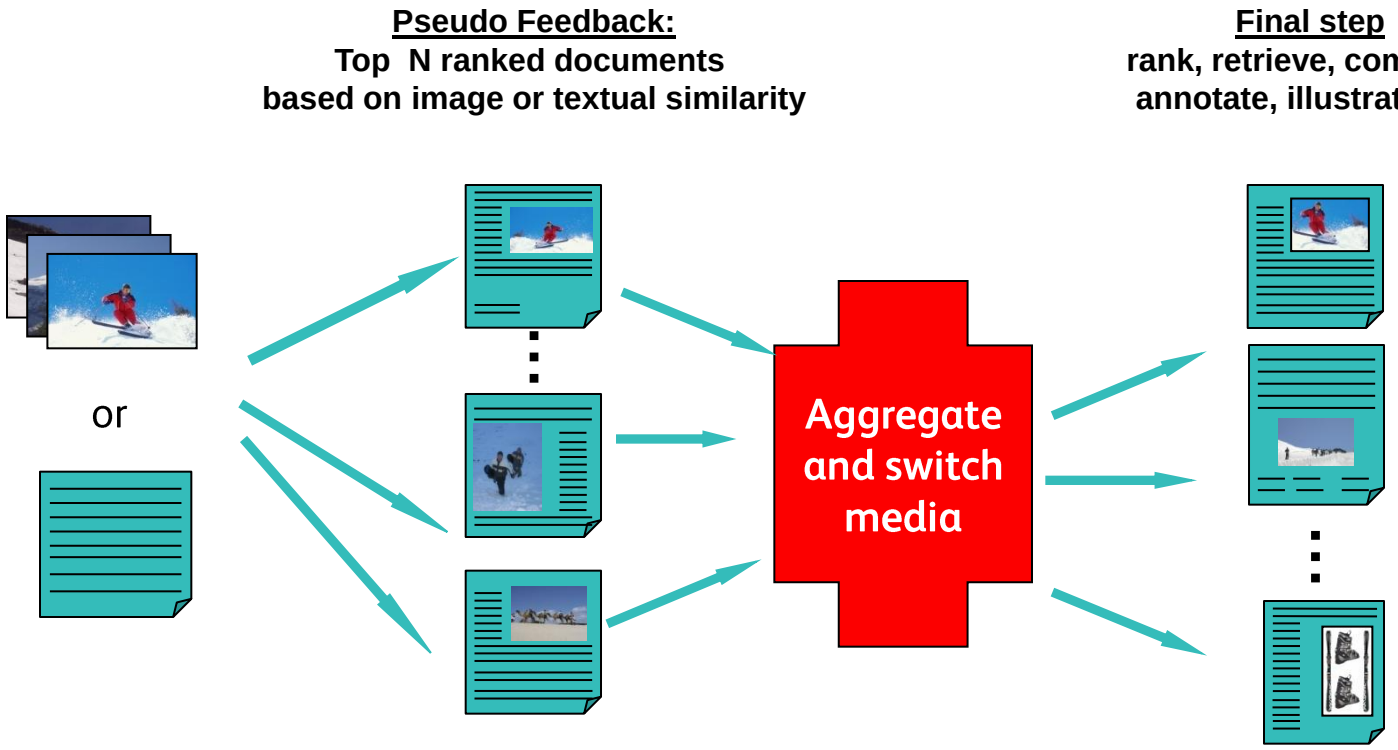
xerox

Merging visual with text- SOA

- Early fusion:
  - Feature concatenation of low level features
  - Co-occurences or joint probabilities between textual and visual features (Mori et al 1999, Duygulu et al 2002, Vinokourov et al 2003, Blei et al 2003, etc)
- Late fusion
  - Late score combination of mono-media results (Maillot et al 2006, Clinchant et al 2007)
- Intermediate level fusion
  - Relevance models (Jeon et al 2003)
  - Trans-media (or intermedia) feedback (Maillot et al 2006, Chang et al 2006)
  - Cross-media similarity measure (Clinchant et al 2007)

Intermediate level fusion

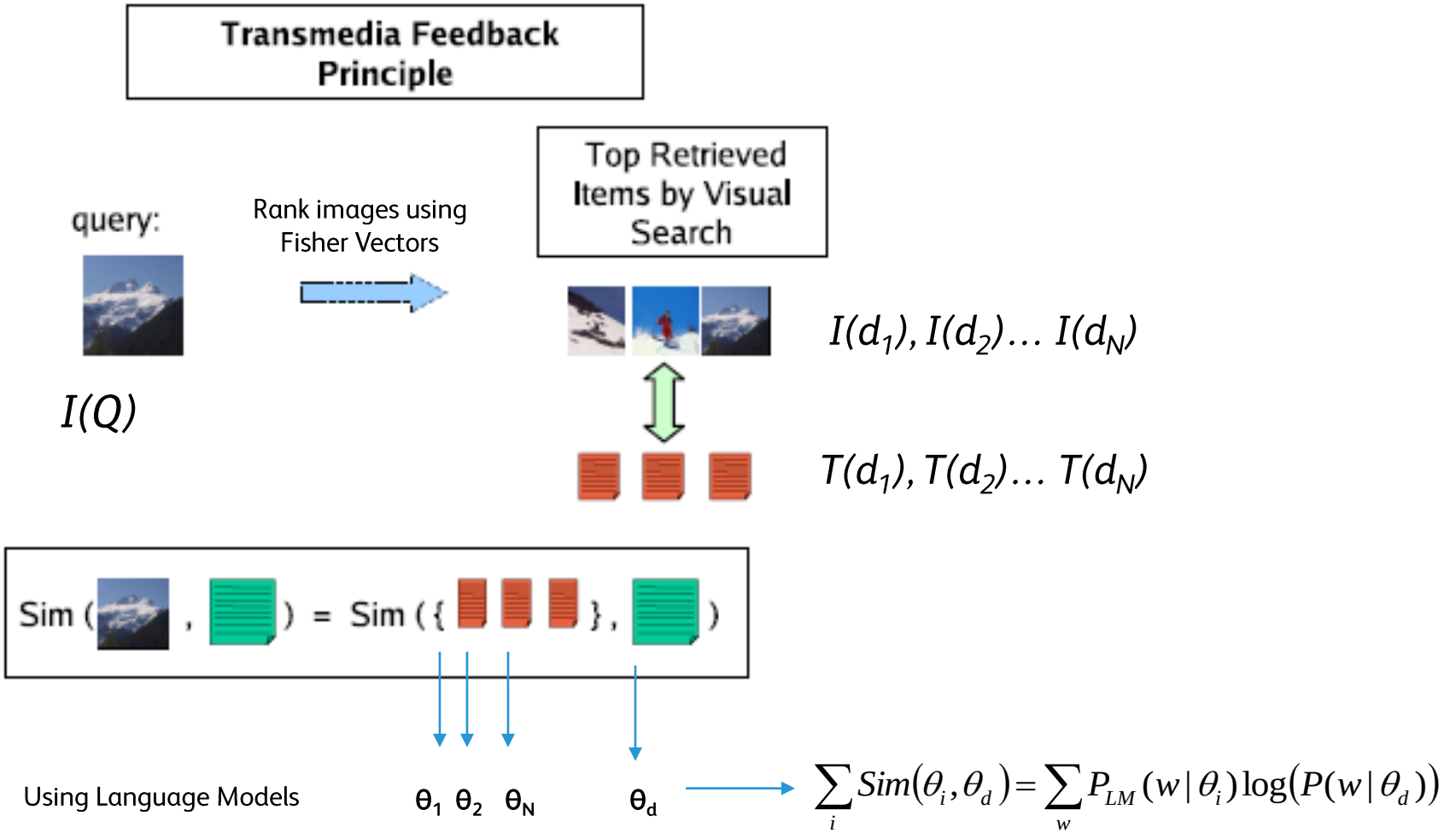
- The main idea is to switch media during using pseudo feedback process:
  - use one media type to gather relevant multimedia objects from a repository
  - use the dual type to step further (retrieve, annotate, etc)



*Crossing textual and visual content in different application scenarios, Ah Pin et al, MTPA (42)1, March 2009

xerox

The main idea (e.g. visual query)



*Crossing textual and visual content in different application scenarios, Ah Pin et al, MTPA (42)1, March 2009

Aggregating Similarity Measures

- Using image I as query in the PF:

$$sim_{\text{IMG-TXT}}(I,T) = \sum_{d_i \in N_{\text{IMG}}(I)} sim_{\text{IMG}}(I, I(d_i)) \cdot sim_{\text{TXT}}(T(d_i), T) = \kappa\left(\boldsymbol{\omega}_{\text{IMG}}(I), Id_{K_{\text{IMG}}}\right) \cdot \mathbf{W}_{\text{TXT}}$$

- where
  - sim_{TXT} is the ***textual*** similarity measure
  - $\mathbf{W}_{TXT}$ is the mono-modal (here **textual**) similarity matrix of the repository
  - $N_{VIS}(I)$ contains the K_{VIS} nearest ***visual*** neighbors of I
  - $\omega_{TXT}(T)$ is a **textual** distance vector between T and the repository
  - and κ is a thresholding function which puts to zero all the distances not corresponding to documents in $N_{VIS}(I)$

- Similarly using text T as query in the PF:

$$sim_{\text{TXT-IMG}}(I,T) = \sum_{d_i \in N_{\text{TXT}}(T)} sim_{\text{TXT}}(T(d_i), T) \cdot sim_{\text{IMG}}(I, I(d_i)) = \kappa\left(\boldsymbol{\omega}_{\text{TXT}}(T), Id_{K_{\text{TXT}}}\right) \cdot \mathbf{W}_{\text{IMG}}$$

*Crossing textual and visual content in different application scenarios, Ah Pin et al, MTPA (42)1, March 2009

Aggregating Similarity Measures

- Similarly we have:

$$sim_{\text{TXT-TXT}}(T',T) = \sum_{d_i \in N_{\text{TXT}}(I)} sim_{\text{TXT}}(T',T(d_i)) \cdot sim_{\text{TXT}}(T(d_i),T) = \kappa\left(\boldsymbol{\omega}_{\text{TXT}}(T'), Id_{K_{\text{TXT}}}\right) \cdot \mathbf{W}_{\text{TXT}}$$

- and:

$$sim_{\text{IMG-IMG}}(I,I') = \sum_{d_i \in N_{\text{IMG}}(I)} sim_{\text{IMG}}(I',I(d_i)) \cdot sim_{\text{IMG}}(I,I(d_i)) = \kappa\left(\boldsymbol{\omega}_{\text{IMG}}(I'), Id_{K_{\text{IMG}}}\right) \cdot \mathbf{W}_{\text{IMG}}$$

- So finally the similarity between two multi-modal objects D and D' is defined as:

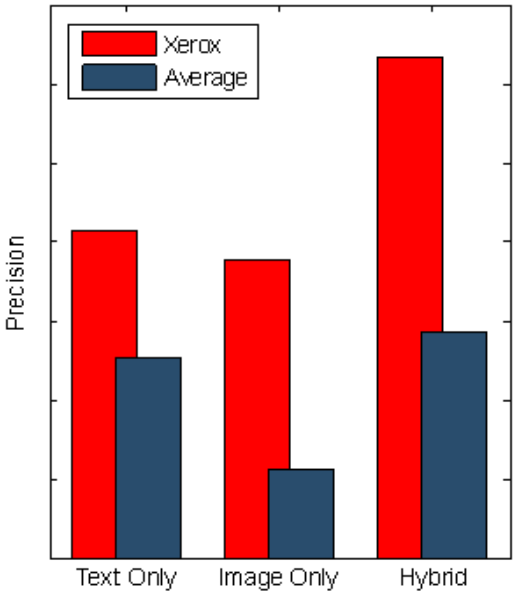
$$\begin{aligned} sim_{\text{X}}(D,D') = {} & \alpha \cdot sim_{\text{IMG-TXT}}(I(D),T(D')) + \beta \cdot sim_{\text{TXT-IMG}}(T(D),I(D')) \\ & + \gamma \cdot sim_{\text{TXT-TXT}}(T(D),T(D')) + \delta \cdot sim_{\text{IMG-IMG}}(I(D),I(D')) \end{aligned}$$

*Crossing textual and visual content in different application scenarios, Ah Pin et al, MTPA (42)1, March 2009

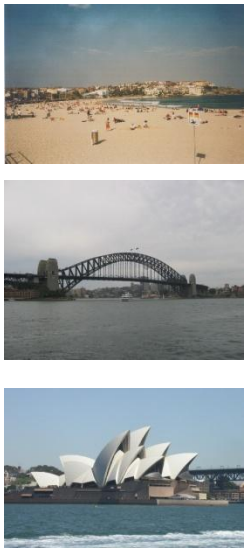
Multimedia Information Retrieval

- The multimodal documents U in the repository are ranked according to their cross-modal similarity with the multimodal query Q : $sim_X(Q, U)$.
- Winner 3 years consecutively of Image Clef **Cross-Modal** Photo Retrieval Task

ImageClef 2007

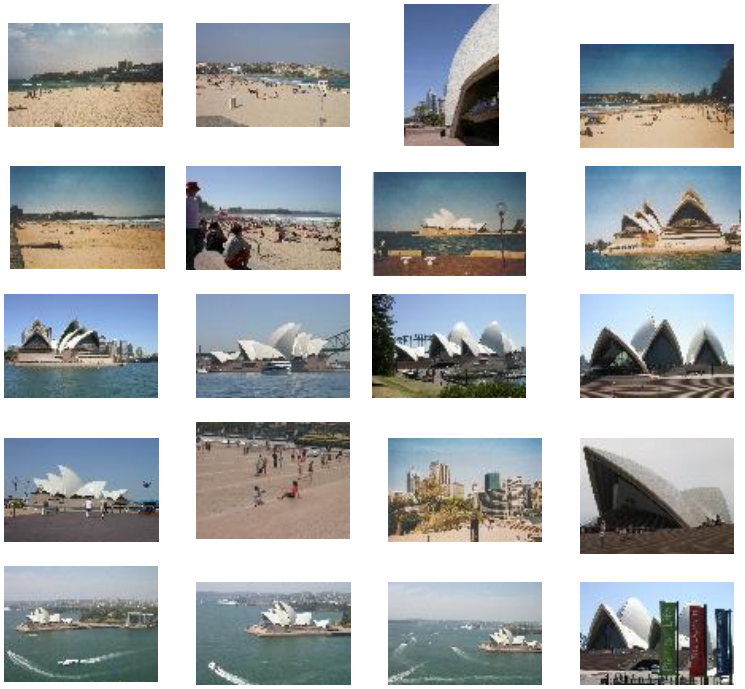


Q



views of Sydney's world-famous landmarks

TOP 20



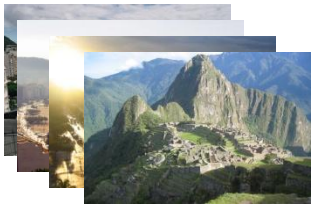
*** Xrce's participation to ImageClefPhoto 2007, 2008, 2009 S. Clinchant, G. Csurka, Ah Pine and J.M. Renders**

xerox

Assisting Hybrid Content Generation

– For example: a “simulated “ Travel Blog Assistant System

Upload user images



Downloaded Flickr images

Multimodal Knowledge base

e.g. Internet
Wikipedia
Travel Blog
Proprietary…

IAPR TC12 Benchmark Photo Repository

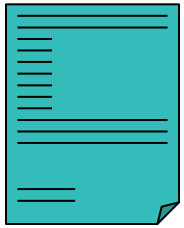
TBAS

Image auto-annotation

Texts illustration

Link text with images

Write blog text

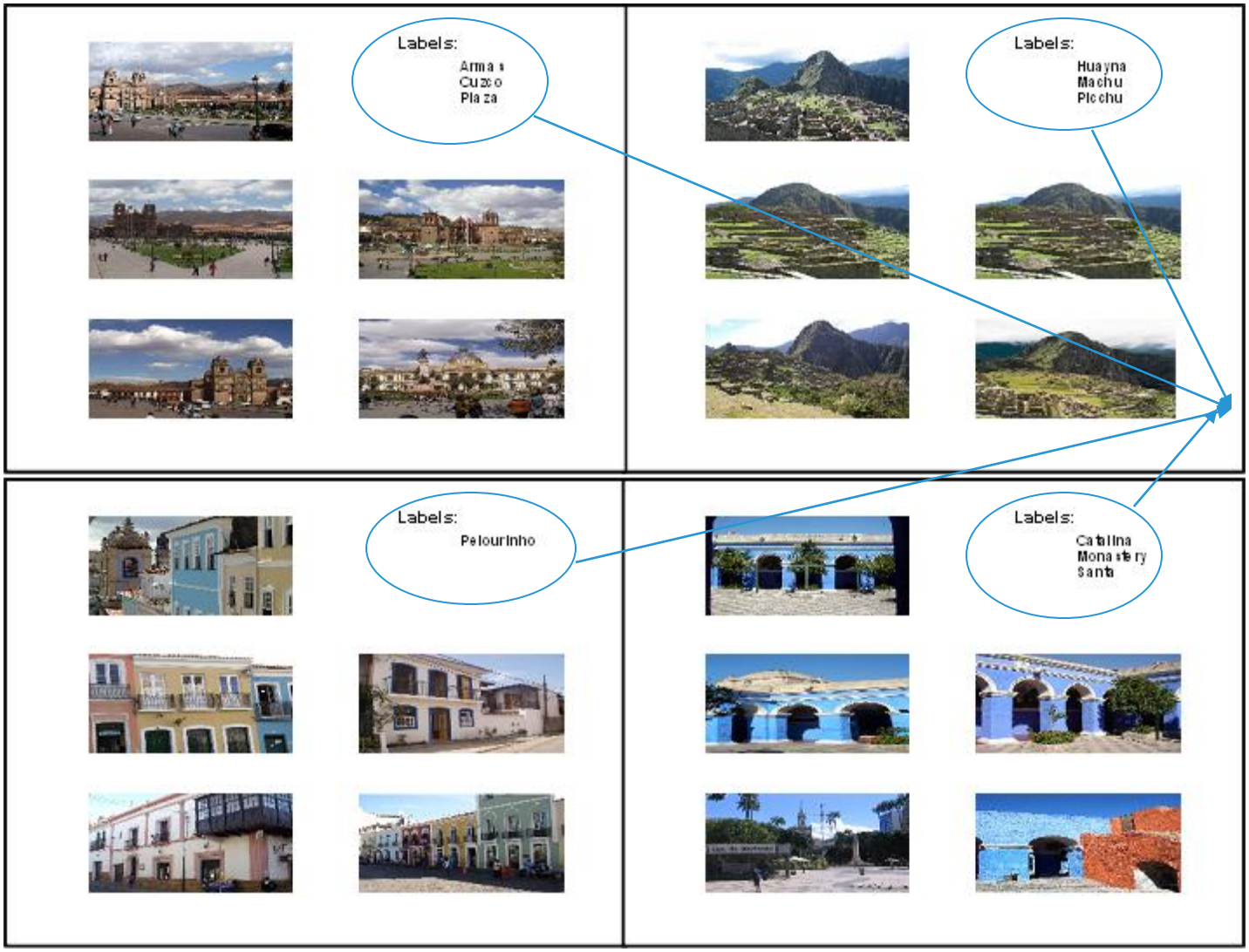


Real blog paragraphs

Published travel blog

*Travel Blog Assistant System, Csurka et al,, MMIU workshop, MMIU Workshop,VISAPP 2008

Image annotation using the repository



Annotations obtained for test (flickr) images from the aggregated text of the 4 top ranked images

*Travel Blog Assistant System, Csurka et al,, MMIU workshop, MMIU Workshop,VISAPP 2008

xerox

Text illustration

- Given a text *T* find the closest image(s) $I(d_i)$ in the repository according to $\text{sim}_X(T, I(d_i))$
- Example: *T* a paragraph from a Travel blog page (www.travelpod.com) and the repository is the IAPR TC12 Benchmark Photo Repository

Blog text

After dumping our bags at our pousada (two blocks from the beach) and flinging on our swim suits, we headed down to the world´s most famous beach... Copacabana. Along with its neighbour Ipanema, it´s been immortalised in a song and is synonymous with glamour and beautiful bodies.



Images from the Repository (IAPR)

*Travel Blog Assistant System, Csurka et al,, MMIU workshop, MMIU Workshop,VISAPP 2008

xerox

Examples of text and images linked through the repository

Our plans to hit Copacabana beach the next day and check out hot Brazilian girls in skimpy bikinis were ruined by the weather. It rained all day! Can you believe that. I think we'll be heading to another place mid-week for some beach time.

There is a lot of tourists there from around ten until three, but it didn't feel as crowded as we'd feared. We started there for 12 hours- saw the sunrise and sunset, and walked the citadel twice. It is an awesome site in the proper sense of the word (Yanks take note). Bloody magic. Some archeologists reckon that Machu Picchu could have predated the Inca but that they did a lot of improvements.




Blog texts

$$sim_X(I,T)$$




Flickr images

*Travel Blog Assistant System, Csurka et al,, MMIU workshop, MMIU Workshop,VISAPP 2008

Conclusion

We have shown that :

- Fisher Vector generally higher performance than BOV.
- A strong "*model-dependent*" (visual vocabulary), but "*class-independent*" representation.
- We successfully applied it (often state-of-the art performance) to:
  - Image categorization
  - Semantic image segmentation
  - Large Scale Image Retrieval
  - Intelligent Thumbnailing
  - Image and multi-modal document ranking and information retrieval
  - Assisting Hybrid Content Generation

Thank you for your attention!

xerox