BOV and Fisher Vectors in Image Classification and Retrieval

Gabriela Csurka

Xerox Research Centre Europe
Meylan, France

xerox

Growth of digital images

- It has never been easier to create images and videos.
  - 81% of worldwide mobile phones have cameras
  - 80% of U.S. households own digital cameras
  - 11% of Americans have more than 10,000 digital photos
- It has never been easier to share images and videos:

4 B images

8 B images

15 B images

20 B images

- *How to efficiently access information in such large repositories?*
  - Using a generic image representation.
  - Combining with extra information related to images (e.g. text)

Generic Image Representation

- Independent on the object class:

- Handling view, lighting changes, occlusion as well as intra-class variability:

- Able to represent different type of images:

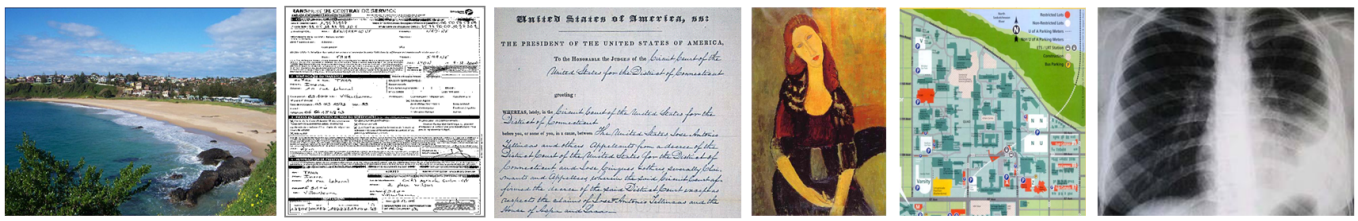
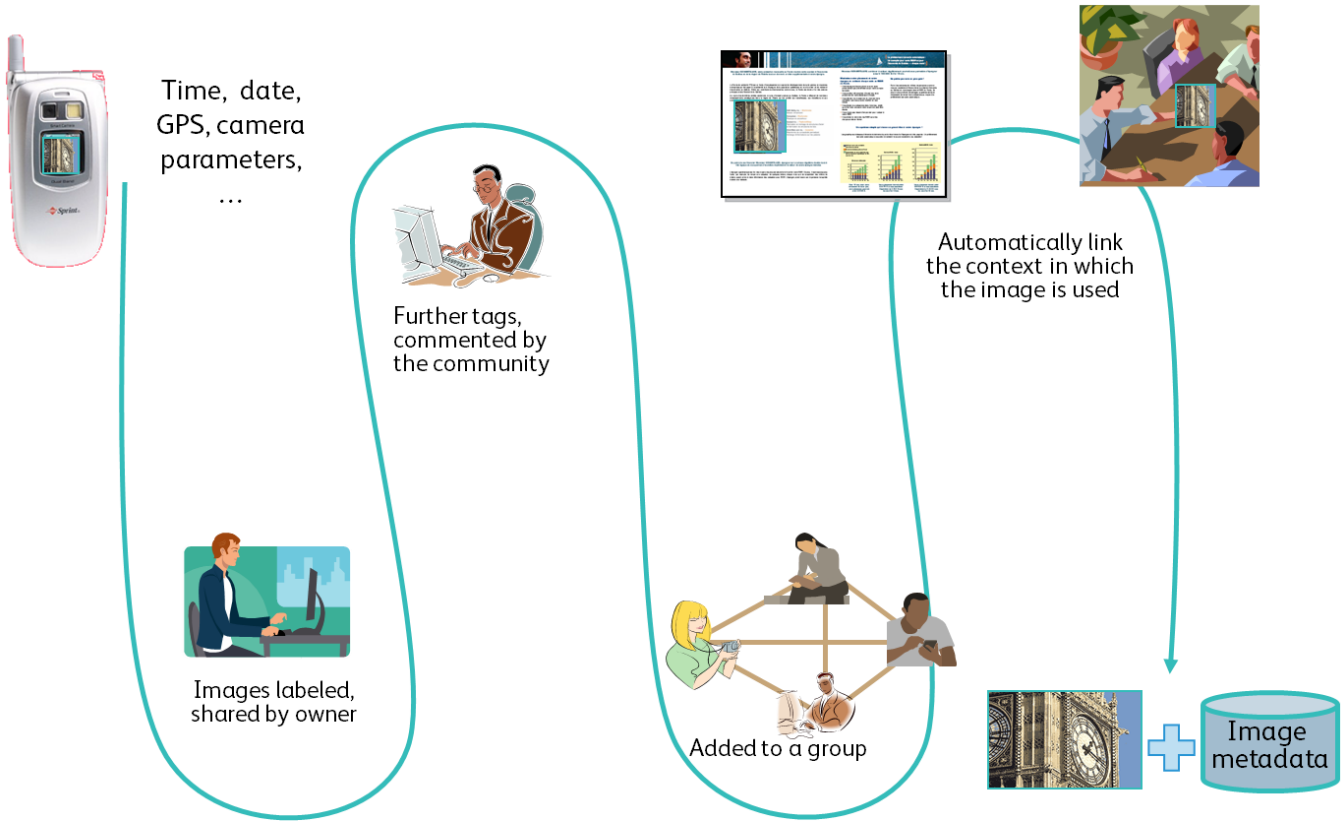


Image metadata



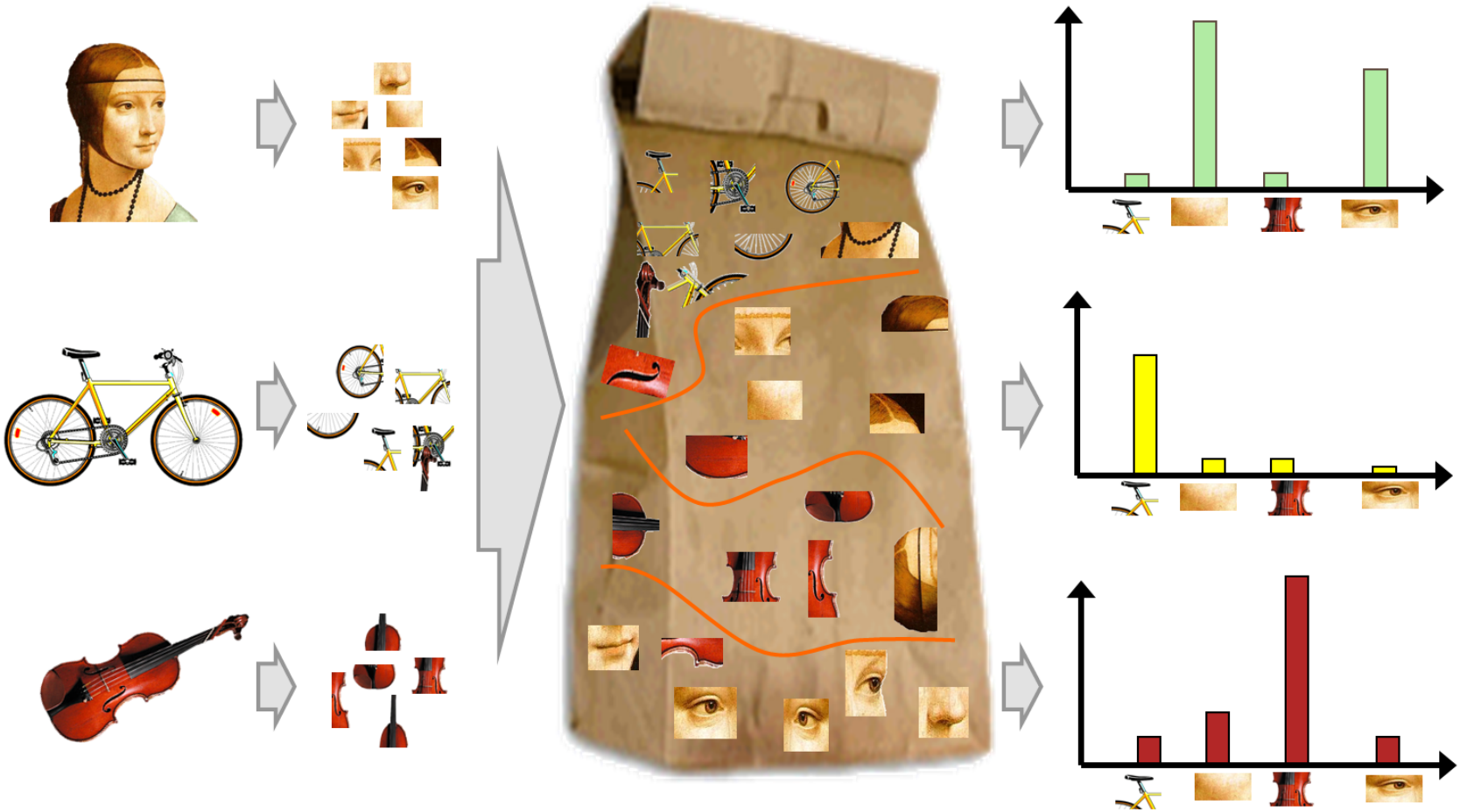
The images are:

- grouped, linked, geo-tagged, annotated, commented,
- related to other types of media, used in context...

These provide a rich source of extra information.

Outline

Bag of Visual-Words (BOV)* †



* J. Sivic and A. ZissermanVideo Google: A Text Retrieval Approach to Object Matching in Videos, ICCV03

† G. Csurka, C. Dance,et al, Visual categorization with bags of keypoints, SLVC04

@ Slide Credit: Marco Bressan and Li Fei-Fei

Inspired by text

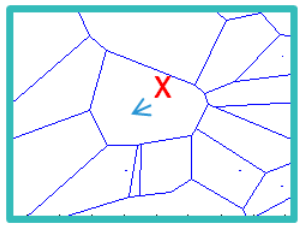
Order-less document representation:

- frequencies of words from a dictionary (Salton & McGill 1983)

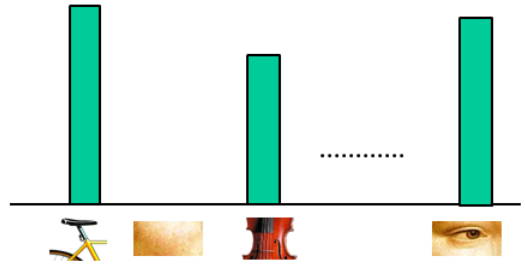
© Slide Credit: Svetlana Lazebnik

BOV: Visual Vocabulary (K-means vs GMM*)

K-means

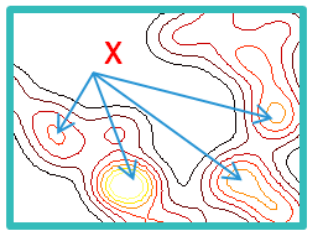


Hard assignment

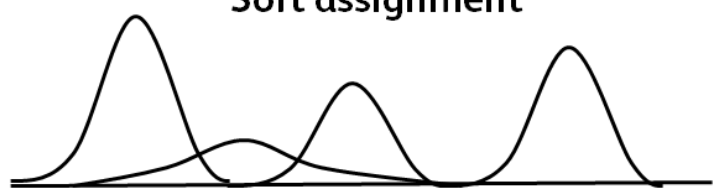


$$\gamma(I) = \sum_{t=1}^{T} [0, 0, \ldots, 1, \ldots, 0]$$

GMM



Soft assignment



$$\gamma(I) = \sum_{t=1}^{T} [\gamma_1(x_t), \gamma_2(x_t), \ldots, \gamma_N(x_t)]$$

* J. Farquhar, S. Szedmak, H.Meng, and J. Shawe-Taylor. Improving 'bag-of-keypoints' image categorisation. Technical report, University of Southampton. 2005.

The Fisher Vectors*

- The *Fisher Vector* extends the BOV by going beyond counting (0-order statistics) to encoding second order statistics.
- It describes in which direction the parameters of the model (GMM) should be modified to best fit the data (an image).

The main idea:

- Characterize a sample of low level features extracted from the image $I = \{x_t, t = 1 \dots T\}$ by its deviation from the GMM distribution:

$$G_\lambda(I) = \frac{1}{T} \sum_{t=1}^{T} \nabla_\lambda \log \left\{ \sum_{j=1}^{N} w_j \mathcal{N}(x_t | \mu_j, \Sigma_j) \right\}.$$

  where we consider diagonal Σ_j matrices.

---

* **F. Perronnin and C. Dance, Fisher Kernels on Visual Vocabularies for Image Categorization, CVPR07.**

The Fisher Kernel

- To compare two images:
  - We use the Fisher Kernel on these gradients:

$$K(I,J) = G_\lambda(I)^\top F_\lambda^{-1} G_\lambda(Y) = (L_\lambda G_\lambda(I))^\top (L_\lambda G_\lambda(J)) = \Gamma_\lambda(I)^\top \Gamma_\lambda(I)$$

    where F_λ is the Fisher Information Matrix and $F_\lambda^{-1} = L_\lambda^\top L_\lambda$.
  - Hence, learning a classifier with the Fisher Kernel is equivalent to learning a linear classifier on the Fisher Vectors $\Gamma_\lambda(I)$.

- We further normalize the Fisher Vector*:
  - Power normalization: to make the distribution of features in a given dimension m less peaky around zero:

$$f(z) = \text{sign}(z)|z|^\alpha \text{ with } \alpha = 0.5$$

  - L2 normalization: to cancel dependence on the proportion of image specific information w.r.t. proportion of background.

---

* F. Perronnin, J. Sanchez and T. Mensink, Improving the Fisher Kernel for Large-Scale Image Classification, ECCV10.

Compressed Fisher Vectors

- Binary representation*:
  - Consider only the gradient according to the mean μ_i

$$\Gamma_{\mu_i^d}(I) = \underbrace{\frac{\psi_i(I)}{T\sqrt{w_i}}}_{b_i(I)} \underbrace{\left(\frac{\phi_i^d(I) - \mu_i^d}{\sigma_i^d}\right)}_{\delta_i^d(I)}$$

    where $\gamma_i(I) = \sum_{t=1}^{T} \gamma_i(x_t)$ and $\phi_i^d(I) = \frac{1}{\psi_i(I)} \sum_{t=1}^{T} \gamma_i(x_t) x_t^d$.
  - Then $b_i(I)$ is encoded by 1 if $\gamma_i(I) > 0$ and 0 otherwise;
  - and $\delta_i^d(I)$ is encoded by its sign.
- Product quantization†:
  - split FV into small sub-vectors of size m (*e.g.* $m = 8$).
  - perform Vector quantization for each subvector.
  - FV is represented as a vector of codebook indices.

---

* F. Perronnin, Y Liu, J. Sanchez, H. Poirier, Large-scale image retrieval with compressed Fisher Vectors, CVPR10

† F. Perronnin and J. Sanchez, High-Dimensional Signature Compression for Large-Scale Image Classification, CVPR2011.

Spatial Pyramid of BOVs or FVs*

- The spatial pyramid takes into account the rough geometry of a scene:
  - consider different spatial splits of the image (1x1, 1x3, 2x2);
  - use power and L2 normalization on each of the 8 BOV or FVs independently;
  - concatenate the BOVs or FVs on all spatial layouts.

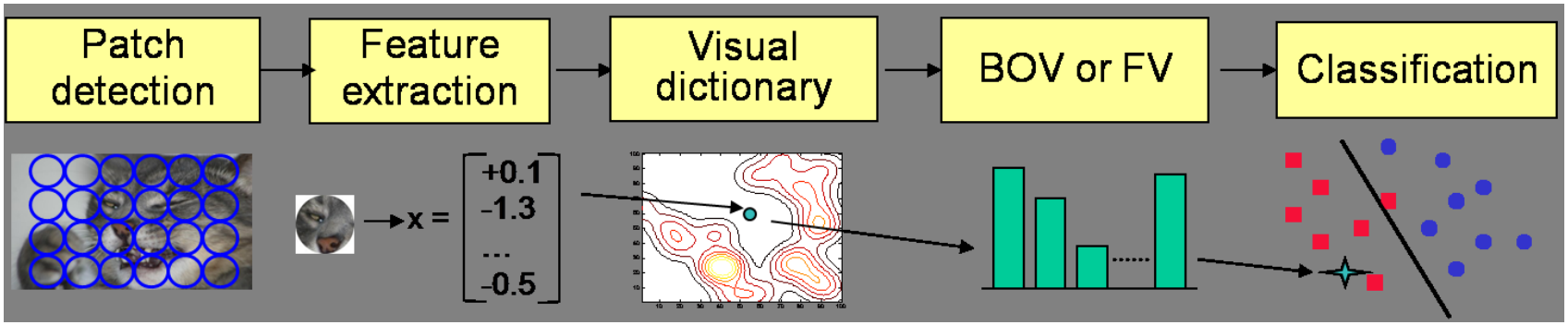
*S. Lazebnik, C. Schmid, and J. Ponce, Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories, CVPR06

Fisher Vector vs. BOV

- The Fisher Vector similarly to the BOV:
  - transforms a variable length feature set into a fixed sized representation
  - is *model-dependent* (visual vocabulary), but *class-independent* representation (image signature)
  - hence is suitable for *supervised* (classification) and *unsupervised* (clustering, retrieval) tasks

- However the Fisher Vector is:
  - a much richer representation of the low level feature distribution in the image than the BOV
  - for the same vocabulary size (N) adds almost no extra CPU cost
  - generally it performs better than a BOV of similar size (significantly higher cost)

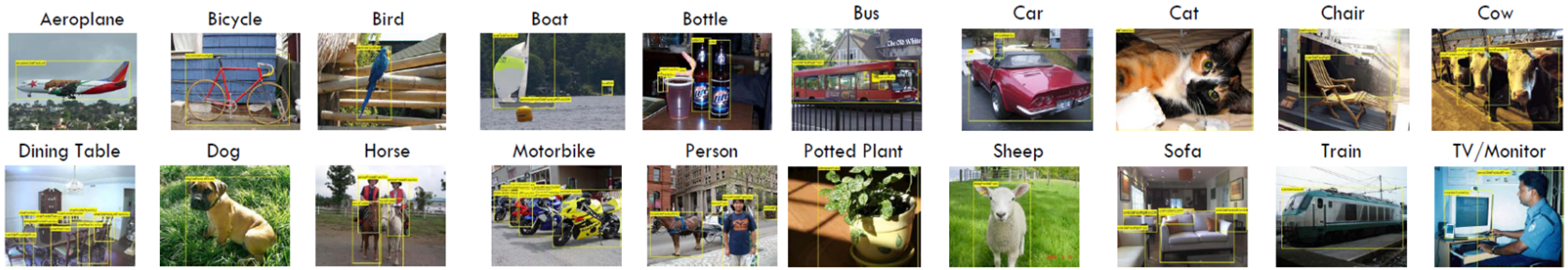
Generic Visual Categorization (GVC)*



- Patch detection: interest points, segments, regular patches, ...
- Feature extraction: SIFT, color statistics, moments, ...
- Visual dictionary: Kmeans, GMM, Random Forest, ...
- Image representations: BOV, FV, ...
- Classification: generative (pLSA, LDA) or discriminative (SVM, SLR)

* G. Csurka, C. Dance,et al, Visual categorization with bags of keypoints, SLVC 2004

Pascal VOC Challenges*

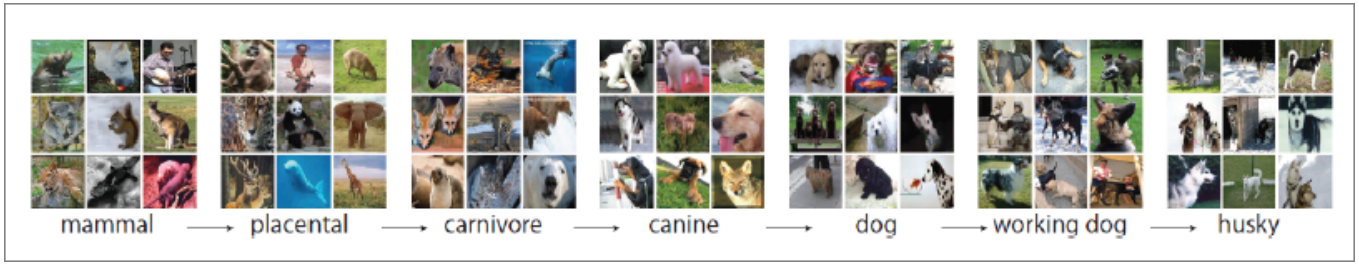


- Most participating systems in the last few years are using BOV.
- The methods mainly vary in:
  - the number of used local feature types and kernels
  - the way multiple features are combined (voting, MKL)
- The 2010 winning system further integrates object detection and segmentation with the global image classifier †.

† **Boosting Classification with Exclusive Context, Qiang Chen et al, VOC 2010**

* `http://pascallin.ecs.soton.ac.uk/challenges/VOC`

ImageNet Large Scale Visual Recognition Challenge*



- 1000 object classes, 200,000 mono-labeled images.
- Best performing systems uses:
  - extended BOV representations (FV or similar)
  - linear classifier

Cost	NEC-UIUC	XRCE	ISIL	UCI	FV-since*
accuracy at Top5	71.8	66.4	55.5	53.4	**74.3**

- Compression is important:
  - Storing all features is 2.8TB, compressed only 45GB
  - x64 compression leads to almost no accuracty loss

---

* **J. Sánchez and F. Perronnin,High-Dimensional Signature Compression for Large-Scale Image Classification, CVPR 2011**

* `http://www.image-net.org/challenges/LSVRC/2010`

Generic concepts: using same representation

- ImageClef Photo Annotation
  - 25000 multi-labeled images,
  - diverse concepts: scene, object, season, event, style, ...

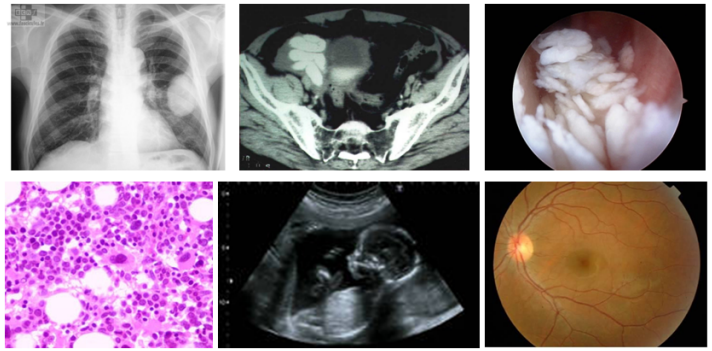
Run	MAP	AUC	F-ex	OS
FV	38.9	80.5	63.9	**64.5**
UVA	**40.7**	**82.6**	**68.0**	59.1

- Predicting the aesthetic quality
  - rated photos from `photo.net`,
  - BOV/FV with local color statistics better then designed aesthetics features

Datta	Ke	BOV	FV
75.85	76.53	81.86	89.90

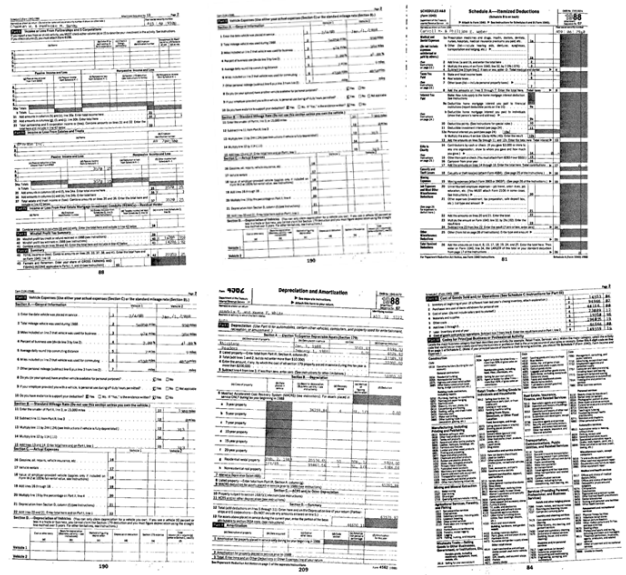
Same representation for diverse image types

- Medical Modality Detection
  - 4000 medical images,
  - 8 modalities (CT, MR, X-ray, PET, etc)



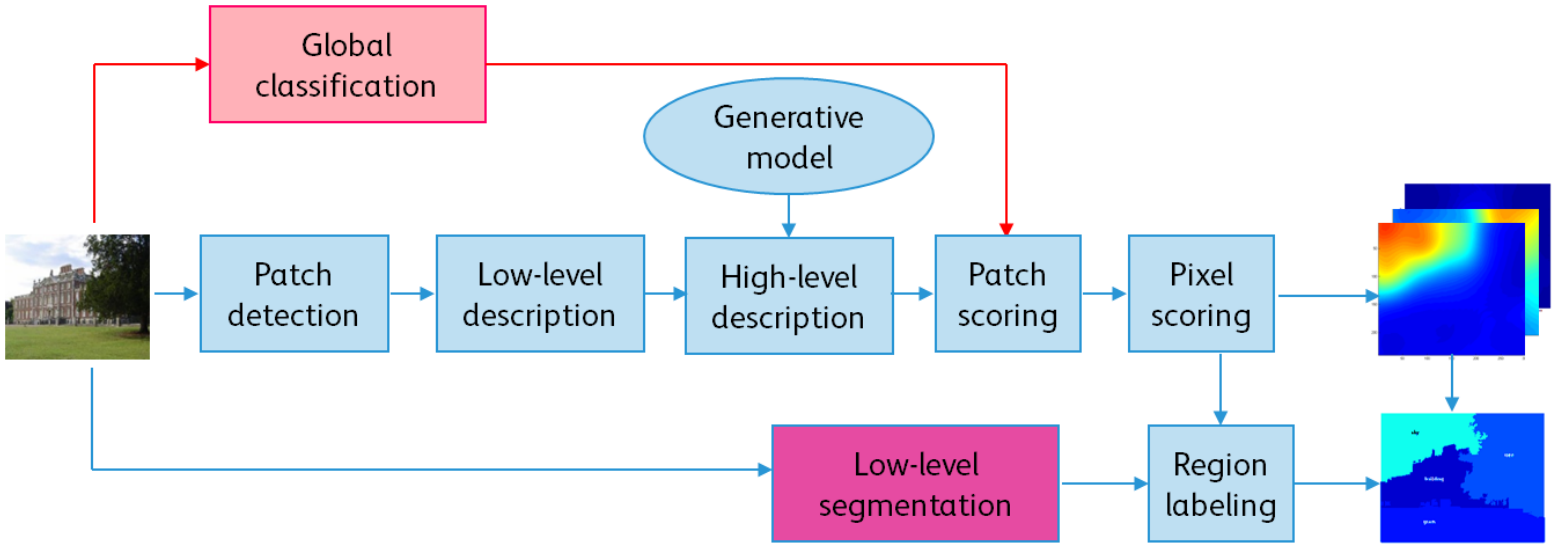
BOV	FV	Best others
80.4	**86.9**	82

- Categorizing US taxforms
  - 5590 scanned images,
  - 20 categories



FV	RunLenght
100	99.8

Similar framework for semantic image segmentation*

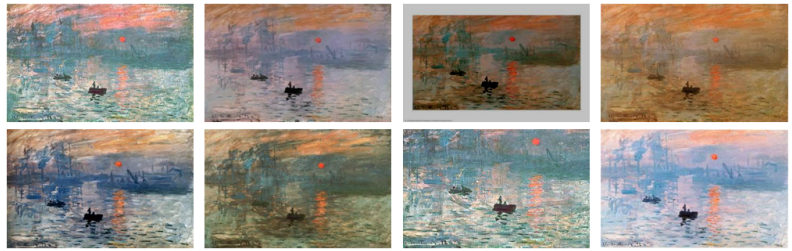


- Similar pipeline as for image categorization:
  - The classifiers are learned at the patch level
  - Combined with low level segmentation improves the borders
  - Combined with global scores can improve the accuracy

*A Simple High Performance Approach to Semantic Segmentation, G. Csurka and F. Perronnin, BMVC08.

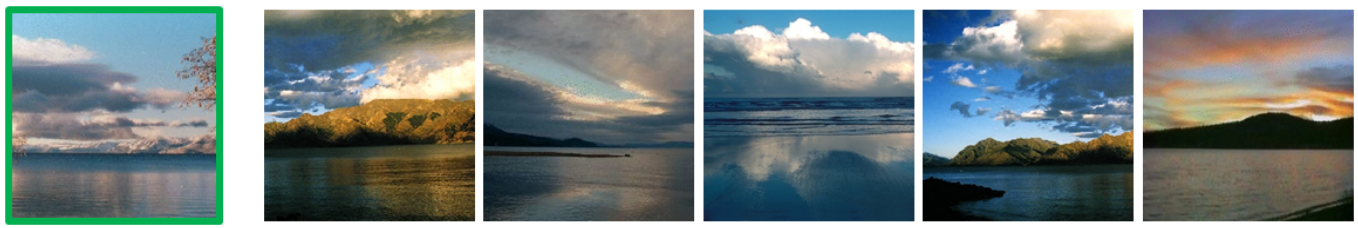
Same representation is suitable for unsupervised tasks

- Near Duplicate Detection



- Same object detection (Google Goggles type application): logos, landmarks, artworks, etc

- Content based image retrieval

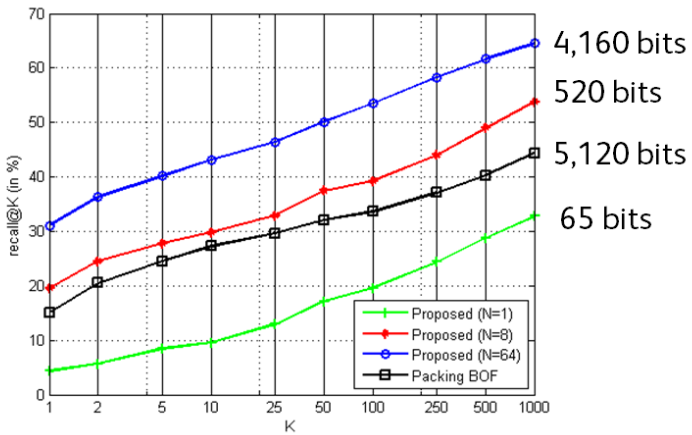


xerox

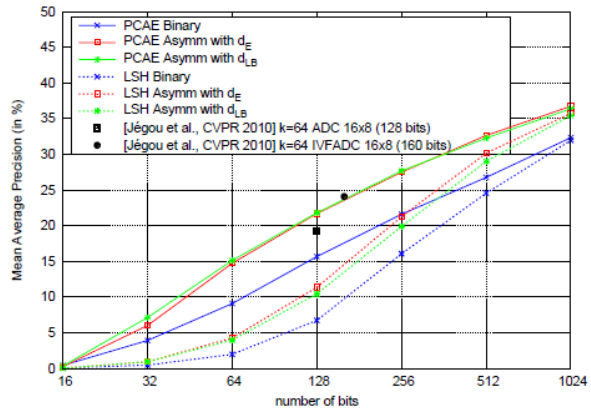
Suitable for large scale CBIR

- The INRIA Holiday dataset:
  - 1,491 images of 500 scenes / objects plus 1M of random Flikr images.

- Binary FV[a] compared with packed BOV[b]:



- Product quantized FV[c] compared with VLAD[d] features:



[a] Large-scale image retrieval with compressed Fisher Vectors, F. Perronnin et al, CVPR10

[b] Packing bag-of-features, H. Jégou et al, ICCV09

[c] Asymmetric Distances for Binary Embeddings, A Gordo and F Perronnin, CVPR11

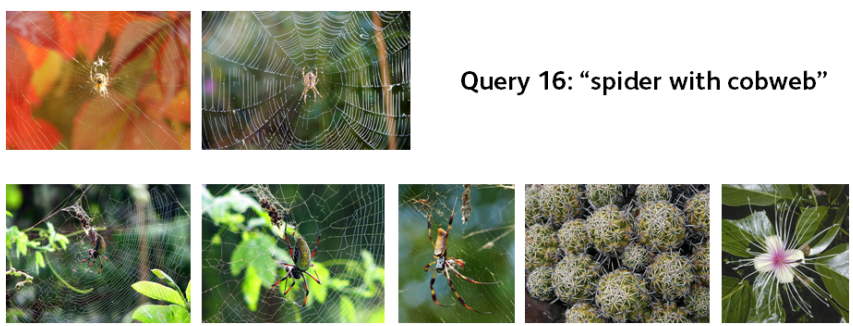
[d] Aggregating local descriptors into a compact image representation, H. Jégou et al, CVPR10.

Semantic Image Retrieval

- ImageClef* Photo Retrieval
  - 20,000 still natural images, 60 query topics.
  - Best (MAP=22%) visual only system both in 2007 and 2008.

Query 7: "group standing in salt pan"

- ImageClef Wikipedia Retrieval:
  - 237,000 Wikipedia images, 70 diverse query topics.
  - Best (MAP=5.5%) visual only system in 2010.



Query 16: "spider with cobweb"

* **ImageClef Evaluation Forum:** `http:www.imageclef.org`

xerox

Semantic ambiguity

- It is generally said that *"a picture is worth a thousand words"*,
  - but in the context of information retrieval, which "word" is meant when an image is used as a query ?
  - e.g. (below) if only images are used as query, what we mean:
    - sharks underwater (the actual query) or
    - blue background and fish-like shape (what was rather found)

Query: "sharks underwater"

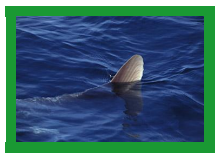

CBIR






TEXT+IMAGE

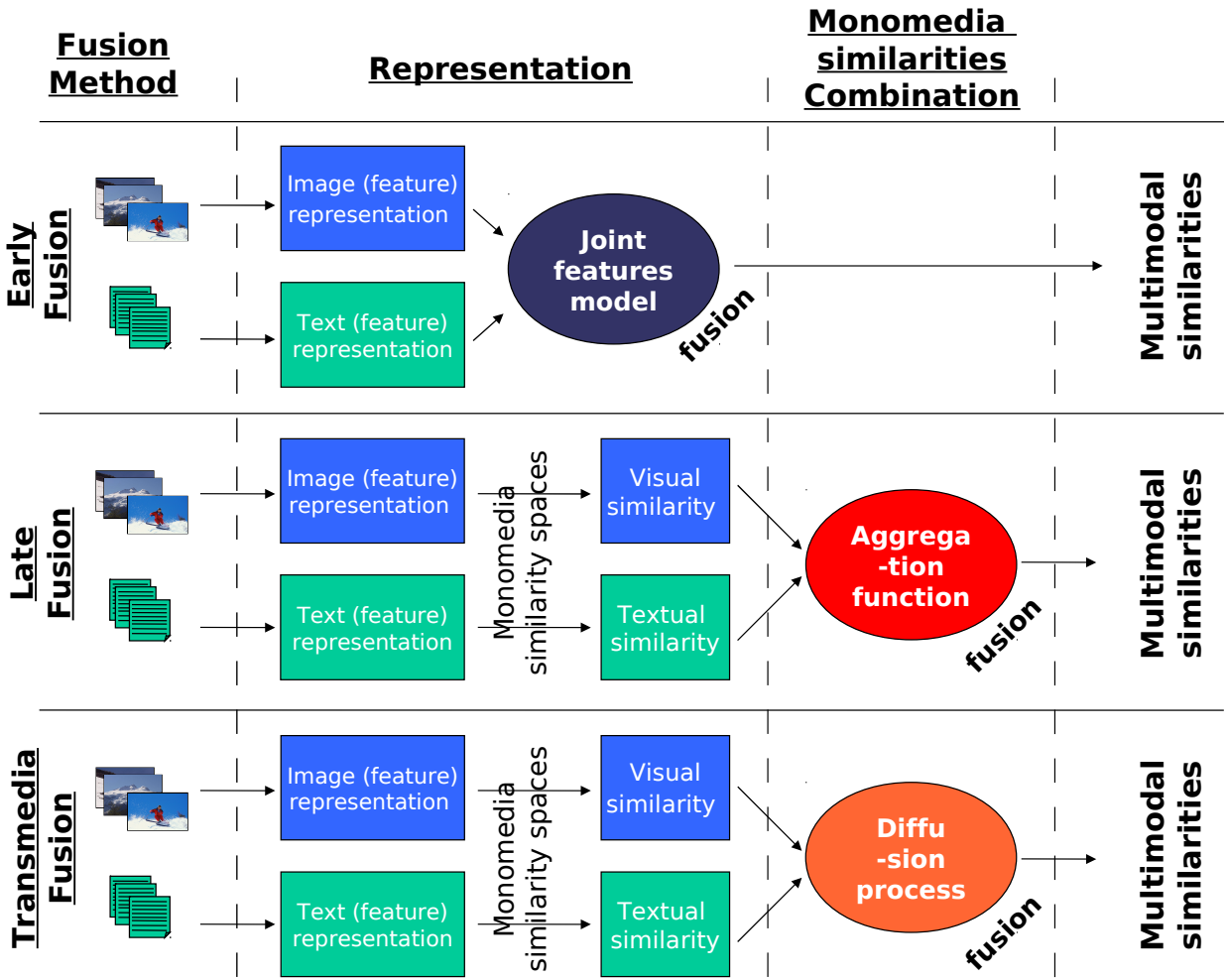





- Combining images with text always helps!

Outline

Visual and Textual Information Fusion*



* S. Clinchant, J. Ah-Pine, G. Csurka, Semantic Filtering for Textual and Visual Information Fusion, ICMR11

Fusion Level

- Early Fusion:
  - With or without feature weighting (Deselaers et al '04, Ferecatu & Sahbi '08, Moulin et al '08)
  - Joint models such as (K)CCA, multi-view (Mori et al '99, Lavrenko et al '03)
  - Translation models (Duygulu et al '02)

- Late Fusion:
  - Linear combination: most often and most succesfully used.
  - CombSum (Ho et al '94), CombProd (Martinez-Fernandez '04), CombMNZ (Shaw & Fox '94),...
  - Image Reranking (Hoi et al 05, Zhou et al '08, Popescu '10)
  - Late Semantic Combination (S. Clinchant et al '11, Csurka et al '11)

- Intermediate level fusion:
  - Relevance models (Jeon et al '03, Lavrenko et al '03)
  - Transmedia query expansion (Maillot et al '06, Chang et al '06)
  - Cross-media similarities (Clinchant et al '07, Ah-Pine et al '09)

Late fusion and image reranking

- Late Fusion:

$$s_{LSC}(q,d) = \alpha_t s_t(q,d) + \alpha_v s_v(q,d)$$

  - where $\alpha_t = \alpha$ and $\alpha_v = 1 - \alpha$ are positive weights that sum to 1.

- Image reranking:

$$s_{IR}(q,d) = \begin{cases} s_v(q,d) & \text{if } d \in \mathrm{KNN}_t(q) \\ 0 & \text{otherwise} \end{cases}$$

  - where $\mathrm{KNN}_t(q)$ denotes the set of the K most similar objects to q according to the textual similarities.

Late Semantic Combination (LSC)*

- Combines image reranking with late fusion:

$$s_{LSC}(q,d) = \alpha_t s_t(q,d) + \alpha_v s_{IR}(q,d)$$

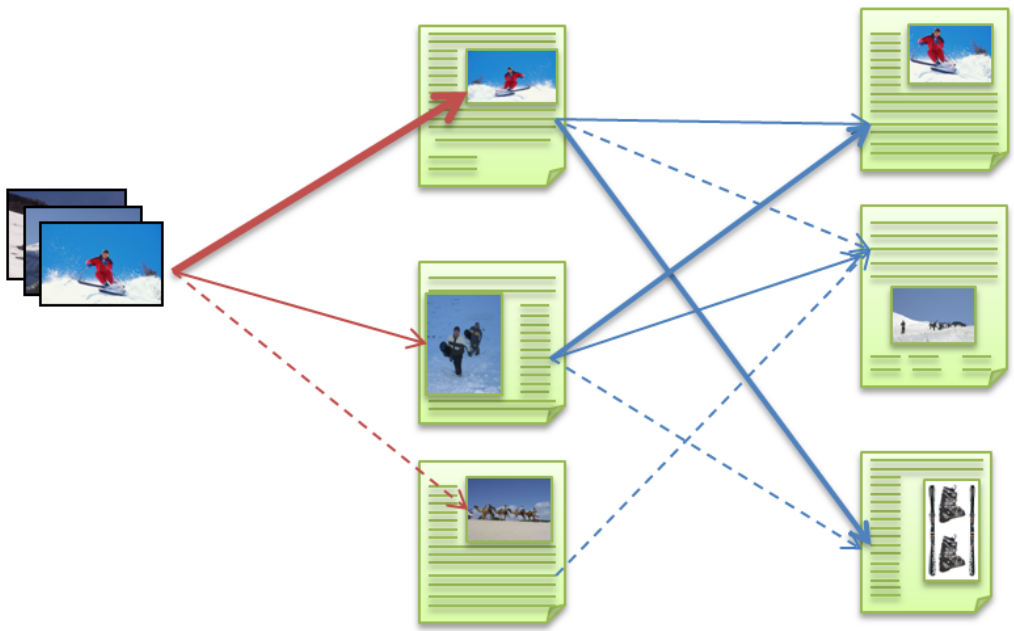
- Several advantages:
  - Similarly to image reranking, it exploits the fact that text and images are semantically expressed at different levels.
  - It further exploits the text expert by combining the semantically filtered image scores with the text scores,
  - It is highly scalable as image similarities has to be computed only between the query and images of the top K selected documents.

*S. Clinchant, J. Ah-Pine, G. Csurka, Semantic Filtering for Textual and Visual Information Fusion, ICMR11

Intermediate level fusion*

- The main idea is to switch media during pseudo feedback process:
  - use one media to gather top relevant objects from a repository
  - use the dual modality to rerank

  It can be seen as transmedia pseudo relevance feedback.



* J. Ah-Pine et al, Crossing textual and visual content in different application scenarios. Multimedia Tools and Applications, 2009

Trans-media pseudo relevance feedback (TMRF)*†

- Use first image to query and then aggregate text similarities from top retrieved elements to rank:

$$s_X(q,d) = \sum_{d' \in \mathrm{KNN}_v(q)} s_v(q,d') s_t(d',d)$$

  - where $\mathrm{KNN}_v(q)$ denotes the set of the K most similar objects d' to q using visual similarities $s_v(q,d')$,
  - $s_t(d',d)$ is the textual similarity between the items d and d' of the collection (query independent)

---

*S. Clinchant, J. Renders, and G. Csurka. XRCE's participation to ImageCLEF 2007

†J. Ah-Pine et al, Crossing textual and visual content in different application scenarios. Multimedia Tools and Applications, 2009

Trans-media pseudo relevance feedback (TMRF)*†

- Use first image to query and then aggregate text similarities from top retrieved elements to rank:

$$s_X(q,d) = \sum_{d' \in \mathrm{KNN}_v(q)} s_v(q,d') s_t(d',d)$$

  - where $\mathrm{KNN}_v(q)$ denotes the set of the K most similar objects d' to q using visual similarities $s_v(q,d')$,
  - $s_t(d',d)$ is the textual similarity between the items d and d' of the collection (query independent)

- Note that:
  - better when further combined with textual similarities $s_t(q,d)$;
  - it can exploit multi-modality even for mono-modal queries;
  - we can begin with textual query and aggregate visual similarities:

$$s_X(q,d) = \sum_{d' \in \mathrm{KNN}_t(q)} s_t(q,d') s_v(d',d)$$

*S. Clinchant, J. Renders, and G. Csurka. XRCE's participation to ImageCLEF 2007

†J. Ah-Pine et al, Crossing textual and visual content in different application scenarios. Multimedia Tools and Applications, 2009

Multi-modal Retrieval Experiences

- FV combined with text using LSC or TMRF was winning system in several ImageCLEF* Tasks between 2007-2010:
  - ImageCLEF Photo 2007-2008 (IAPR):
    - 20,000 natural images with associated semantic descriptions;
    - 60 query topics each with 3 example images.
  - Photographic retrieval 2009 (BELGA):
    - 498,920 news photographs with short English captions
    - 50 query topics each with 1-10 example images.
  - Wikipedia Retrieval 2010 (WIKI)
    - 237,434 images with associated captions;
    - original Wikipedia pages from where the images were extracted;
    - 70 query topics with 1-3 query images.
  - Medical Retrieval 2010 (MED)
    - 77,477 medical images of different modalities,
    - we used the image modality classifier as filter
    - 16 ad-hoc query topics with 2-3 sample images.

---

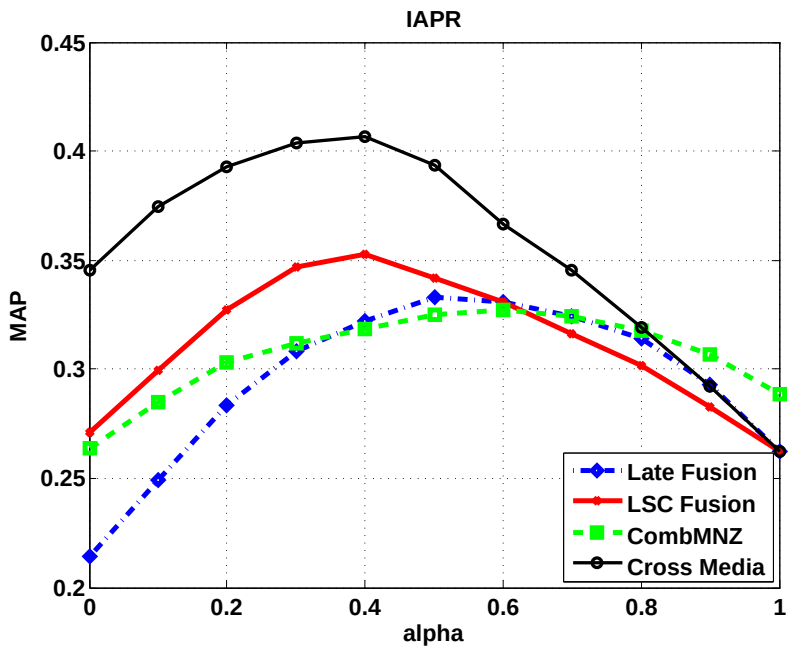
* **ImageClef Evaluation Forum:** `http:www.imageclef.org` **Photo Retrieval**

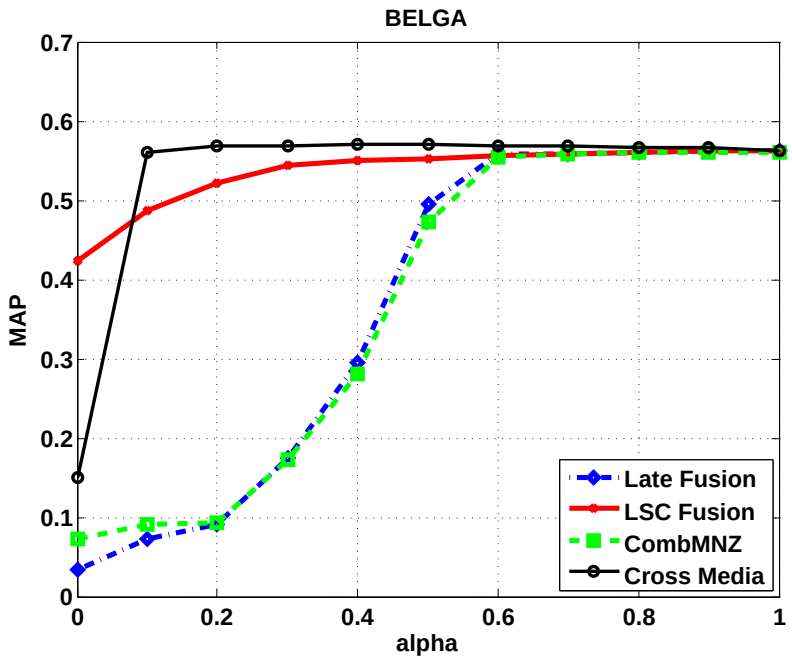
Comparative retrieval results*

MAP	IAPR	BELGA	WIKI	MED
s_t	26.3	56.2	20.5	31.4
s_v	22.1	3.3	5.5	0.9
Best Late Fusion	34.0	56.2	21.9	31.4
Image Reranking	27.6	42.4	19.4	8.3
CombMNZ	33.5	56.0	23.7	27.8
Best Cross-Media	**42.1**	**57.0**	21.6	31.4
Cross-Media with K=3	40.5	56.6	20.6	31.4
Best LSC	35.4	56.3	**26.6**	**36.9**

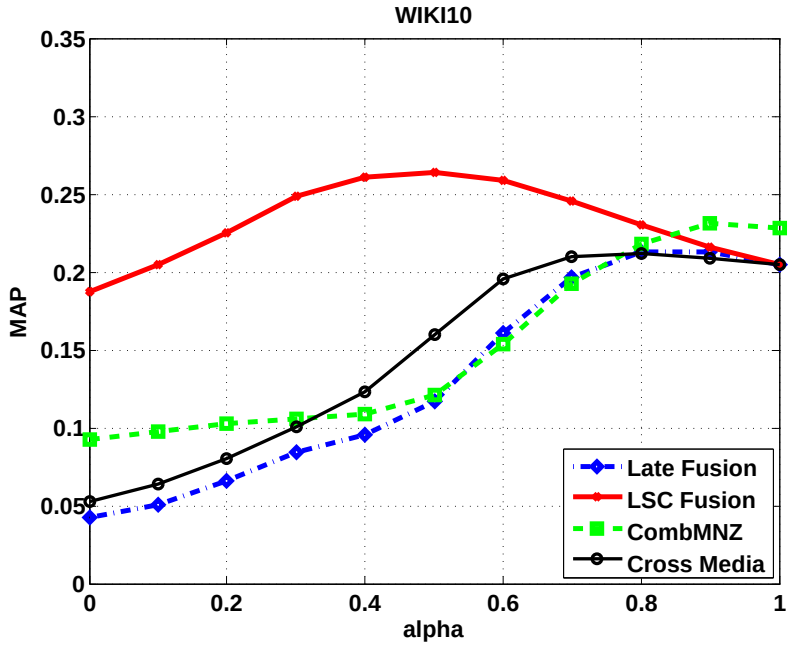
* S. Clinchant, J. Ah-Pine, G. Csurka, Semantic Filtering for Textual and Visual Information Fusion, ICMR11

Varying α.

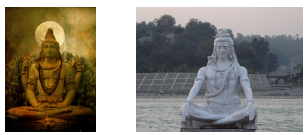




Wikipedia query examples



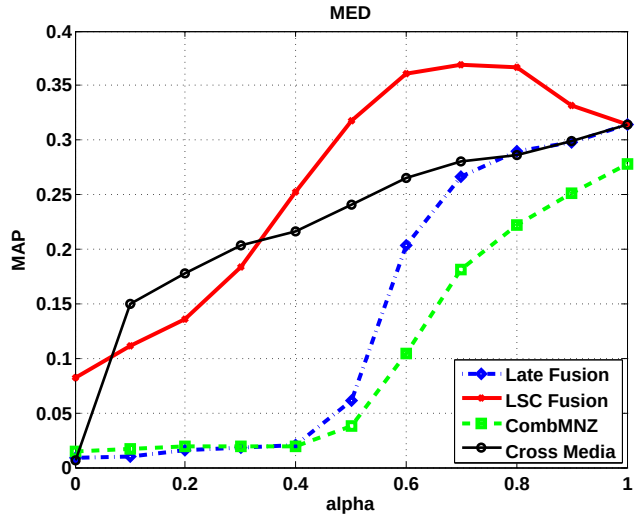
Query: "Shiva painting or scuplture"



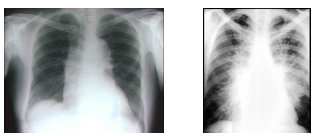
Late

LSC

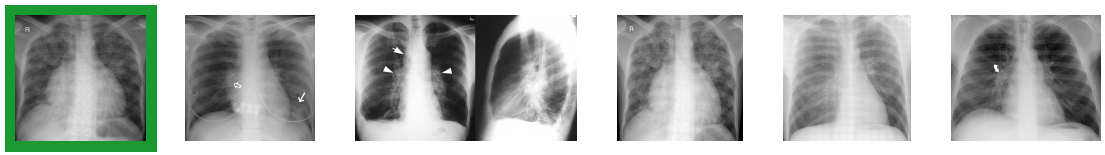
Medical image retrieval with modality detection



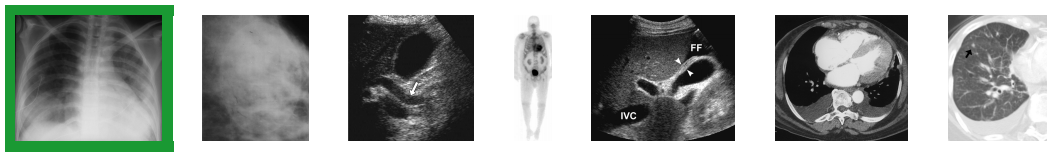
Query 4: "congestive heart failure"



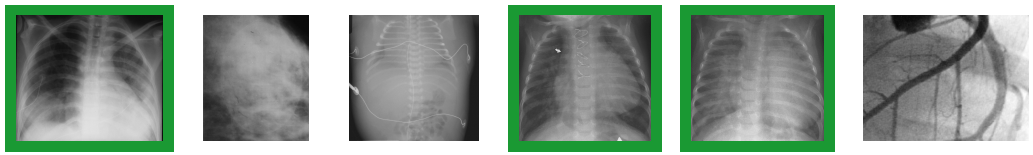
CBIR



TBIR



LSC



Discussion

- Late fusion
  - works well when both CBIR and TBIR have reasonable performance,
  - fails when CBIR performance is too poor.
- Image re-ranking
  - better than image ranking, but performs poorly compared to late and trans-media fusion
- Late fusion with semantic filtering (LSC)
  - works generally better than image reranking, late fusion and combMNZ
  - more stable than cross media similarities
  - is simple, efficient and highly scalable
- Trans-media relevance feedback (TMRF)
  - performs well when image and text experts perform similarly
  - better when re-combined with textual expert
  - it can exploit multi-modality even for mono-modal queries
  - suitable for content generation

Assisting Hybrid Content Generation - toy example

- e.g.: "simulated" Travel Blog Assistant System

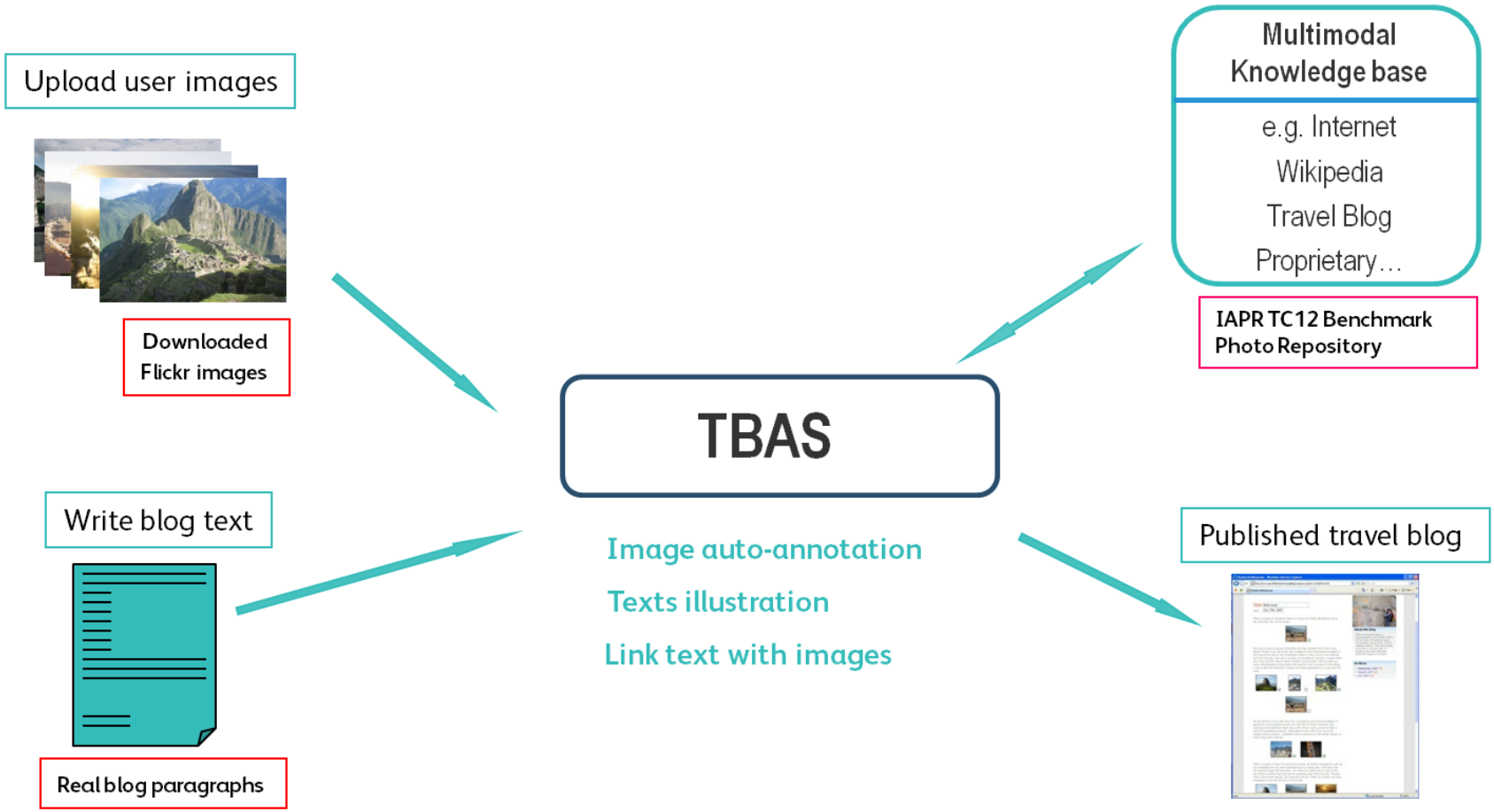
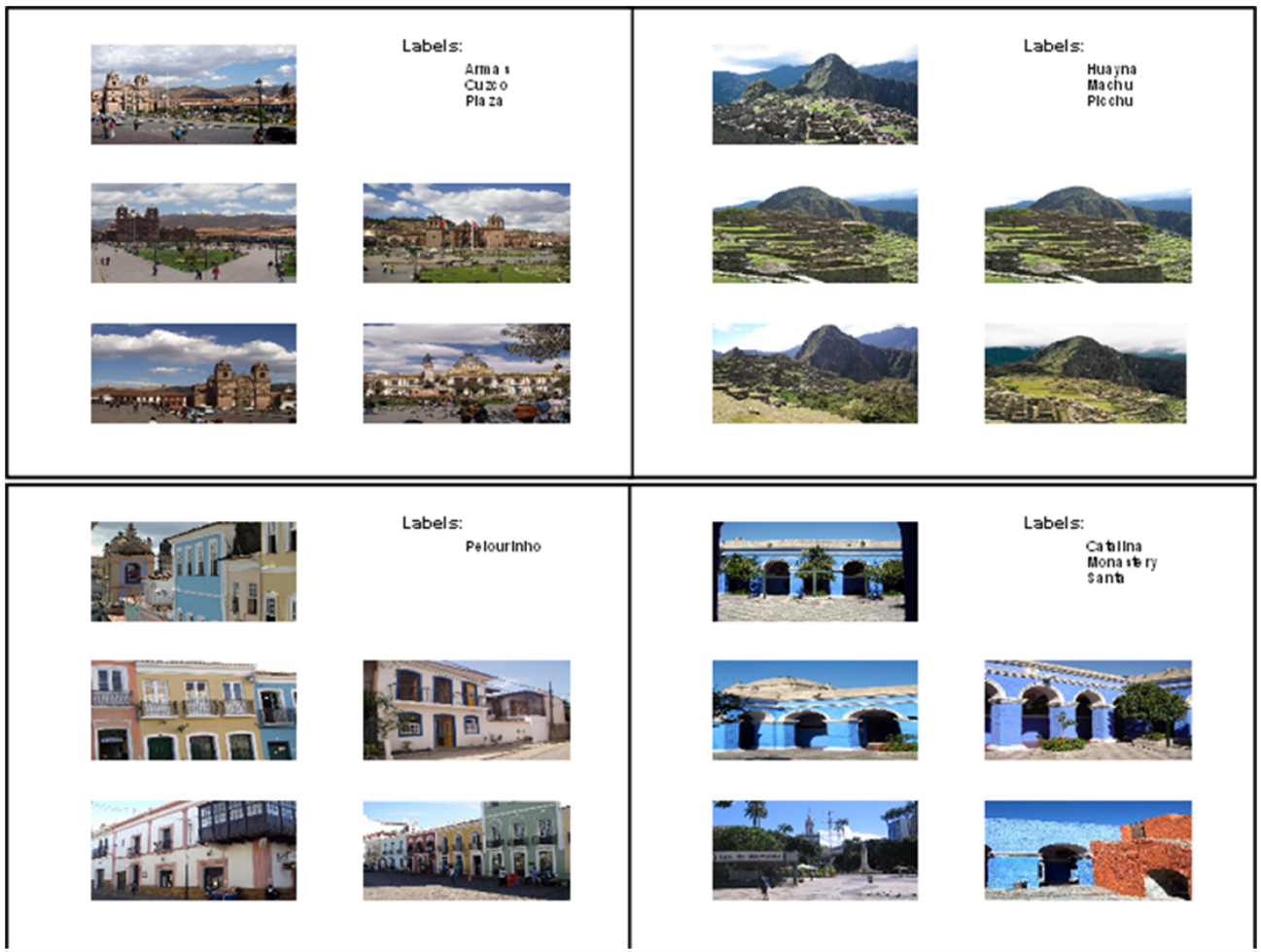


Image auto-annotation

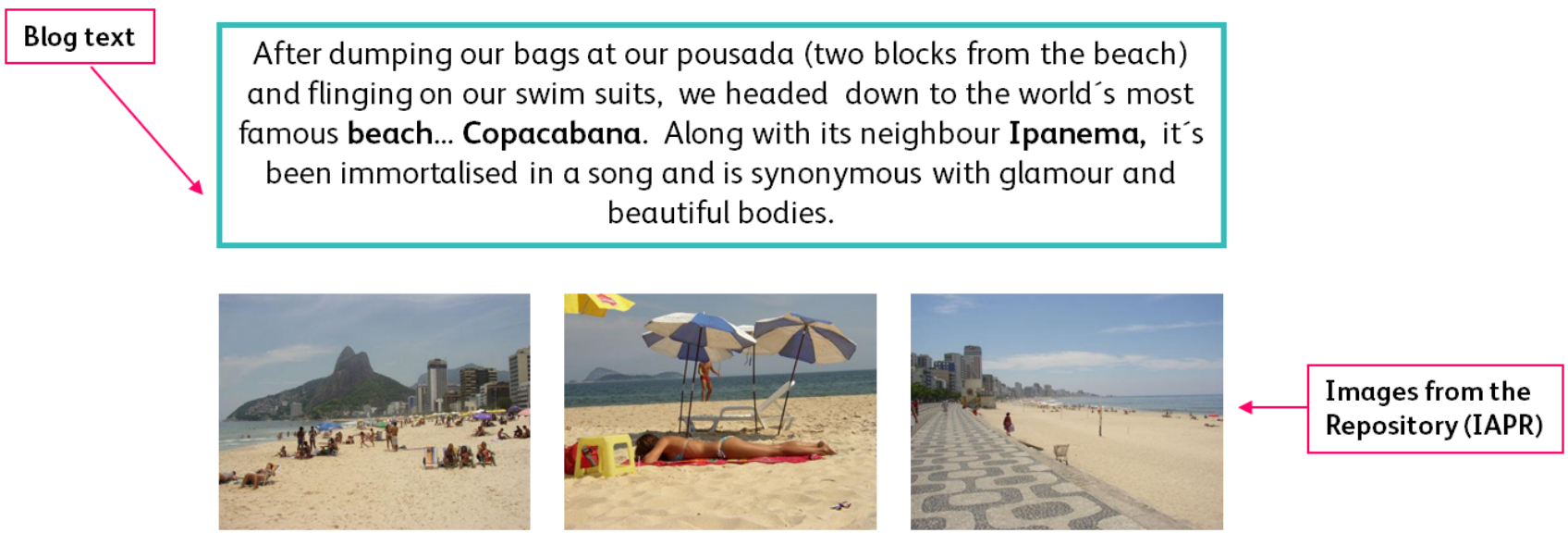
- Given an image, aggregate the texts/annotations of the similar images and extract the most frequent nouns.



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

Text illustration

- Given a text, rank a set of images in the repository according to the trans-media similarity measure.



Link independent images and text through a 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

$$sim_{VT}(I,T)$$




Flickr images

Outline

Conclusion

- Fisher Vectors (and similar BOV extentions) are :
  - are state-of-the-art *"model-dependent"*, but *"class-independent"* image representations
  - The FV can be used efficiently with the Linear Kernel
  - They allows for scalable image categorization and retrieval
- Combined with text, the multi-modal system:
  - outperforms both mono-modal (CBIR and TBIR) systems
  - using the LSC fusion, the retrieval is both efficient and highly scalable
- Hence, the FV can be successfully used in many applications:
  - Image categorization and semantic image segmentation
  - Mono and multi-modal image retrieval
  - Assisted hybrid content creation, ...