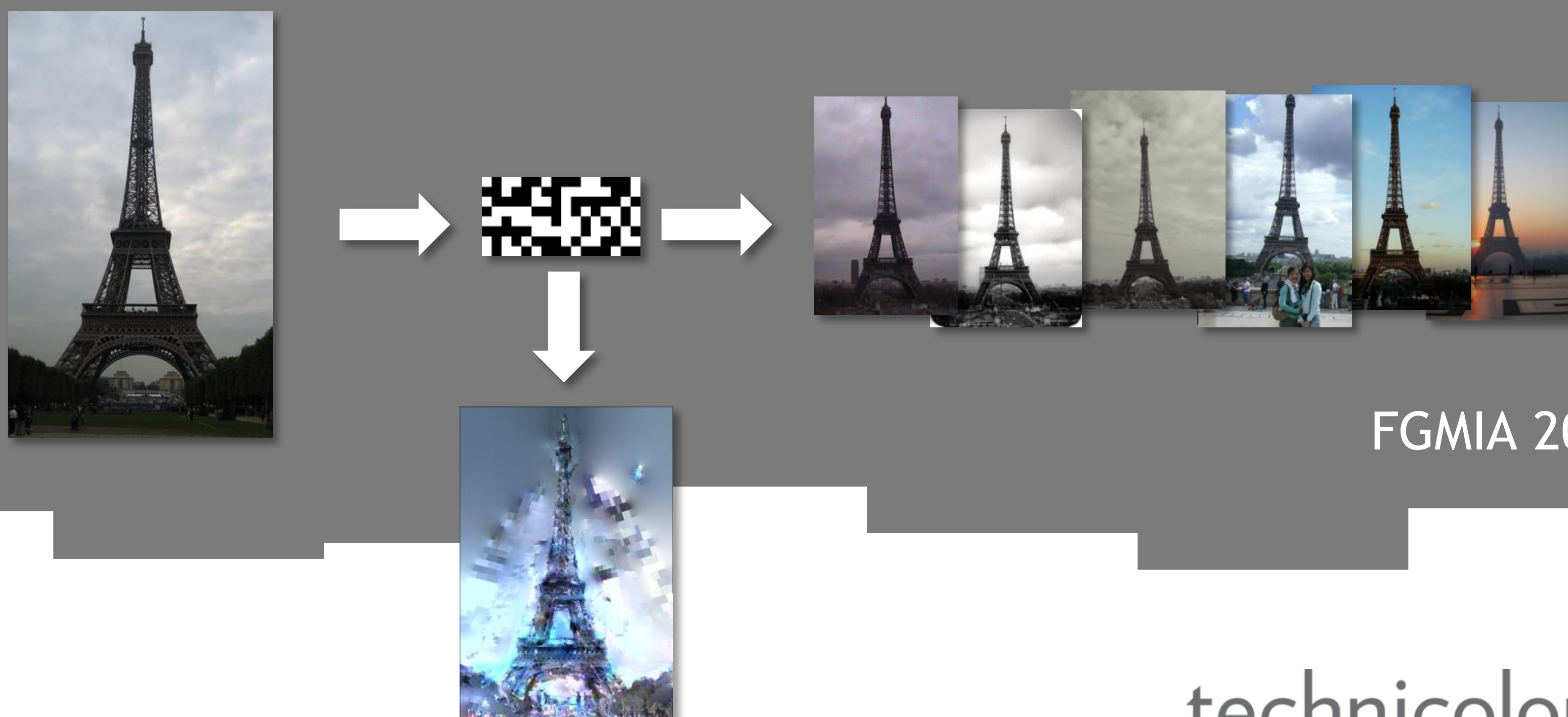
funding:{ quaero AXES Agence Nationale de la Recherche ANR }

From images to descriptors and back again

Patrick Pérez

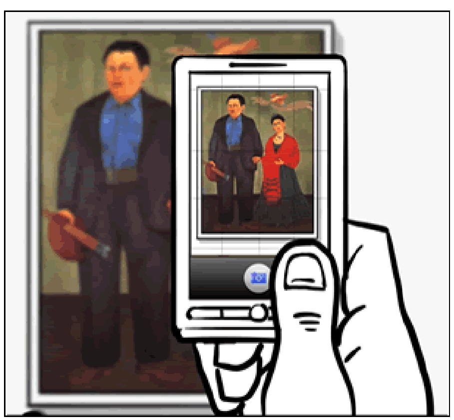


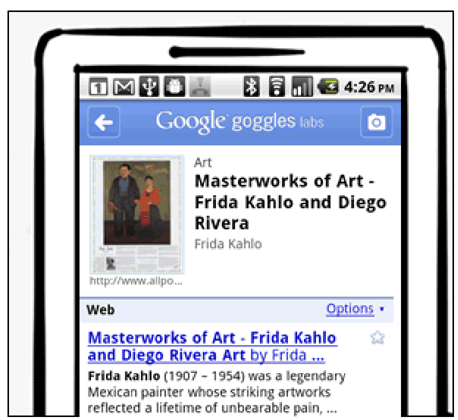
FGMIA 2014

technicolor

Visual search

- Searching in image and video databases
- One scenario: *query-by-example*
  - Input: one query image
  - Output
    - Ranked list of “relevant” visual content
    - Information on object/scene visible in query
- Some existing systems
  - Google Image and Goggles / Amazon Flow / Kooaba (Qualcom)



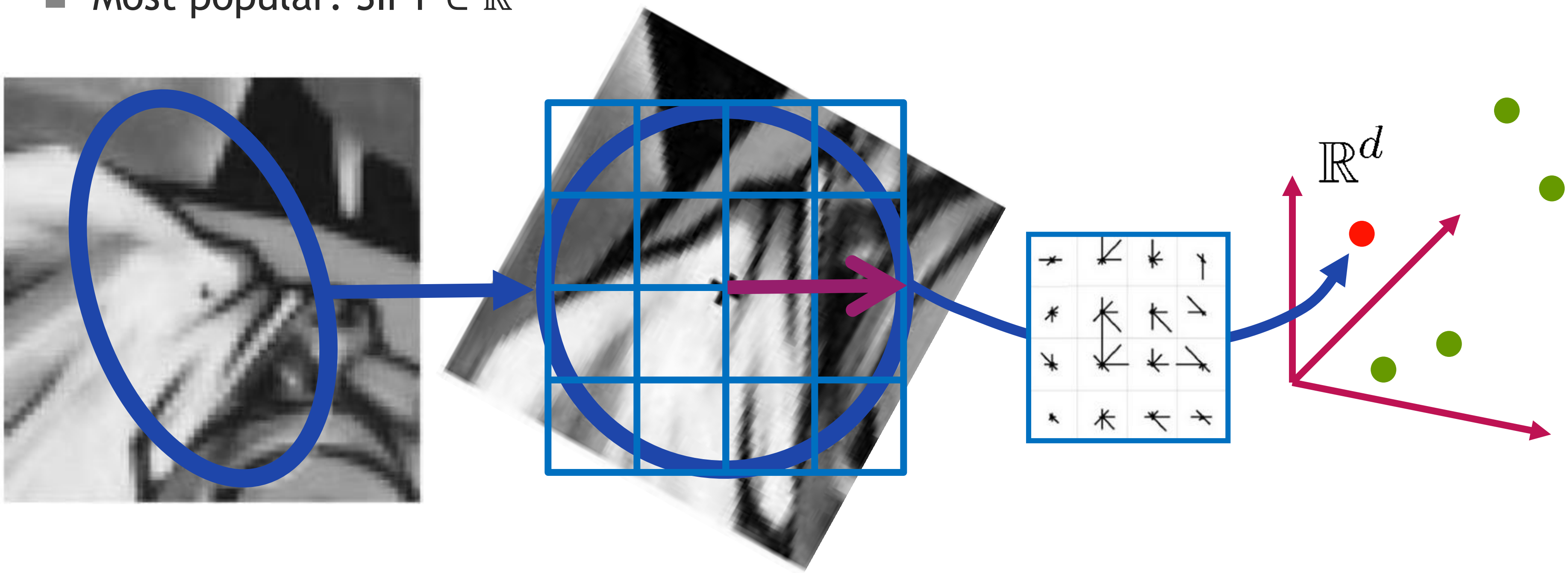


Large scale image comparison

- Raw images can’t be compared pixel-wise
  - Relevant information is lost in clutter and changes place
  - No invariance or robustness
- Meaningful and robust representation
  - Global statistics
  - *Local descriptors aggregated in a global signature*
- Efficient approximate comparisons

Local descriptors

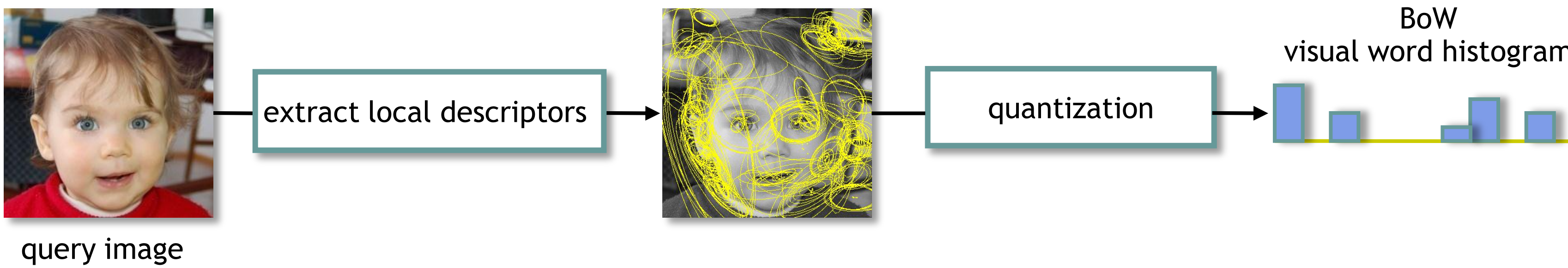
- Select/detect image fragments, normalize and describe them
  - Robust to some geometric and photometric changes
  - Most popular: SIFT $\in \mathbb{R}^{128}$



- Precise image comparison: match fragments based on descriptors
  - Works very well … but way too expensive on a large scale

[Mikolajczyk , Schmid. IJCV 2004]
[Lowe. IJCV 2004]

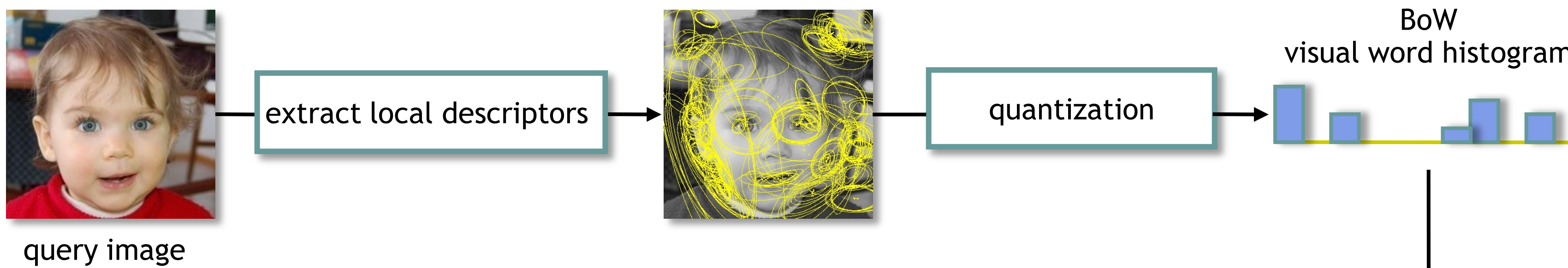
Bag of “Visual Words” pipeline



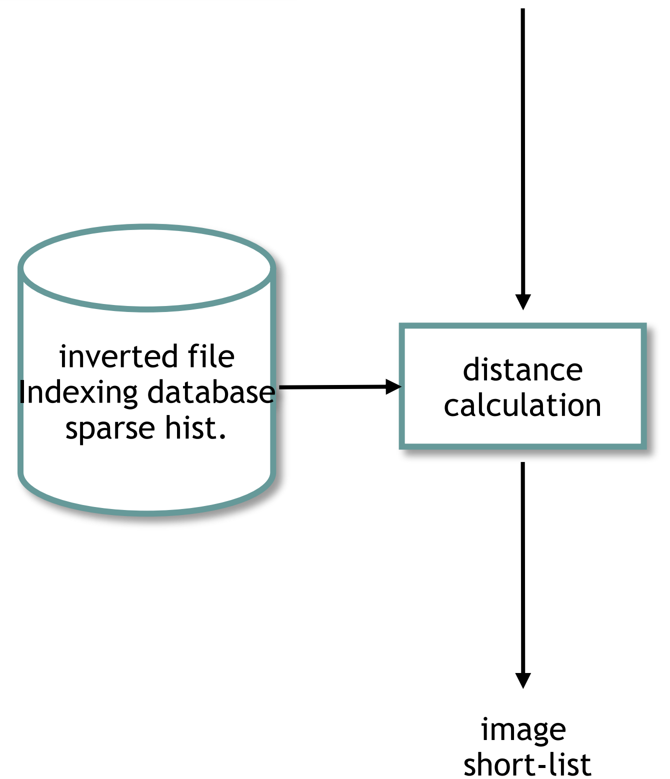
- Forget about precise descriptors
  - Vector-quantization using a dictionary of k “visual words” learned off-line
- Forget about fragment location
  - Counting visual words
- BoW: *sparse fixed size signature by aggregation* of a variable number of quantized local descriptors

[Sivic, Zisserman. ICCV 2003][Csurca *et al*. 2004]

Bag of “Visual Words” pipeline

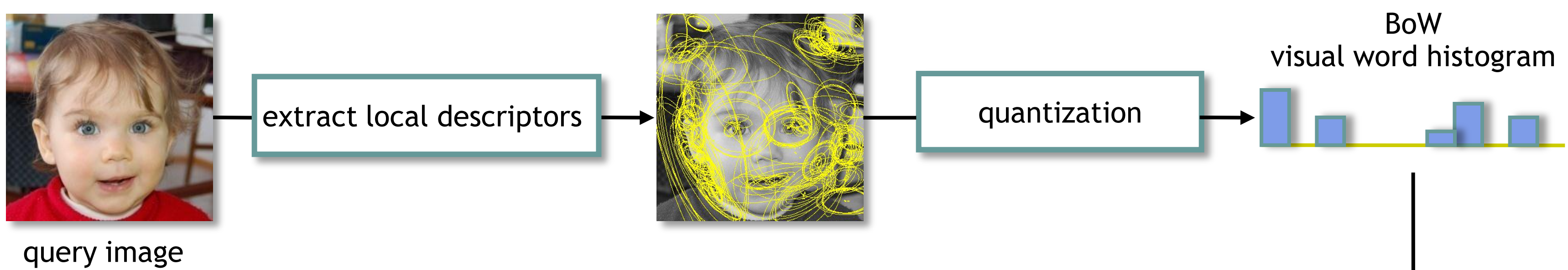


- Efficient search with *inverted files*
  - Search only images that share words with query
  - Short-listing based on histogram distance

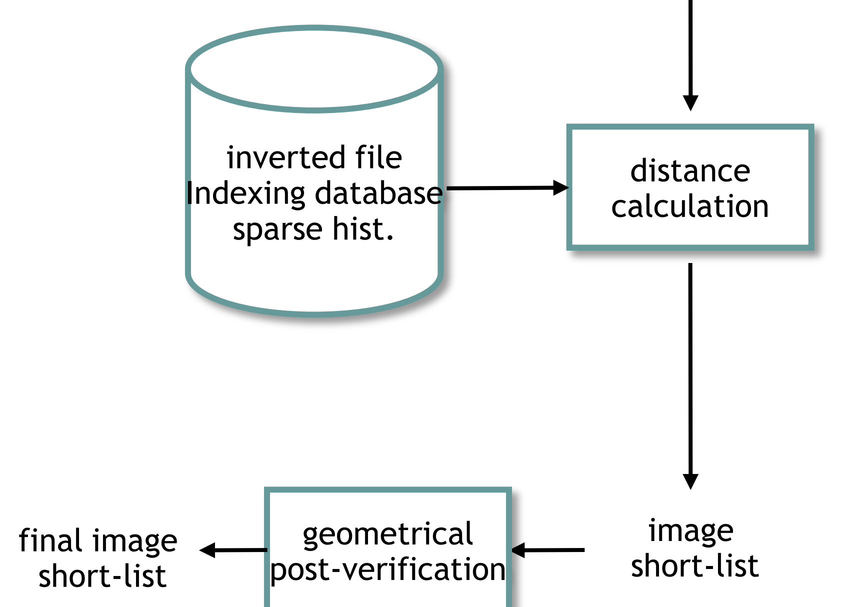


[Sivic, Zisserman. ICCV 2003]

Bag of “Visual Words” pipeline



- Geometrical post-verification
  - Match local features
  - Infer most likely geometric transform
  - Rank short list based on goodness-of-fit



[Sivic, Zisserman. ICCV 2003]

technicolor

Limitations and contributions

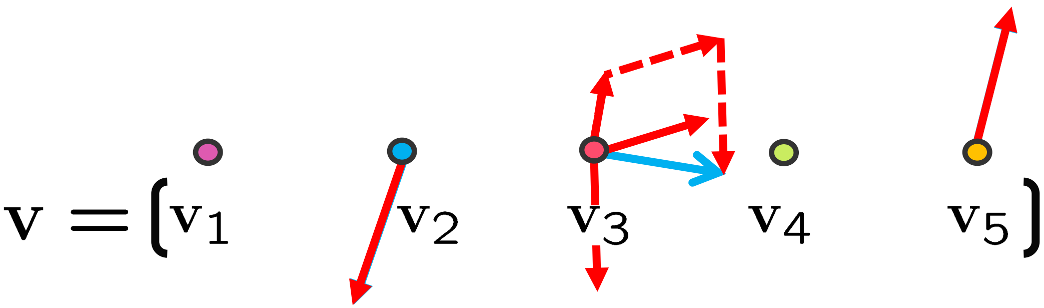
- Precise search requires large dictionary (k ~20,000-200,000 words)
  - Difficult to learn
  - Costly to compute (k distances per descriptor) on database
  - Memory footprint still too large (~10KB per image)
    - With 40GB RAM, search 10M images in 2s
    - Does not scale up to web-scale ($\propto 10^{11}$ images)
- Contribution*
  - Novel aggregation of local descriptors into image signature
  - Combined with efficient indexing
    - Low memory footprint (20B per image, 200MB RAM for 10M images)
    - Fast search (50ms to search within 10M images on laptop)

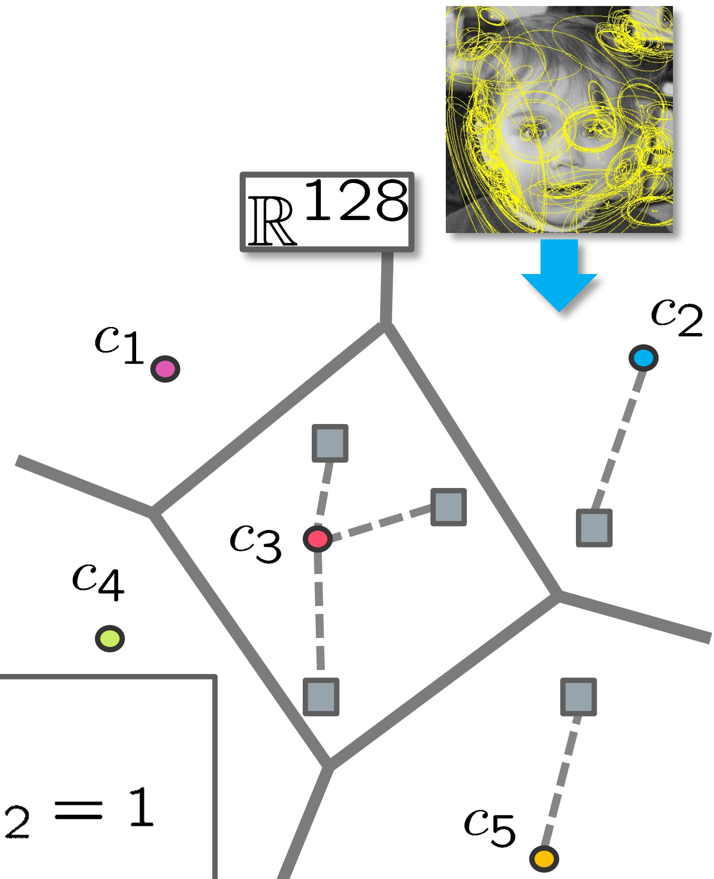
*[Jégou, Douze, Schmid, Pérez. CVPR 2010]

technicolor

Beyond cell counting

- Vector of Locally Aggregated Descriptors (VLAD)
  - Very coarse visual dictionary (e.g., $k = 64$): $\mathcal{C} = \{c_1, \cdots c_k\} \in \mathbb{R}^{128}$
  - But characterize distribution in each cell

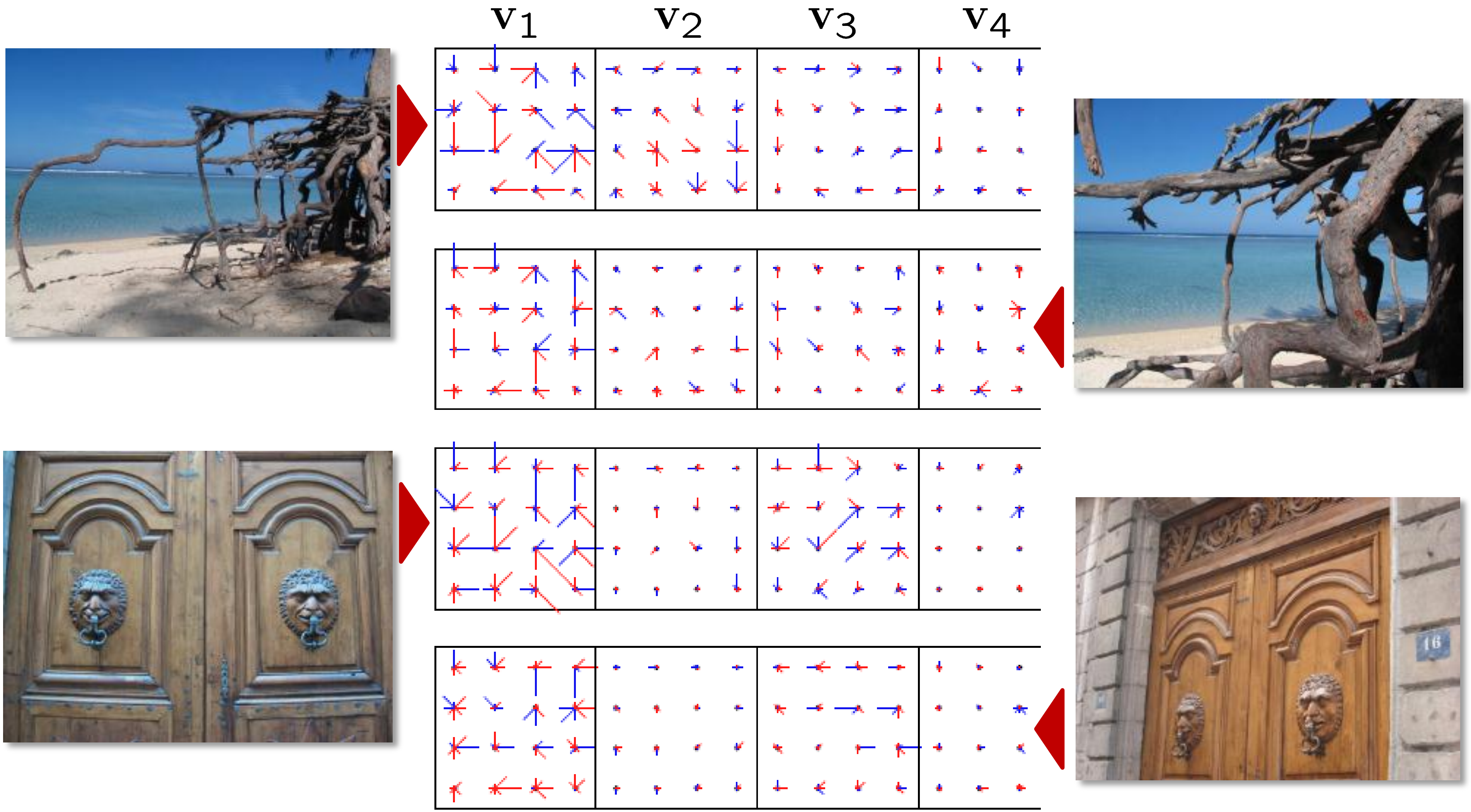




$$\mathbf{v} \propto \begin{bmatrix} \mathbf{v}_1 \\ \vdots \\ \mathbf{v}_k \end{bmatrix} \in \mathbb{R}^{128k}, \ \mathbf{v}_i = \sum_{\mathbf{x} \in \text{cell } i} (\mathbf{x} - c_i), \ \|\mathbf{v}\|_2 = 1$$

VLAD

- Vectors of size $D = 128 \times k$, k SIFT-like blocks



technicolor

Fisher interpretation

- Given parametric family of pdfs $\{p_\theta,\ \theta \in \Theta \subset \mathbb{R}^u\}$
  - Fisher information matrix (size u)

$$F_\theta = \mathbb{E}_{p_\theta}[\nabla_\theta \ln p_\theta \nabla_\theta^T \ln p_\theta]$$

  - Log-likelihood gradient of sample $\{\mathbf{x}_n\}_{n=1\cdots N}$

$$G_\theta(\{\mathbf{x}_n\}) = \frac{1}{N}\sum_{j=n}^{N} \nabla_\theta \ln p_\theta(\mathbf{x}_n)$$

- Fisher kernel: given θ, compare two samples

$$\begin{aligned} K_\theta(\{\mathbf{x}_m\},\{\mathbf{y}_n\}) &= G_\theta(\{\mathbf{y}_m\})^\top F_\theta^{-1} G_\theta(\{\mathbf{x}_n\}) \\ &= \langle F_\theta^{-\frac{1}{2}} G_\theta(\{\mathbf{y}_m\}), \underbrace{F_\theta^{-\frac{1}{2}} G_\theta(\{\mathbf{x}_n\})}_{\mathcal{G}_\theta(\{\mathbf{x}_n\})} \rangle \end{aligned}$$

  - Dot product of *Fisher vectors* (FV)

[Jaakkola, Haussler. NIPS 1998][Perronnin *et al*. CVPR 2011]

technicolor

VLAD and Fisher vector

- Example: spherical GMM with parameters $\theta = (\{\pi_i, \boldsymbol{\mu}_i, \sigma_i\})_{i=1\cdots k}$
  - *Approximate* FV on mean vectors only

$$\mathcal{G}_{\boldsymbol{\mu}_i}(\{\mathbf{x}_n\}) = \frac{1}{N\sqrt{\pi_i}} \sum_{n=1}^{N} \kappa_n(i)\sigma_i^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_i),\ i = 1 \cdots k$$

  with *soft assignments* $\kappa_n(i)$. FV of size $D = d \times k$

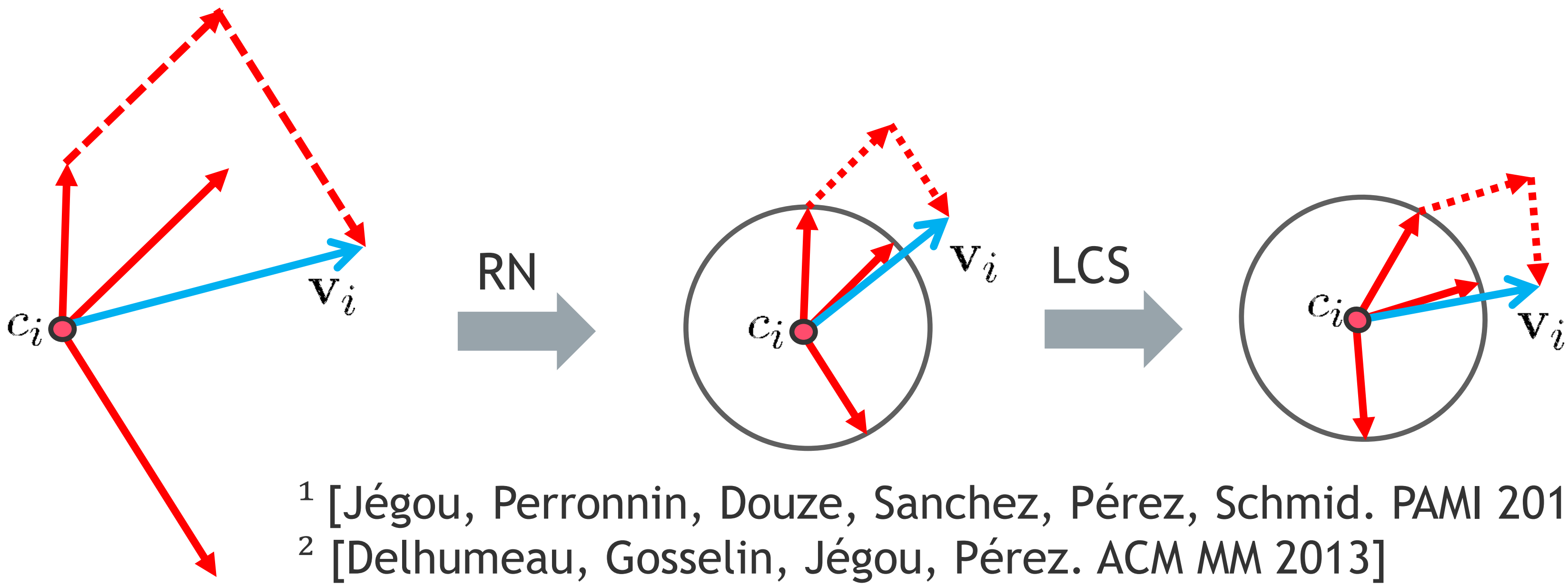
  - If equal weights and variances, hard assignment to code-words, FV = VLAD

$$\mathcal{G}_{\boldsymbol{\mu}_i}(\{\mathbf{x}_n\}) \propto \mathbf{v}_i(\{\mathbf{x}_n\}),\ i = 1 \cdots k$$

technicolor

Additional tricks

- Power-law[1] $v_j \leftarrow \mathrm{sign}(v_j)|v_j|^\alpha,\ j = 1 \cdots D,\ \alpha \in (0,1)$
- Residue normalization ("RN")[2]
$$\mathbf{v}_i = \sum_{\mathbf{x} \in \text{cell } i} \frac{\mathbf{x} - c_i}{\|\mathbf{x} - c_i\|_2},\ i = 1 \cdots k$$
- Intra-cell PCA local coordinate system ("LCS")[2]
$$\mathbf{v}_i = R_i \sum_{\mathbf{x} \in \text{cell } i} \frac{\mathbf{x} - c_i}{\|\mathbf{x} - c_i\|_2},\ i = 1 \cdots k$$
- RootSift ("$\sqrt{SIFT}$")[3]



[1] [Jégou, Perronnin, Douze, Sanchez, Pérez, Schmid. PAMI 2012]
[2] [Delhumeau, Gosselin, Jégou, Pérez. ACM MM 2013]
[3] [Arandjelovic , Zisserman. CVPR 2013]

Exhaustive search

- Comparisons to BoW on Holidays (1500 images with relevance GT)

Image signature	dim	mAP (%)
BoW-20K	20,000	43.7
BoW-200K	200,000	54.0
VLAD-64	8192	51.8
$+\ \alpha = 0.2$		54.9
$+\ \sqrt{SIFT}$		57.3
+ RN		63.1
+ LCS		65.8
+ *dense SIFTs*		76.6

Getting short and compact

- Towards large scale search
  - PCA reduction of image signature to $D' = 128$
  - Very fine quantization with *Product Quantizer* (PQ)*
  - Results on *Oxford105K* and *Holydays*+1M Flickr distractors

Image signature	Ox105K	Hol+1M
Best VLAD-64 (8192 dim)	45.6	–
Reduced (128 dim)	26.6	39.2
Quantized (16 bytes)	22.2	32.3

*[Jégou, Douze, Schmid. PAMI 2010]

technicolor

Quantized signatures

- Vector quantization on k_f values

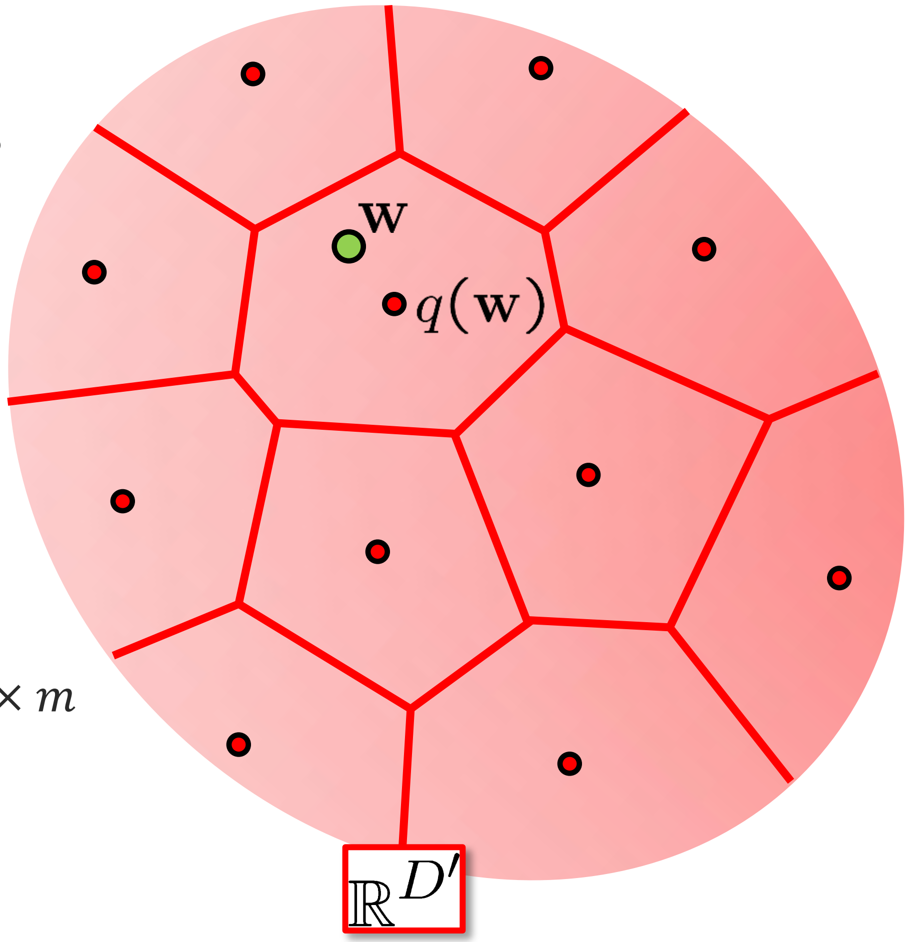
$$\mathbf{w} \approx q(\mathbf{w})$$

- For good approximation, large codes
  - e.g., 128 bits ($k_f = 2^{128}$)
- Practical with *product quantizer**

$$\mathbf{w} = \begin{bmatrix} \mathbf{w}_1 \\ \vdots \\ \mathbf{w}_m \end{bmatrix}, \quad q(\mathbf{w}) = \begin{bmatrix} q_1(\mathbf{w}_1) \\ \vdots \\ q_m(\mathbf{w}_m) \end{bmatrix}$$

with k_r values per sub-quantizer

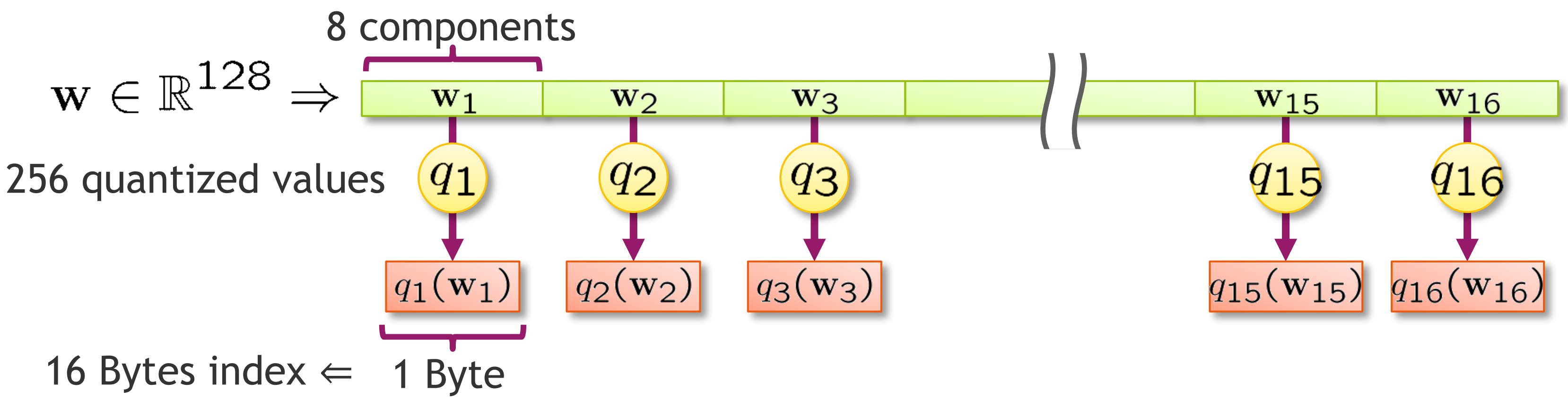
- yields $k_f = (k_r)^m$ with complexity $k_r \times m$



*[Jégou, Douze, Schmid. PAMI 2010]

technicolor

Quantized signatures



$$D' = 128$$
$$m = 16$$
$$k_r = 2^8$$
$$k_f = 2^{128}$$

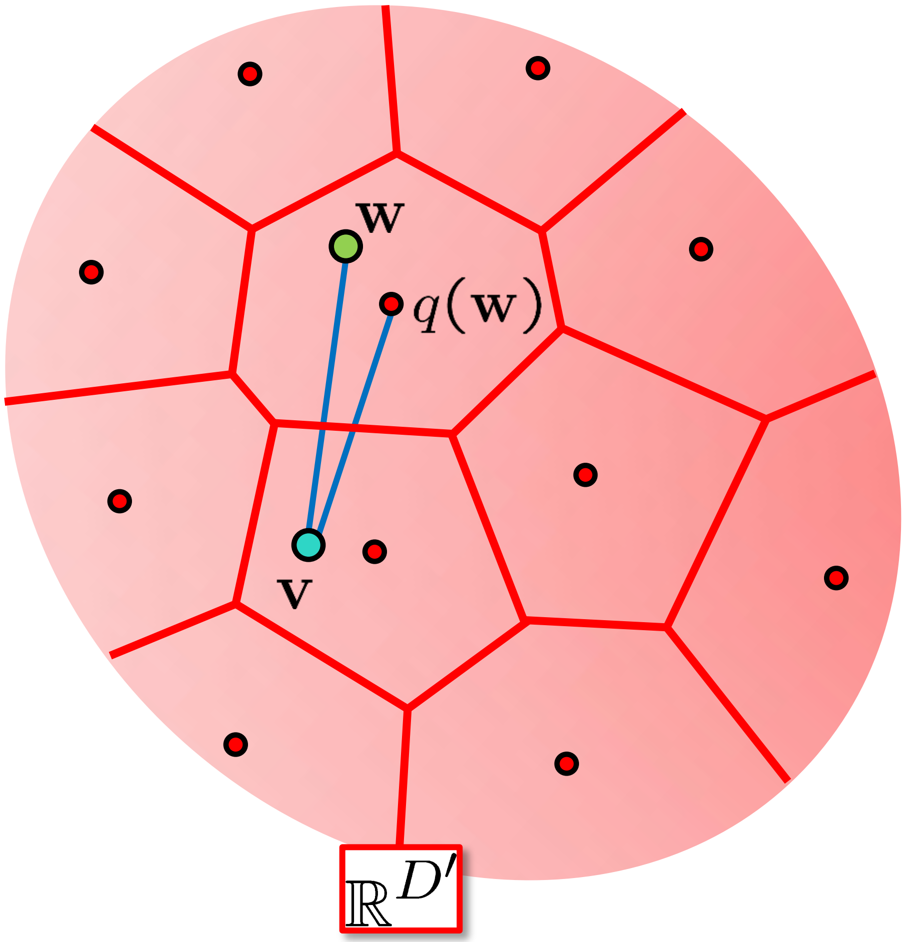
Asymmetric Distance Computation (ADC)

- Given query signature **v**, distance to a basis signature **w**:

$$\|\mathbf{v}-\mathbf{w}\|^2 \approx \sum_{i=1}^{m} \underbrace{\|\mathbf{v}_i - q_i(\mathbf{w}_i)\|^2}_{k_r \text{ possible values}}$$

- Exhaustive search among N_b basis images

$\boxed{mk_r \text{ distances } + (m-1)N_b \text{ sums}}$



ADC with Inverted Files (IVF-ADC)

- Two-level quantization of signatures
  - Coarse quantization (e.g., $k_c = 2^8$ values)
  - One inverted list per code-vector
  - Compare only within lists of w nearest code-vectors to query
  - Fine PQ quantization of *residual* signatures (e.g., $k_f = 2^{128}$)

- Search among N_b basis images

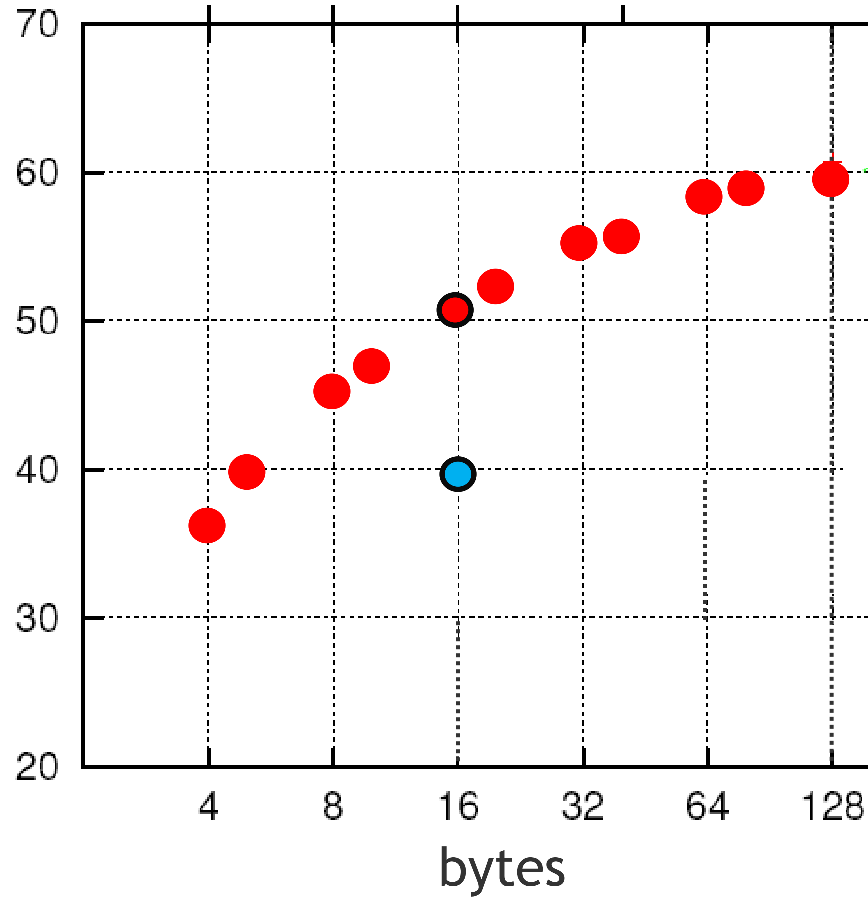
$$\boxed{mk_r \text{ distances} + w(m-1)N_b k_c^{-1} \text{ sums}}$$

$w = 16, m = 16, k_r = k_c = 256 \Rightarrow$ one sum only per image with almost no accuracy change!

technicolor

Performance w.r.t. memory footprint

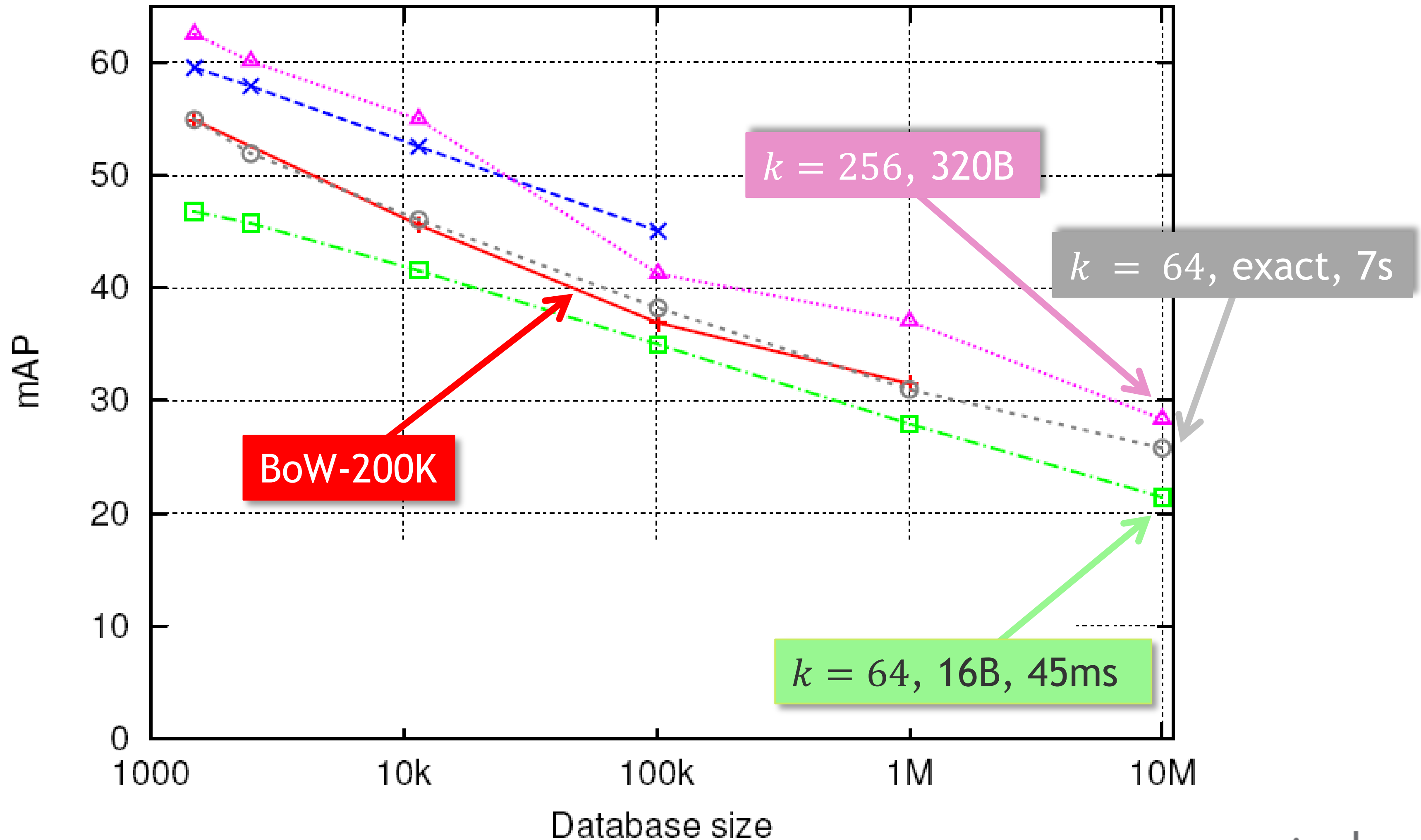
Image signature	bytes	mAP (%)
BoW-20K	10,364	43.7
BoW-200K	12,886	54.0
FV-64		59.5
▪ Spectral Hashing* 128 bits	16	39.4
▪ PQ, $m = 16, k_r = 256$	16	50.6



*[Weiss *et al*. NIPS 2008]

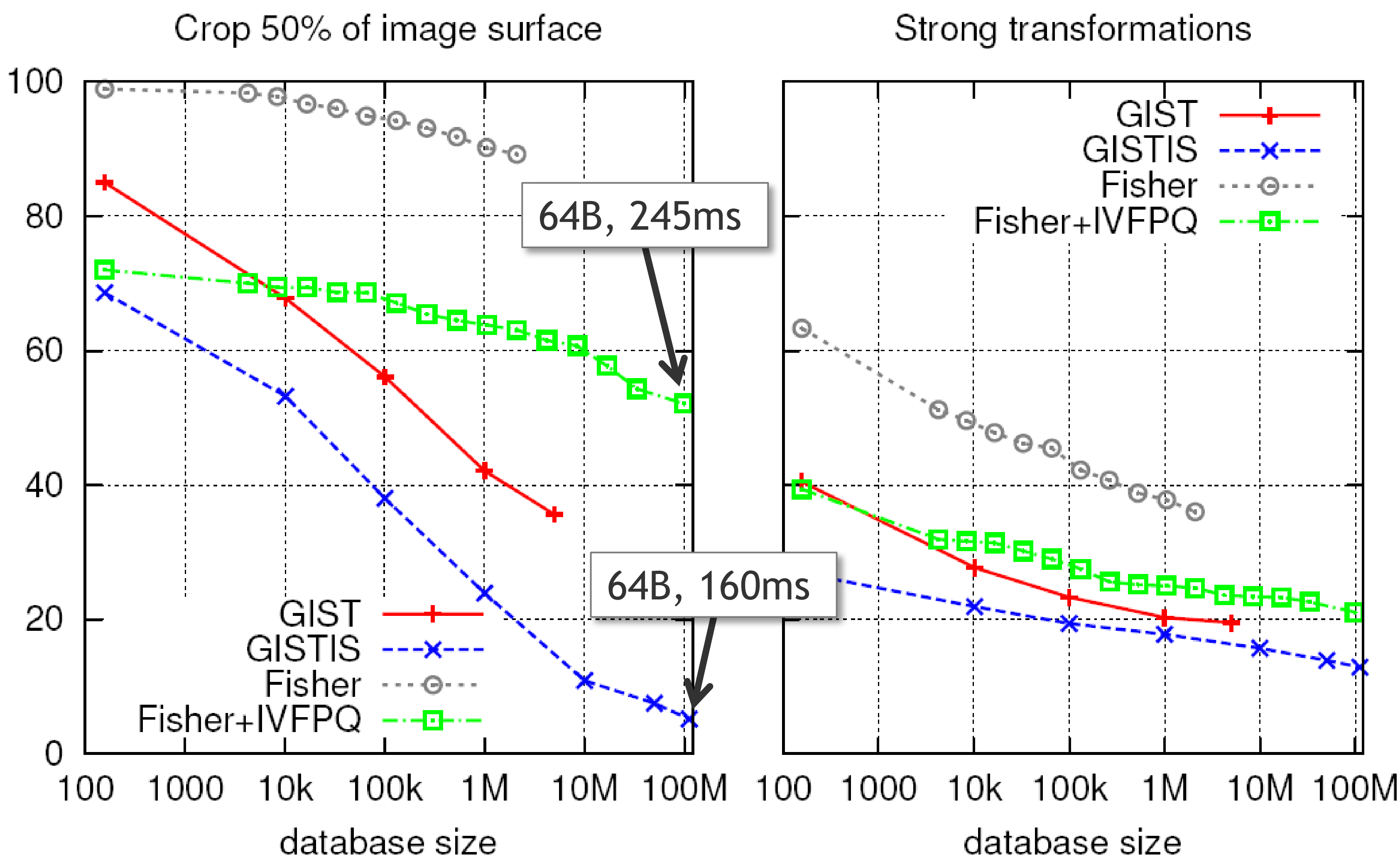
Large scale experiments

- Holidays + up to 10M distractors from Flickr



Larger scale experiments

- Copydays + up to 100M distractors from Exalead



[GIST: Oliva, Torralab. PBR 2006][GISTIS: Douze *et al*. AMC-MM 2009]

technicolor

Beyond Euclidean distance

- Kernel-based similarities
  - Other better but costly kernels
  - For histogram-like signatures: Chi2, histogram intersection (HIK)
- *Explicit embedding* recently proposed for learning[1]
  - Given PSD kernel function $K(\mathbf{x},\mathbf{y}) = \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle$
  - Find an *explicit finite dim.* approximation of implicit feature map

$$K(\mathbf{x},\mathbf{y}) \approx \langle \tilde{\phi}(\mathbf{x}), \tilde{\phi}(\mathbf{y}) \rangle$$

  - Learn linear SVM in this new explicit feature space
  - KCPA[2]: a flexible data-driven explicit embedding

$$[K(\mathbf{x}_i,\mathbf{x}_j)] = U \Lambda U^\top \approx U_{[E]} \mathrm{diag}(\lambda_1, \cdots, \lambda_E) U_{[E]}^\top,\ E < N$$

- What about search?

[1] [Vedaldi, Zisserman. CVPR 2010][Perronnin *et al*. CVPR 2010]
[2] [Schölkopf *et al*. ICANN 1997]

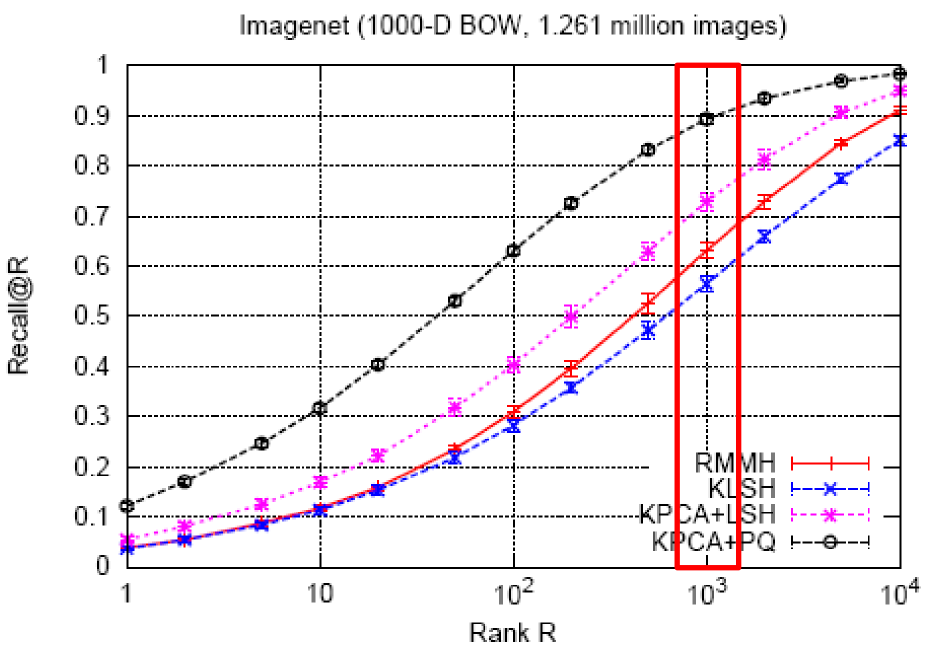
Approximate search with short codes

- Simple proposed approach* ("KPCA+PQ")
  - Embed database vectors with learned KPCA
  - Efficient Euclidean ANN with PQ coding
  - Kernel-based re-ranking in original space
- Competitors: binary search in implicit space
  - Kernelised Locally Sensitive Hashing (KLSH) [Kulis, Grauman. ICCV09]
  - Random Maximum Margin Hashing (RMMH) [Joly, Buisson. CVPR11]
- Experiments
  - Data: 1.2M images from ImageNet with BoW signatures
  - Chi2 similarity measure
  - Tested also: "KPCA+LSH"(binary search in explicit space)

*[Bourrier, Perronnin, Gribonval, Pérez, Jégou. TR 2012]

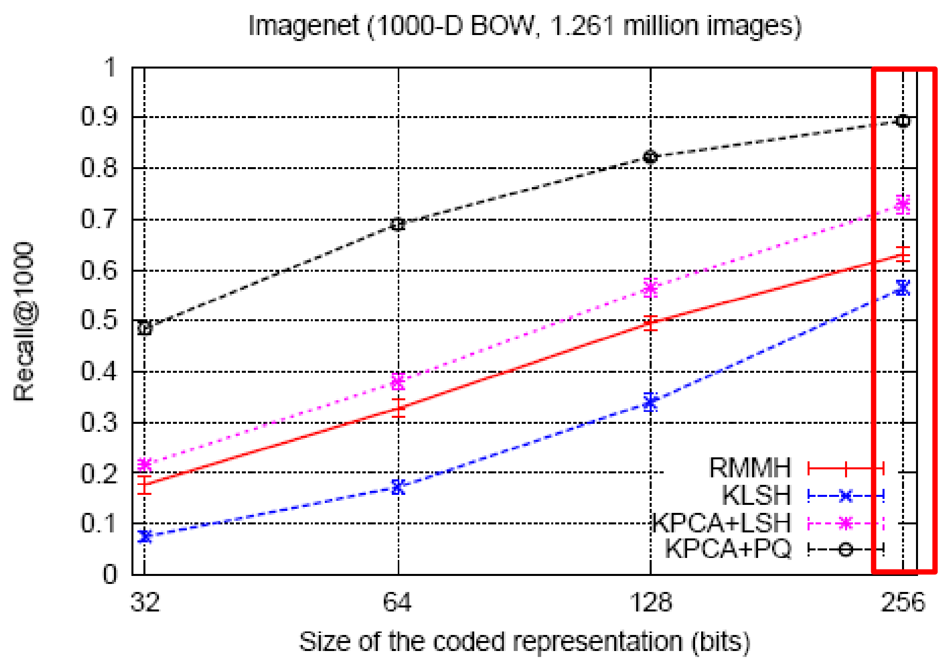
technicolor

Results averaged over 10 runs



Recall@R

$E = 128, B = 256 \text{ bits}, M = 1024$



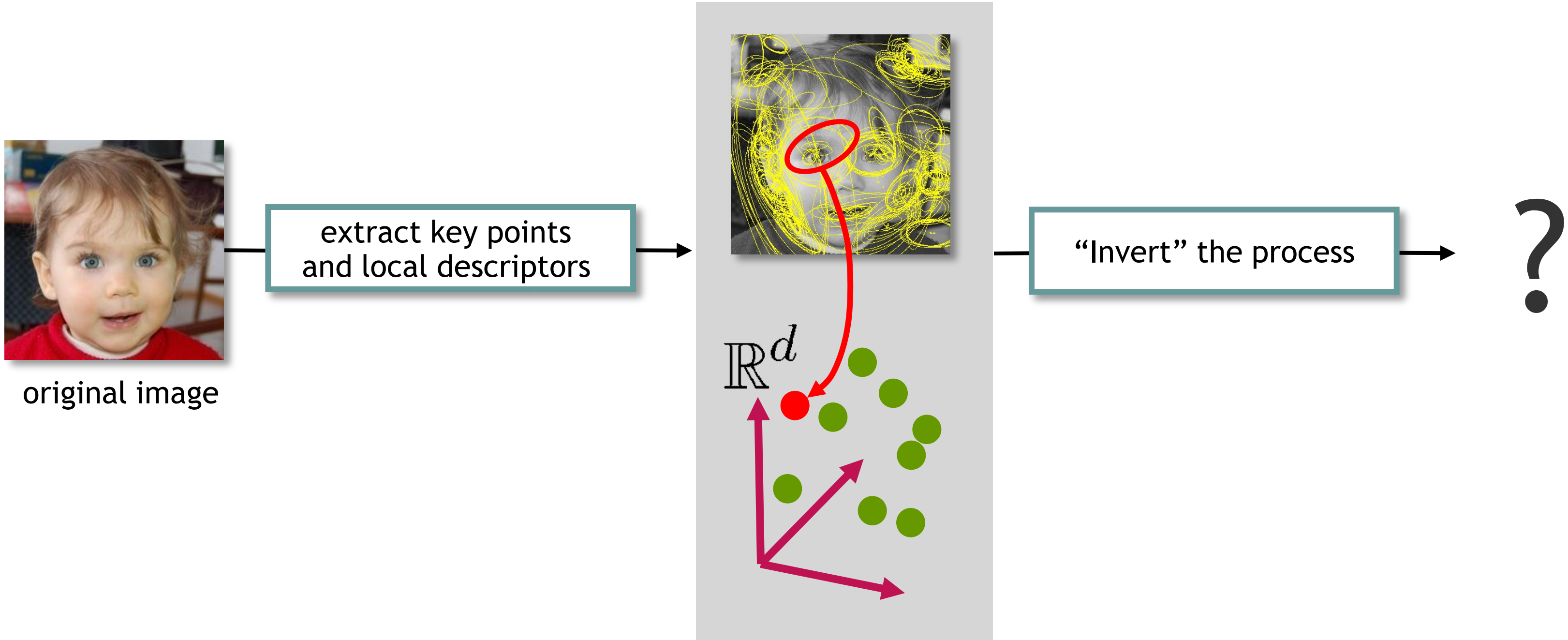
Recall@1000

$B = 32 \rightarrow 256 \text{bits}$

technicolor

Reconstructing an image from descriptors

- If sparse local descriptors only are known

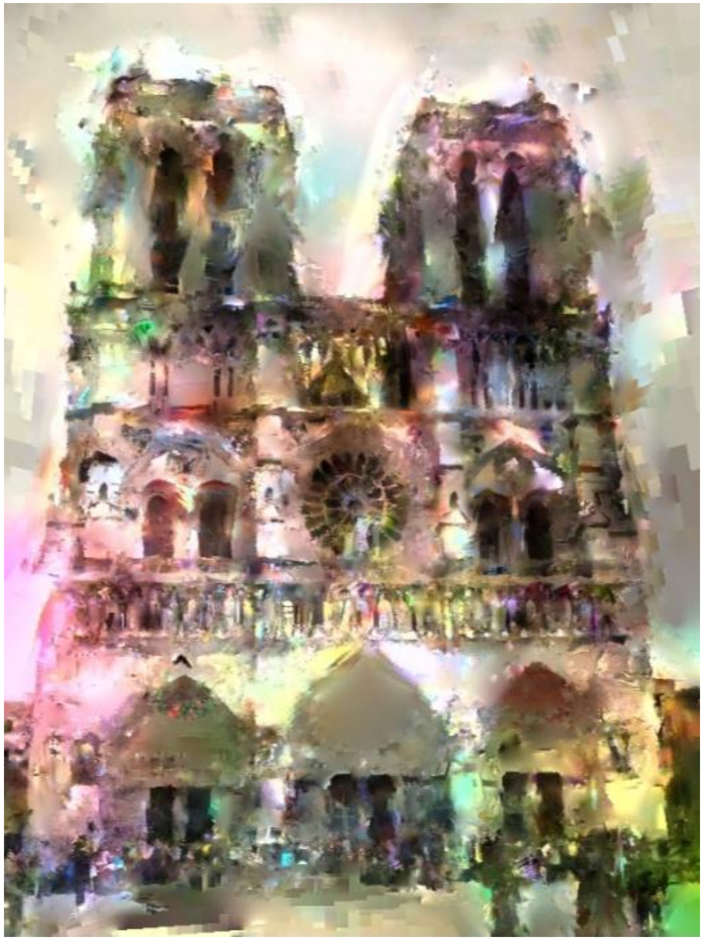


- Better insight into what local descriptors capture, with multiple applications

technicolor

Reconstructing an image from descriptors

- Possible to some extent

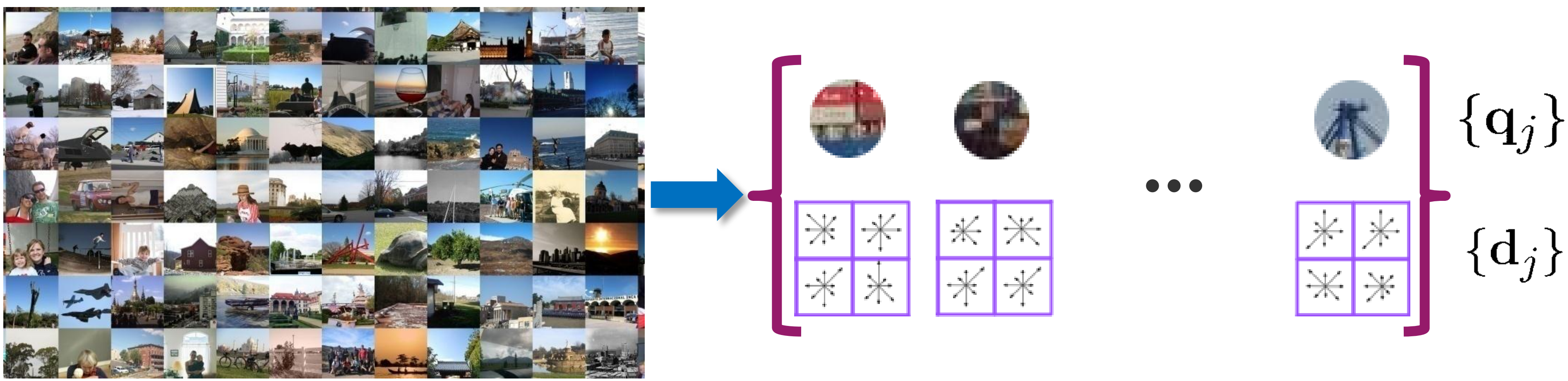


[Weinzaepfel, Jégou, Pérez. CVPR'2011]

technicolor

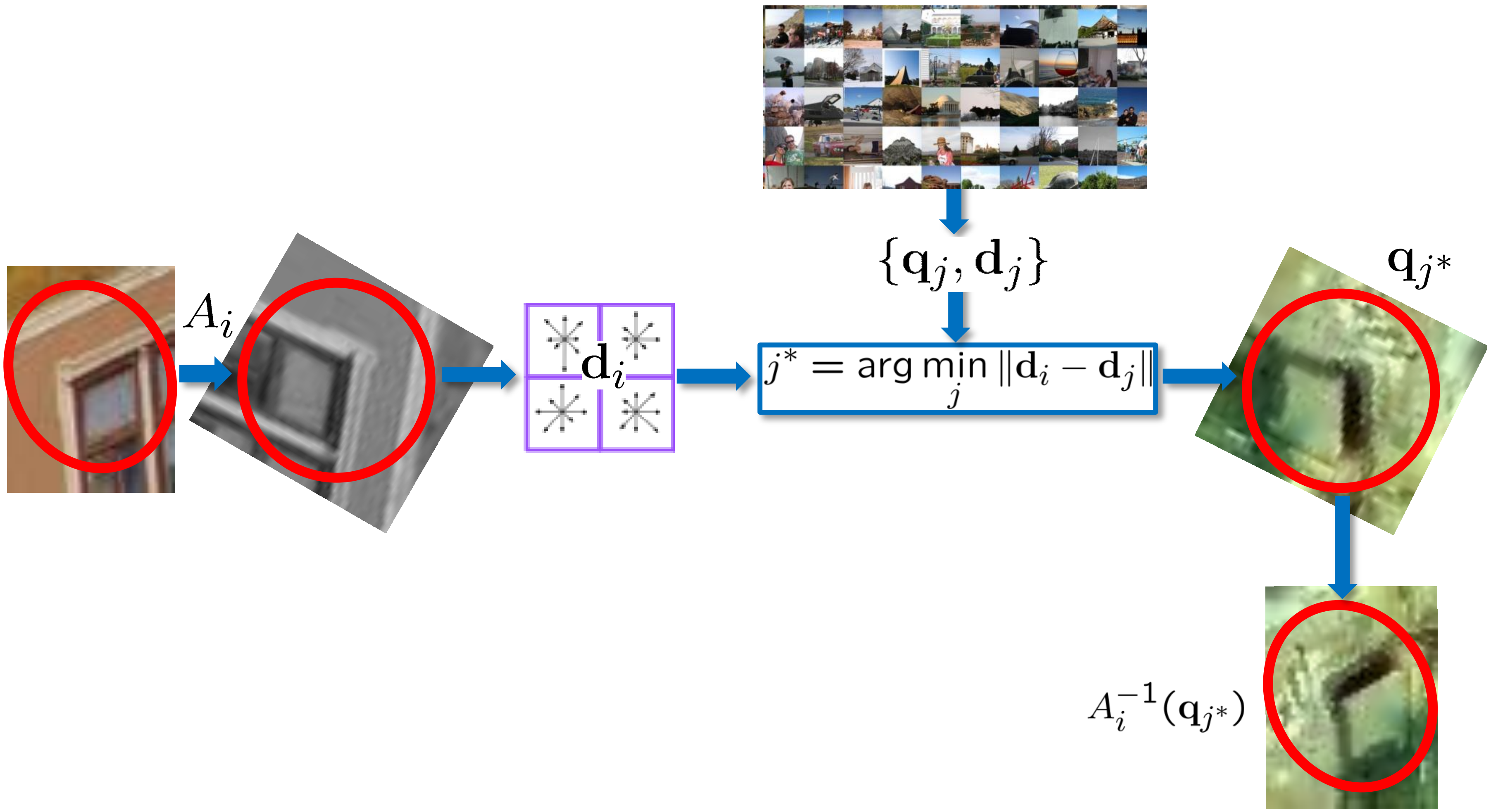
Inverting local description

- Local description, severely lossy by construction
  - Color, absolute intensity, spatial arrangement in each cell are lost
  - Non-invertible many-to-one map
  - Example-based regularization: *use key-points from **arbitrary** images*



  - Patch collection must be large and diverse enough (e.g., 6M)

Inverting local description



Assembling recovered patches

- Progressive collage
  - Dead-leaf procedure, largest patches first

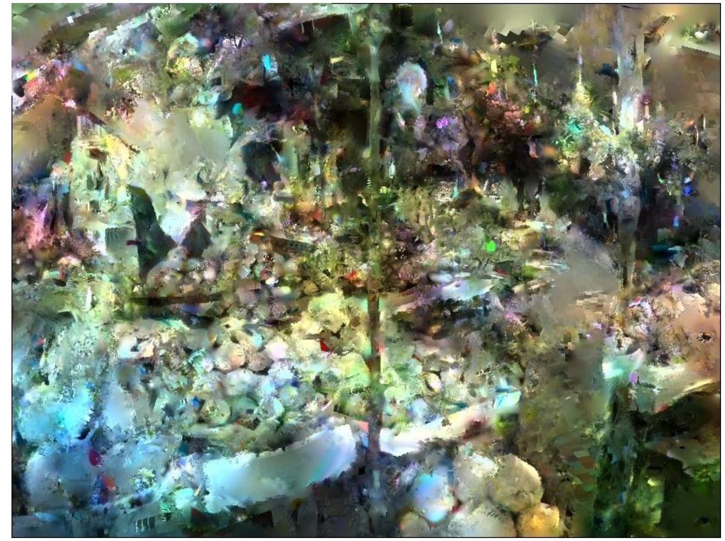
  - Seamless cloning*
    - Harmonic correction: smooth change to remove boundary discrepancies
  - Final hole filling
    - Harmonic interpolation

*[Pérez, Gangnet, Blake. Siggraph 2003]

Reconstruction

Reconstruction

Reconstruction



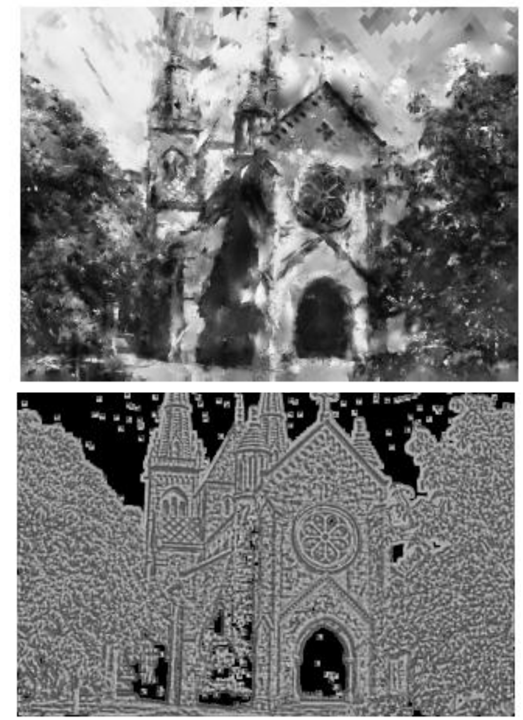
technicolor

Outlook

- New: reconstruction from dense local features

HOG (Hoggles)[2]

local binary pattern[1]



- Human-understandable images can be reconstructed
  - Visual insight into information exploited by detectors and classifiers
  - *Visual information leakage* in image indexing systems: privacy?

[1] [D'Angelo, Alahi, P. Vandergheynst. ICPR 2012]
[2] [Vondrick, Khosla, Malisiewicz, Torralba. ICCV 2013]

technicolor