Image descriptor aggregation for efficient retrieval

Jean-Baptiste Boin

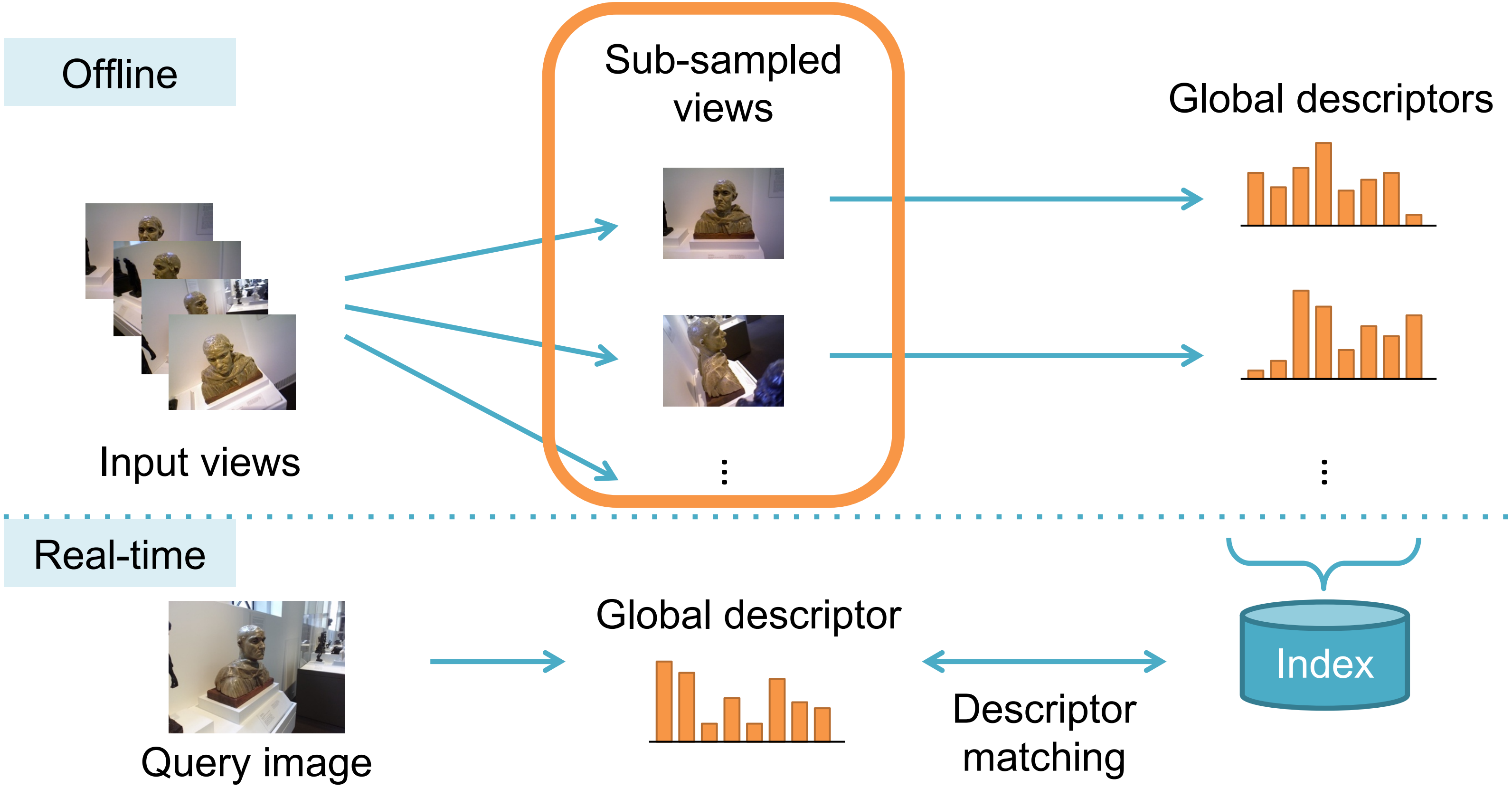
Department of Electrical Engineering
jbboin@stanford.edu

Visual search (Content-based image retrieval)

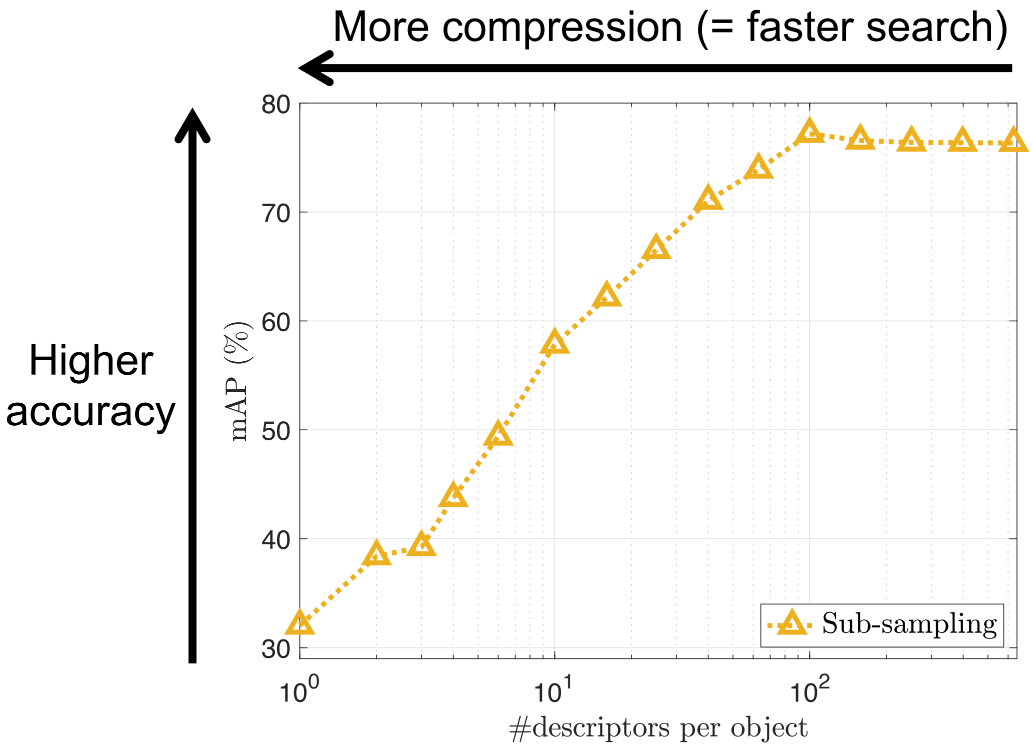
Image query

Database of images

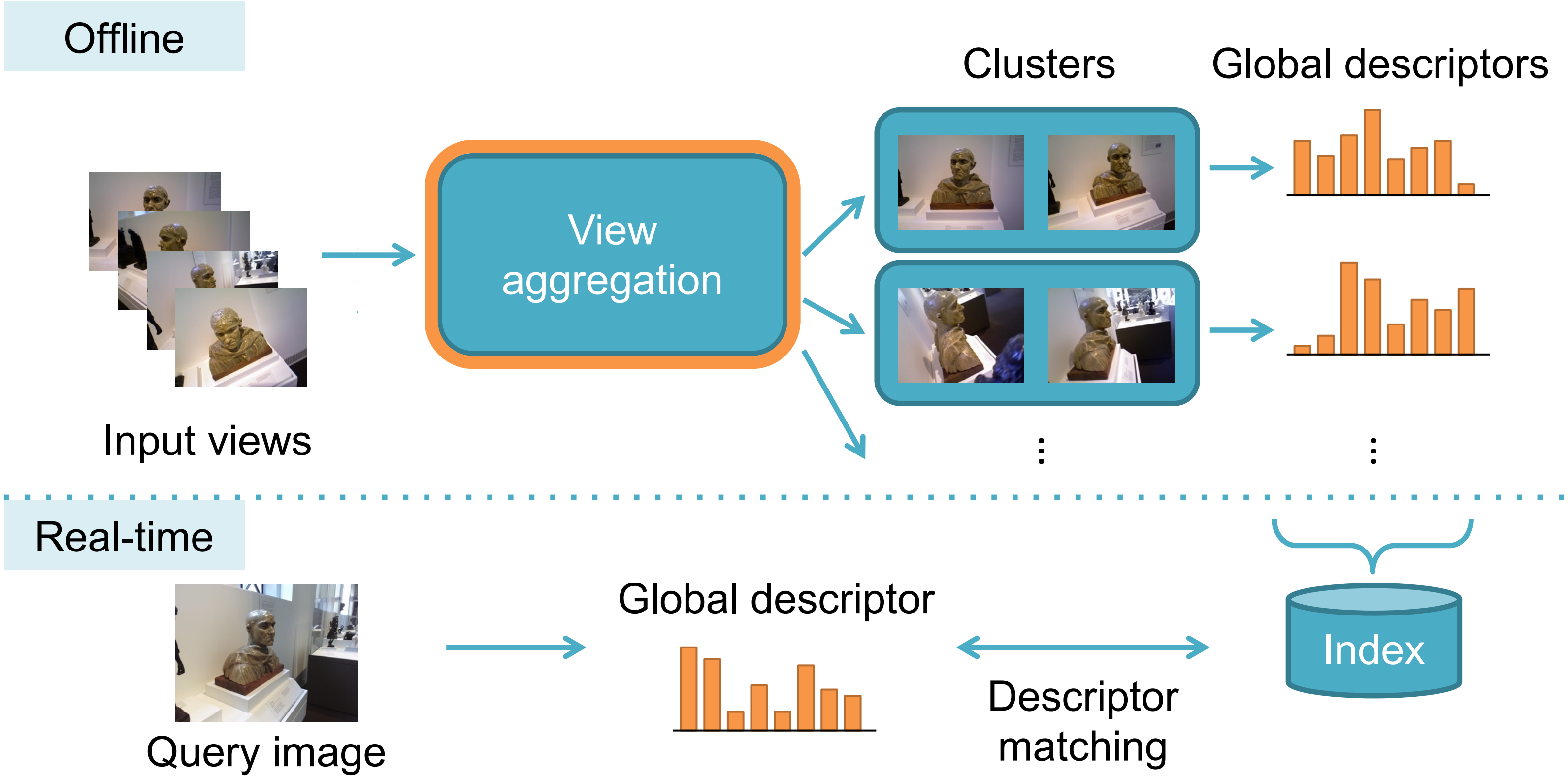
3D object recognition - Retrieval system architecture



Retrieval results

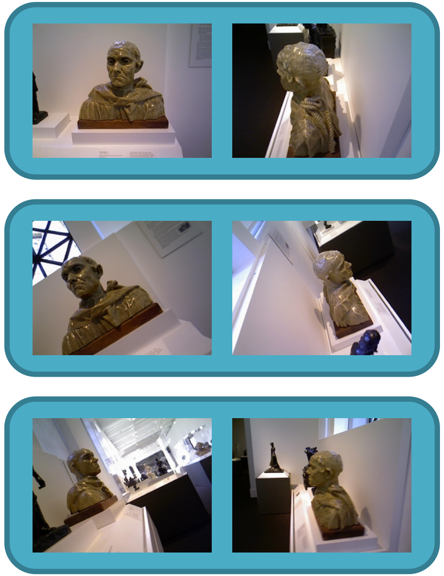


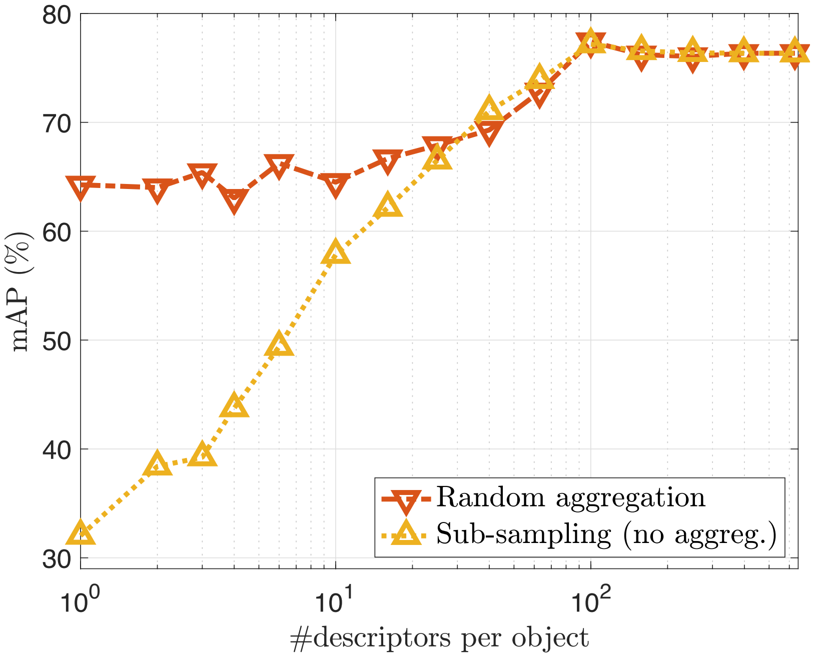
Retrieval system architecture



Retrieval results

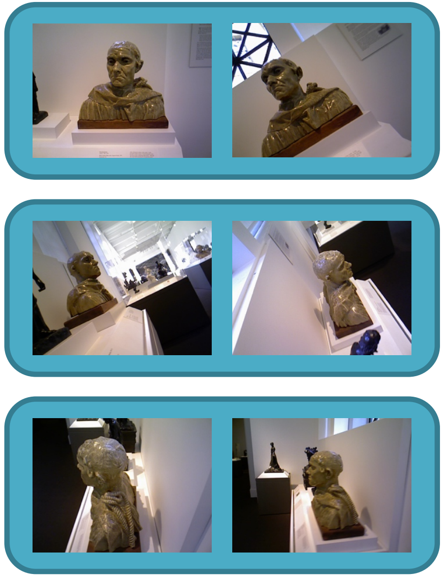
- Random aggregation

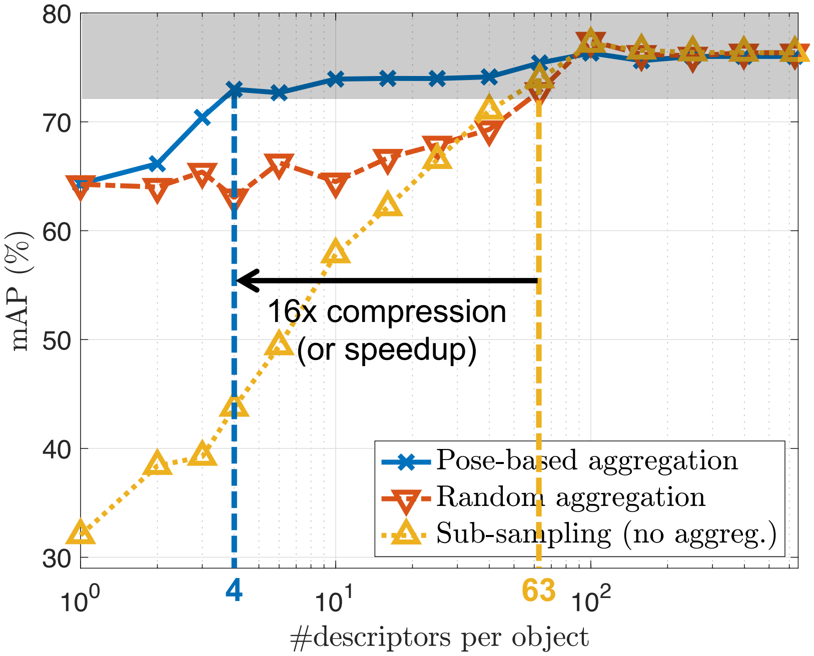




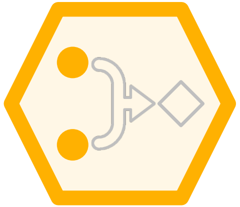
Retrieval results

- Pose-based aggregation

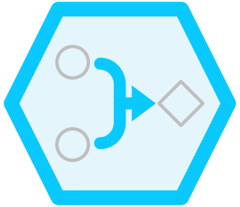




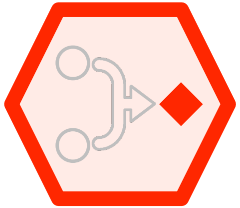
Retrieval with descriptor aggregation



WHAT DESCRIPTORS to aggregate?



WHAT AGGREGATION METHOD to use?

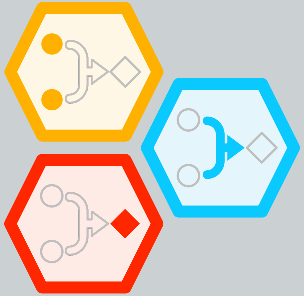


HOW TO USE aggregated descriptors?

Contributions

Theoretical framework

- Theoretical basis for aggregation
- Performance guarantees in ideal scenarios

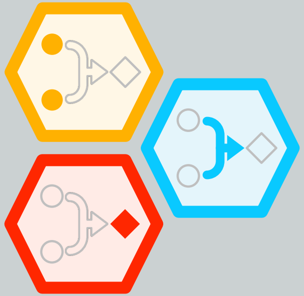


3D object retrieval

- Compact representation of 3D objects
- Comparison of aggregation schemes

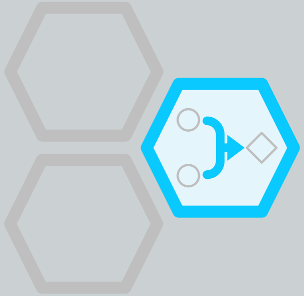
Localization

- Semantic-based descriptor indexing
- Fast accurate panorama retrieval

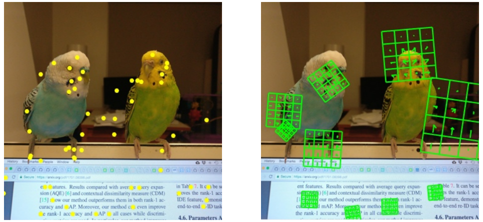


Person re-identification

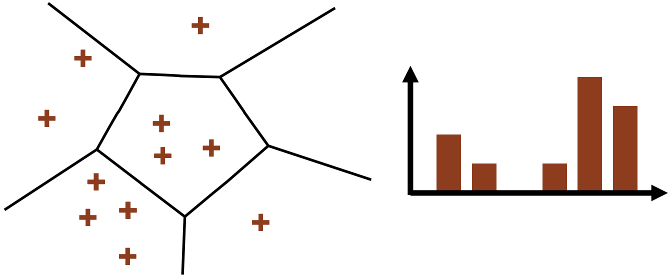
- Query-side aggregation
- Simplification of neural network based methods



Background – Image representation



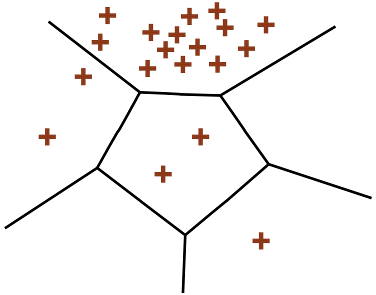
- Traditional pipeline: hand-crafted features
  - › Local patch representation: SIFT *[Lowe, ’04]*
  - › Global descriptor:
    - Bag of Words (BoW) *[Sivic et al., ’03]*
    - Fisher Vectors *[Perronnin et al., ’07]*

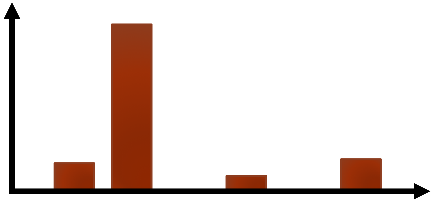


- CNN-based features
  - › Representations extracted from networks trained on other tasks
  - › Can be fine-tuned for improved results

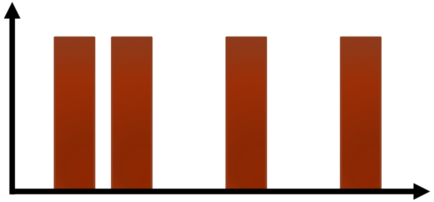
Background – Descriptor aggregation

- Mean/sum pooling (used in Fisher Vectors)
  - "Burstiness" problem *[Jégou et al., '09]*
- Max pooling: preferred for BoW *[Boureau et al., '10]*





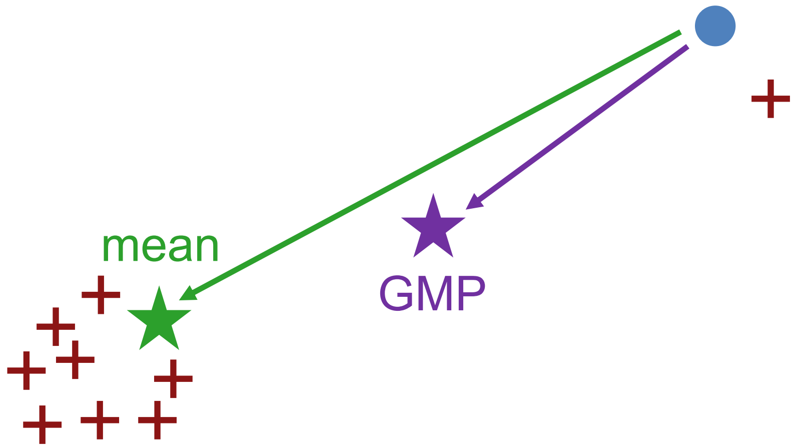
Sum pooling



Max pooling

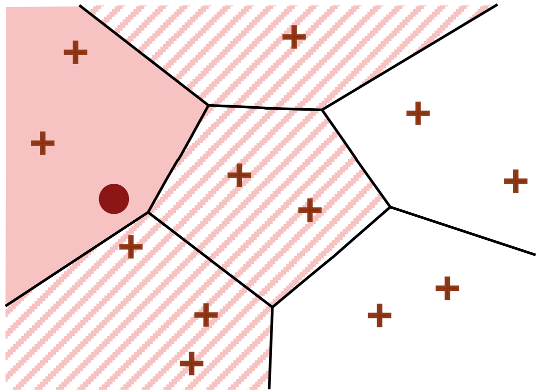
Background – Descriptor aggregation

- Generalized max-pooling (GMP) *[Murray et al., ’14]*
  - Increased similarity to ALL descriptors



Background – Indexing and search

- Task: Nearest Neighbor (NN) search
  - Database: $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$, with $\boldsymbol{x}_i \in \mathbb{R}^d$, $\|\boldsymbol{x}_i\|^2 = 1$
  - Query: $\boldsymbol{q} \in \mathbb{R}^d$, $\|\boldsymbol{q}\|^2 = 1$
  - Find i that maximizes $\boldsymbol{q}^{\mathrm{T}}\boldsymbol{x}_i$
  - Exhaustive search: O(Nd)
- High dimensional exact NN search is hard
  - Space-partitioning indexing: e.g., k-d tree *[Friedman et al., '77]*
  - When $d \geq 10$, no gains compared to exhaustive search *[Weber et al., '98]*



Background – Indexing and search

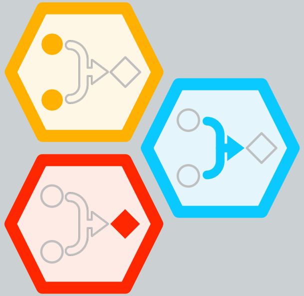
- Approximate Nearest Neighbor (ANN) techniques:
  - Space-partitioning techniques
    - FLANN *[Muja et al., ’14]*
  - Distance approximation: search in a lower-dimensional space
    - $O(Nd) \rightarrow O(Nd')$, $d' \ll d$
    - Locality-Sensitive Hashing (LSH) *[Charikar, ’02]*
      Product Quantization (PQ) *[Jégou et al., ’11]*
  - Aggregate descriptors into groups represented by a single vector
    - $O(Nd) \rightarrow O(N'd)$, $N' \ll N$
    - Group testing *[Shi et al., ’14]*

Contribution 1

Theoretical framework

Theoretical basis for aggregation

Performance guarantees in ideal scenarios



3D object retrieval

Compact representation of 3D objects

Comparison of aggregation schemes

Localization

Semantic-based descriptor indexing

Fast accurate panorama retrieval

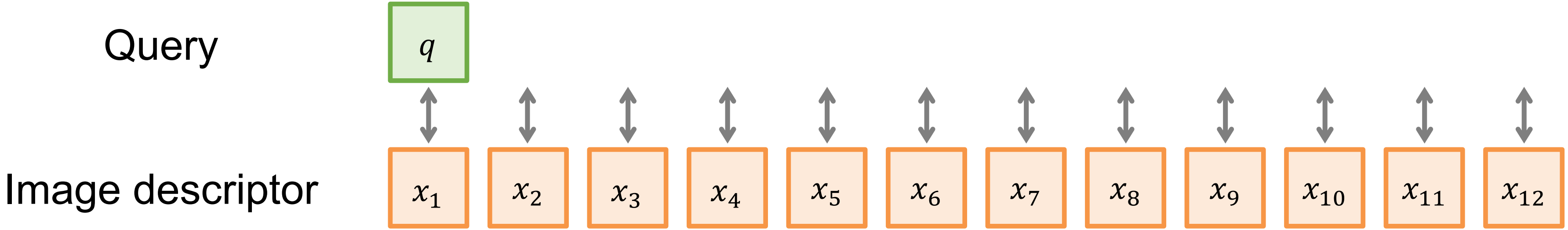
Person re-identification

Query-side aggregation

Simplification of neural network based methods

Memory vectors

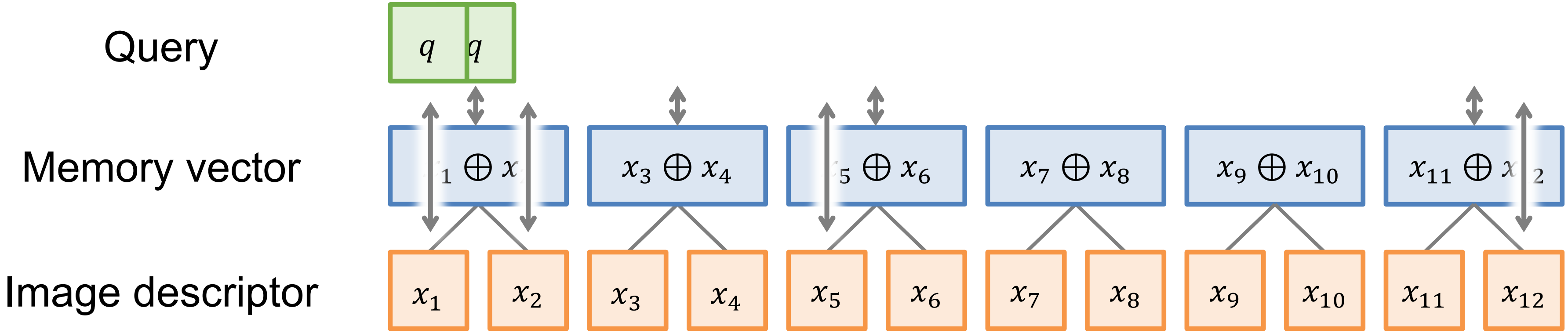
- "Memory vectors for similarity search in high-dimensional spaces" *[Iscen et al., '17]*
- Exhaustive search



- Complexity (number of similarity computations): $C = N$

Memory vectors

- Dataset partitioned into units of size n, represented by a "memory vector" $\boldsymbol{m}$
- Memory vector discarded if similarity is: $\boldsymbol{q}^{\mathrm{T}}\boldsymbol{m} < \tau$



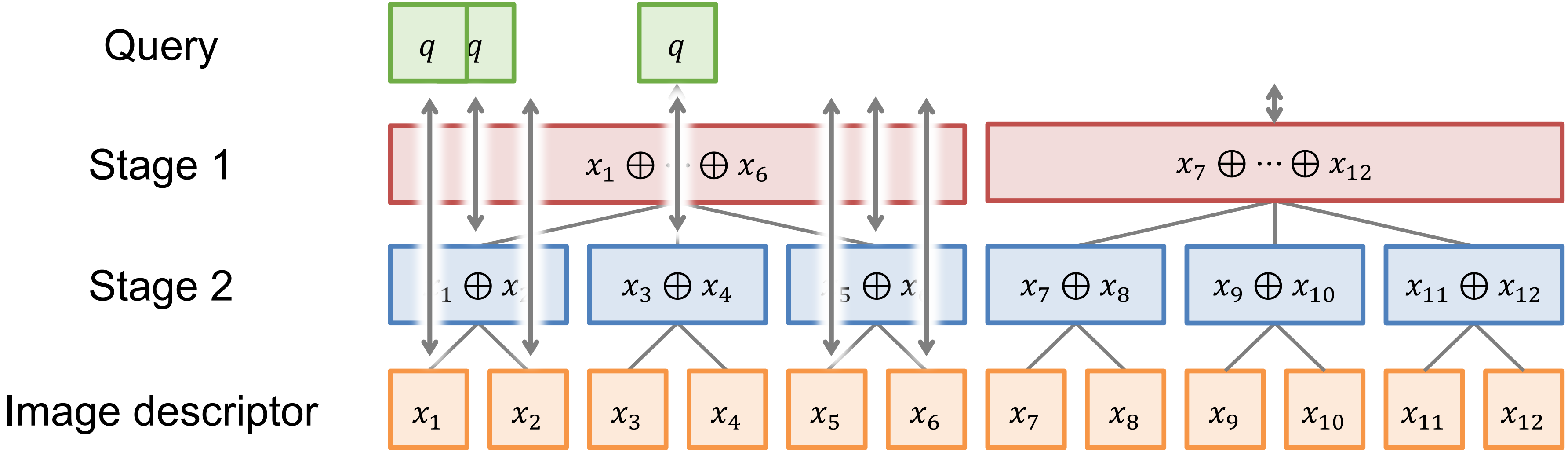
- Complexity: $C \underset{N\to+\infty}{\sim} \frac{\mathrm{N}}{n} + \mathrm{N} \cdot P_{fp}$

$\boxed{n = 2}$

Stanford University

Two-stage retrieval

- Add another stage by aggregating memory vectors themselves

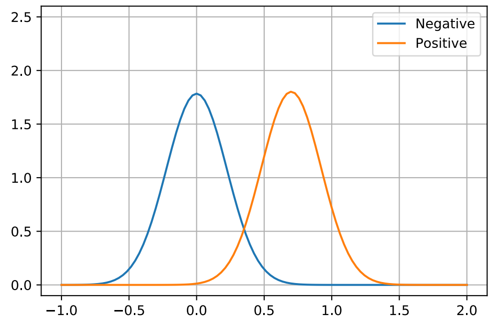


- Complexity: $C \underset{N\to+\infty}{\leq} \frac{N}{n_1} + \frac{N\cdot P_{fp^{(1)}}}{n_2} + N\cdot P_{fp^{(2)}}$

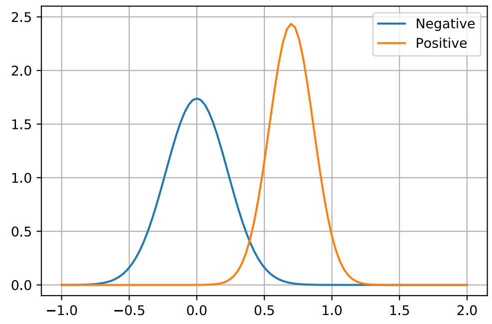
$$\boxed{\begin{array}{l} n_1 = 6 \\ n_2 = 2 \end{array}}$$

Database / query model

- Database: $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$, i.i.d., drawn uniformly over the unit sphere
- Query: random variable $\boldsymbol{Q}$ related to one vector, w.l.o.g. $\boldsymbol{x}_1$
  - $\boldsymbol{Q} = \alpha \boldsymbol{x}_1 + \beta \boldsymbol{Z}$ (such that $\|\boldsymbol{Q}\|^2 = 1$)
- We derive the distributions of scores of positive and negative memory vectors using SUM or GMP aggregation



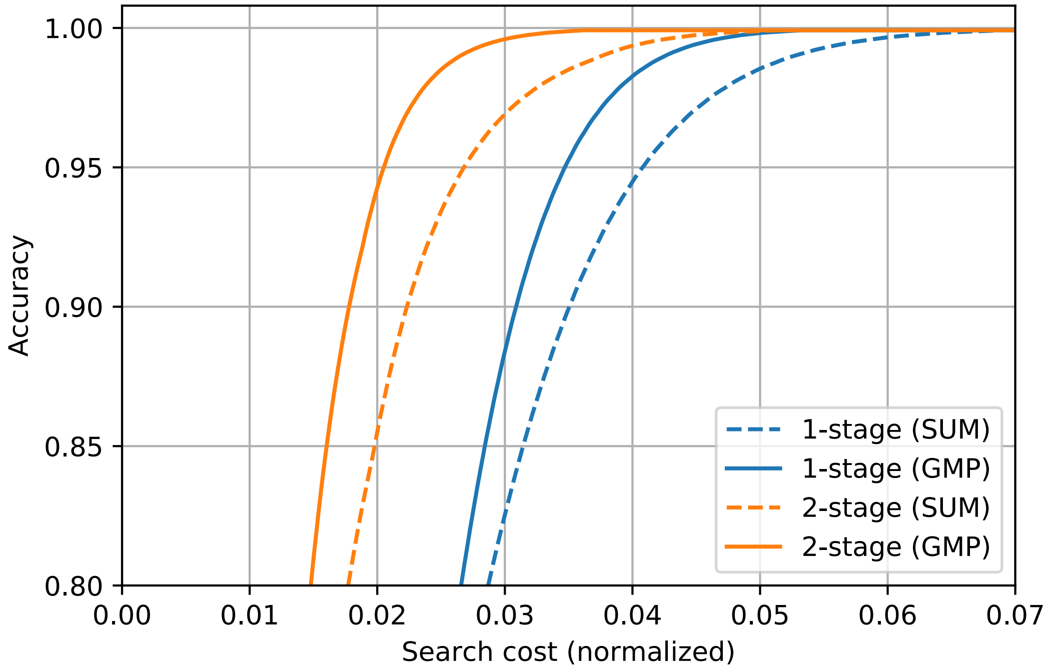
SUM



GMP

Parameters:
d = 1000
α = 0.7
n = 50

Two-stage retrieval – Results

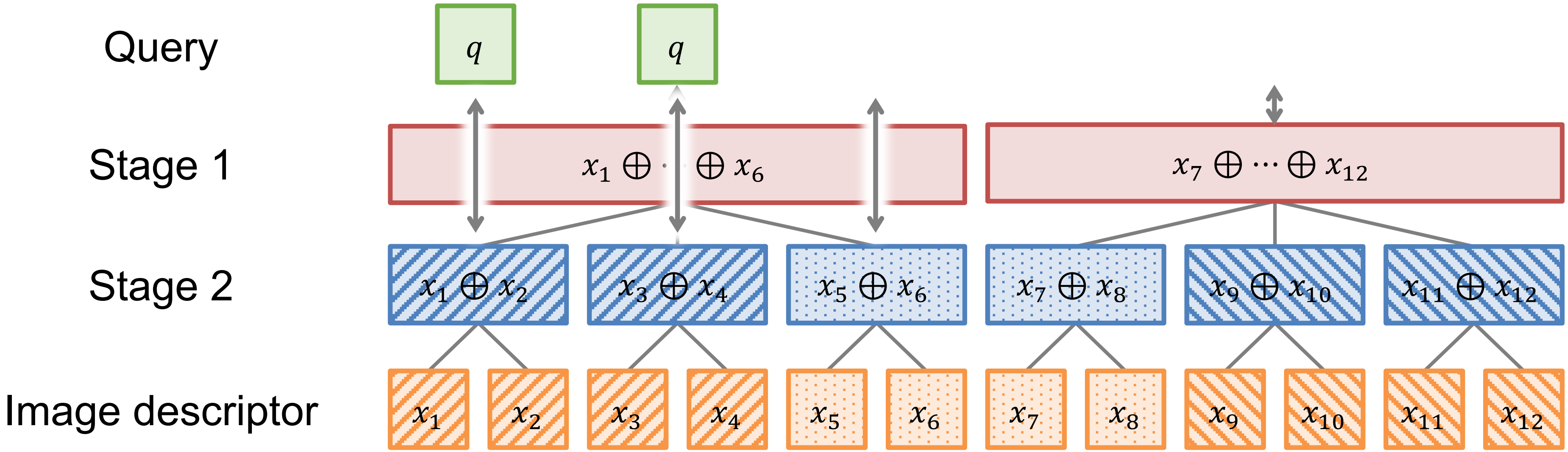


Parameters:
d = 1000
α = 0.7

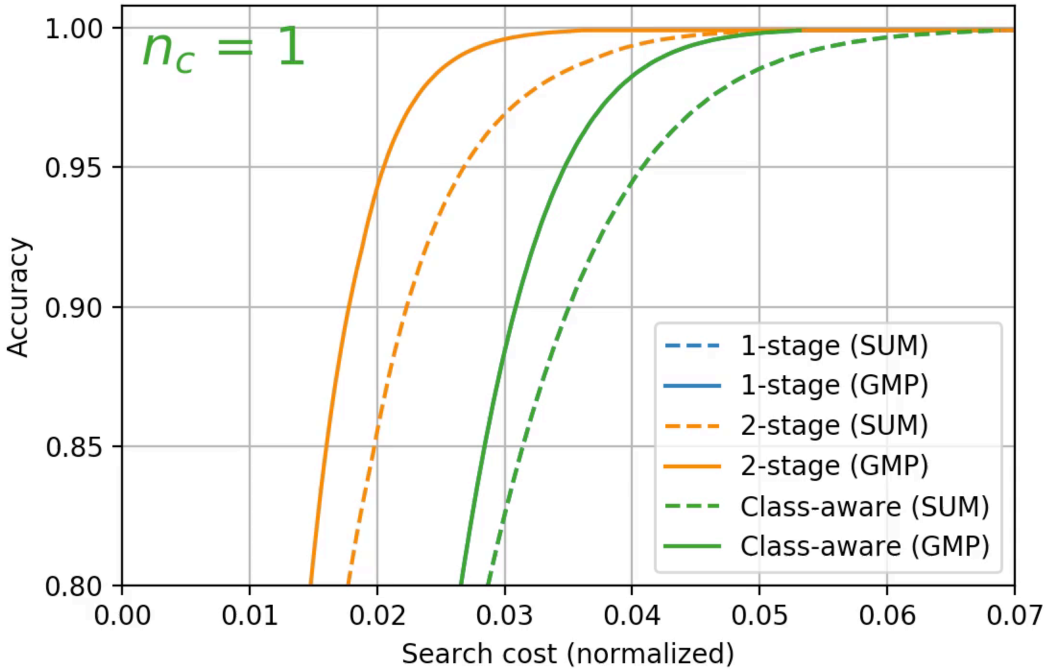
Class-aware retrieval

- Class labels: n_C vectors per class
- Task: class retrieval

- Proposed adaptations:
  - Stage 2 only aggregates **within class**
  - Stop search at stage 2

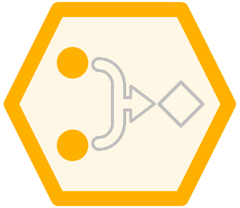


Class-aware retrieval – Results

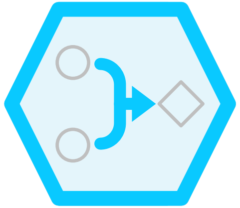


Parameters:
d = 1000
α = 0.7

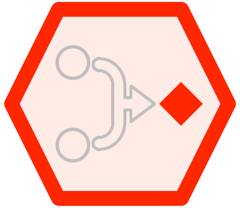
Theoretical framework – Take-away



Within-class aggregation offers speed gains by removing the need for image-level search



GMP provides a better representation for a set of descriptors



Nested aggregation levels yield a better cost/performance trade-off

Contribution 2

Theoretical framework

Theoretical basis for aggregation

Performance guarantees in ideal scenarios

3D object retrieval

Compact representation of 3D objects

Comparison of aggregation schemes

Localization

Semantic-based descriptor indexing

Fast accurate panorama retrieval

Person re-identification

Query-side aggregation

Simplification of neural network based methods

3D object retrieval – Take-away

Descriptors sharing similar characteristics (e.g. camera pose) should be aggregated

Code available on GitHub:
https://github.com/jbboin/fisher_vector_aggregation_3d

Contribution 3

Theoretical framework

Theoretical basis for aggregation

Performance guarantees in ideal scenarios

3D object retrieval

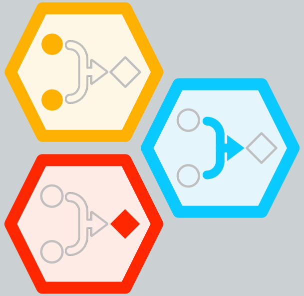
Compact representation of 3D objects

Comparison of aggregation schemes

Localization

Semantic-based descriptor indexing

Fast accurate panorama retrieval



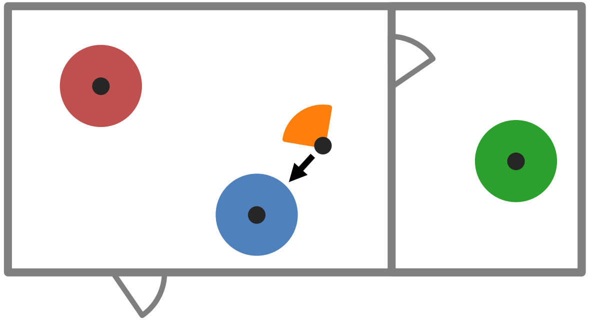
Person re-identification

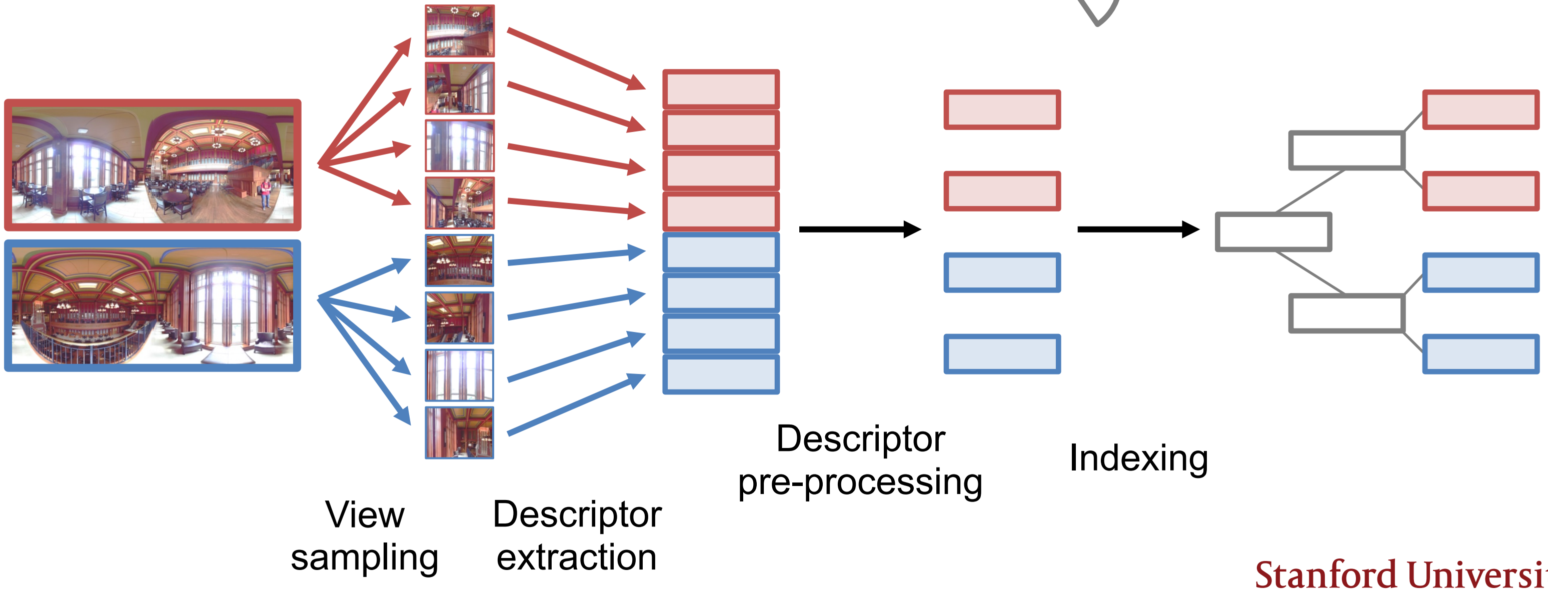
Query-side aggregation

Simplification of neural network based methods

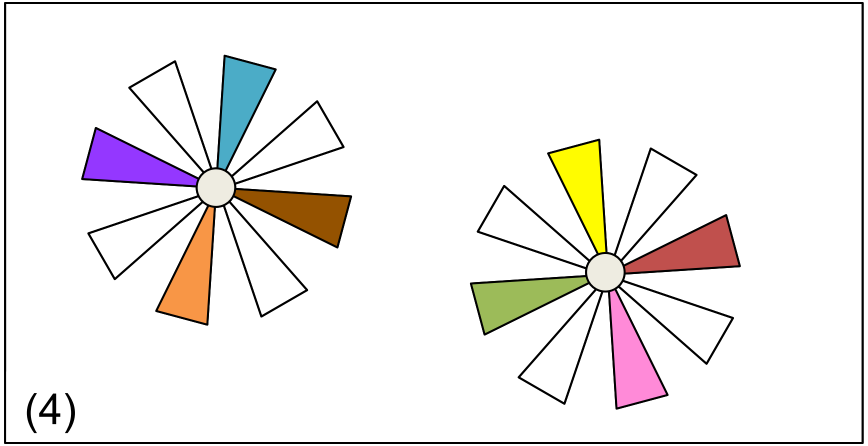
Indoor Localization

- Task: panorama retrieval using a single query image
- Need to choose how to represent a panorama as a set of views

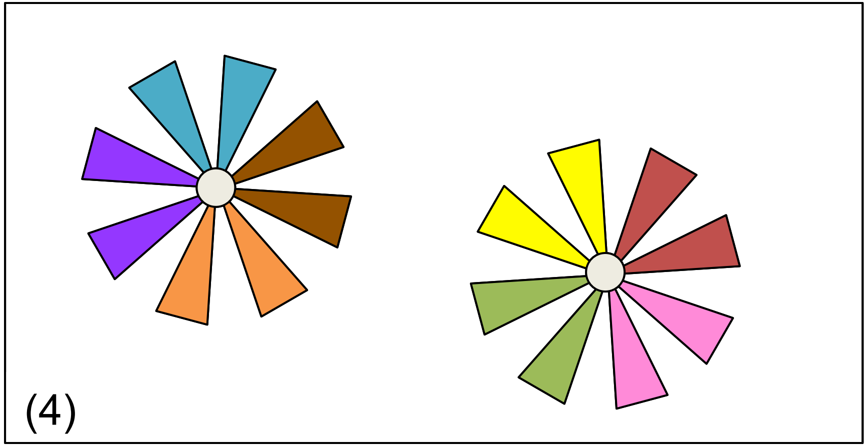




Descriptor pre-processing

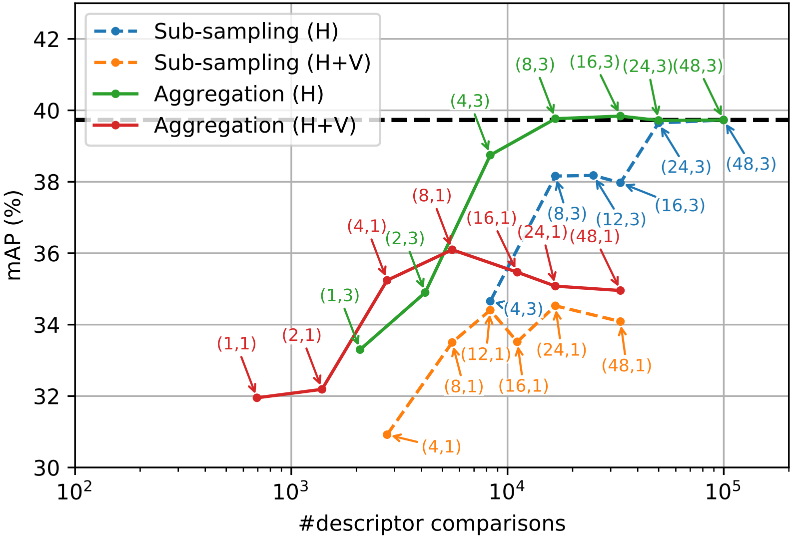


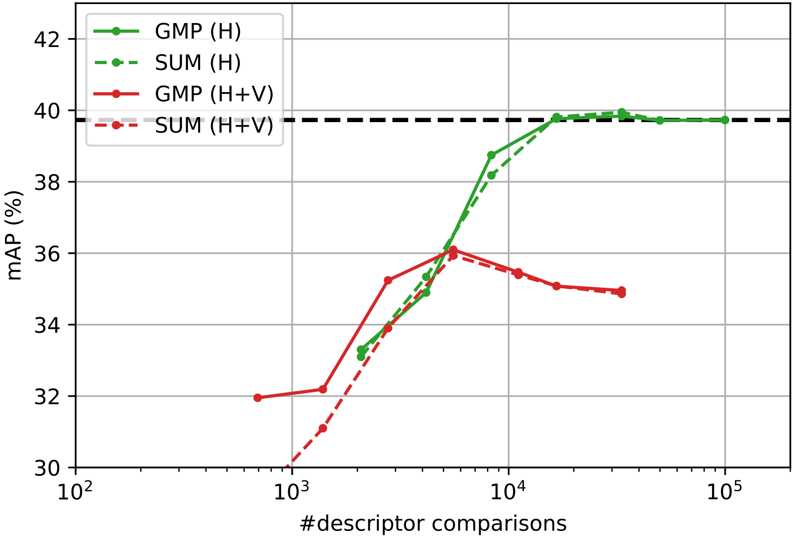
Sub-sampling



Aggregation

Results (exhaustive search)



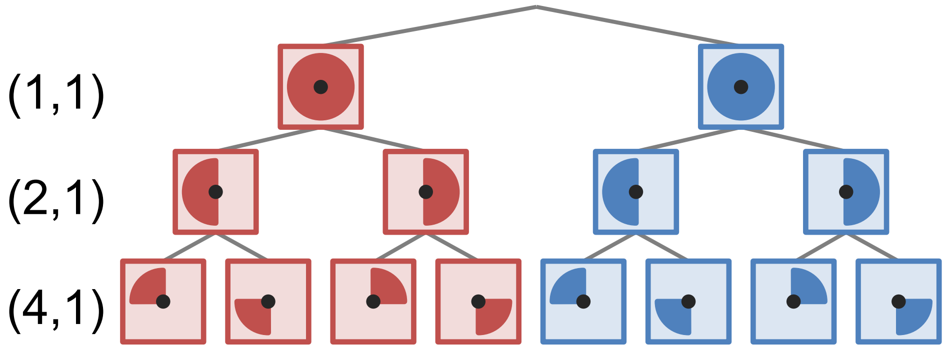


Indexing – Hierarchical aggregation

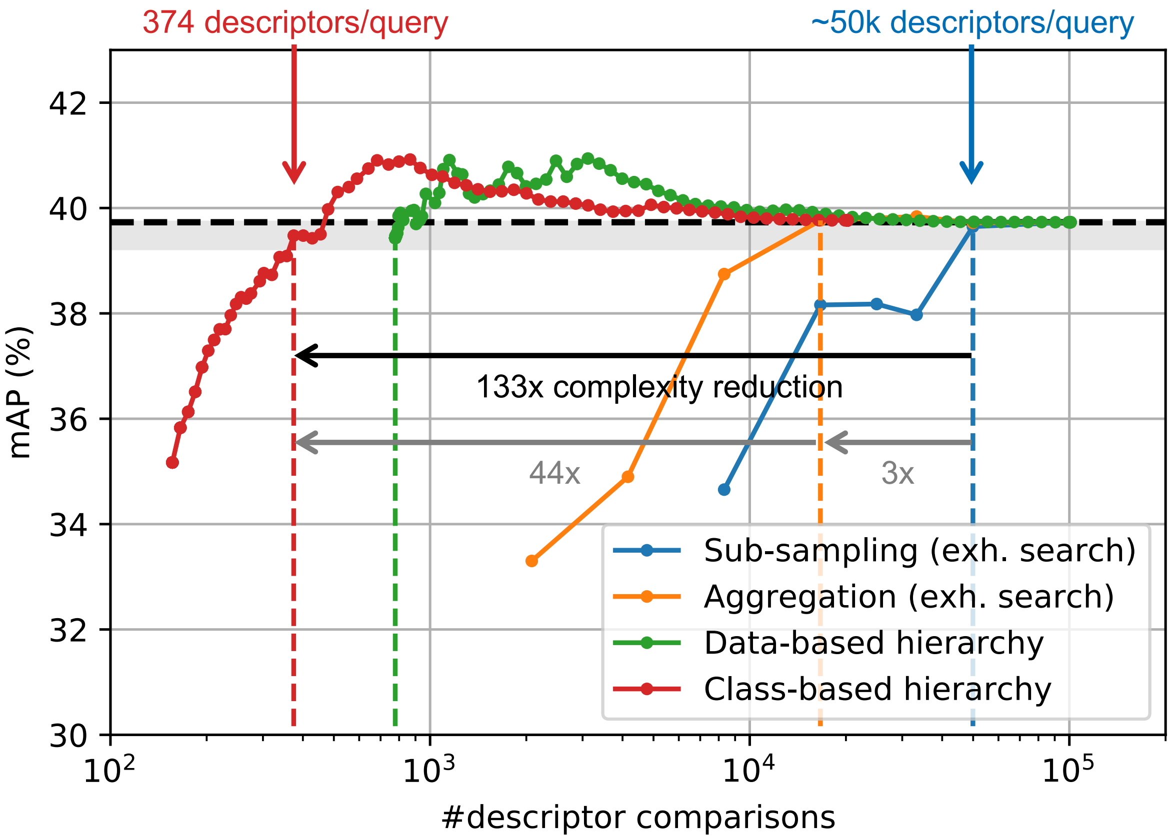
- Best of both worlds
  - Upper levels: coarse search = large complexity gains
  - Lower levels: fine search = higher retrieval performance

- Data-based hierarchy
  - Based on FLANN (k-means tree)
  - Internal node descriptors:
    - Pooled with GMP
    - Normalized

- Class-based hierarchy
  - Based on view orientation
  - + room-level aggregation



Results



374 descriptors/query
~50k descriptors/query
42
40
38
36
34
32
30
mAP (%)
133x complexity reduction
44x
3x
Sub-sampling (exh. search)
Aggregation (exh. search)
Data-based hierarchy
Class-based hierarchy
10^2
10^3
10^4
10^5
#descriptor comparisons

Localization – Take-away

Within-class aggregation offers speed gains by removing the need for image-level search

GMP provides a better representation for a set of descriptors

Nested aggregation levels yield a better cost/performance trade-off

Code available on GitHub:
https://github.com/jbboin/panorama-indexing-localization

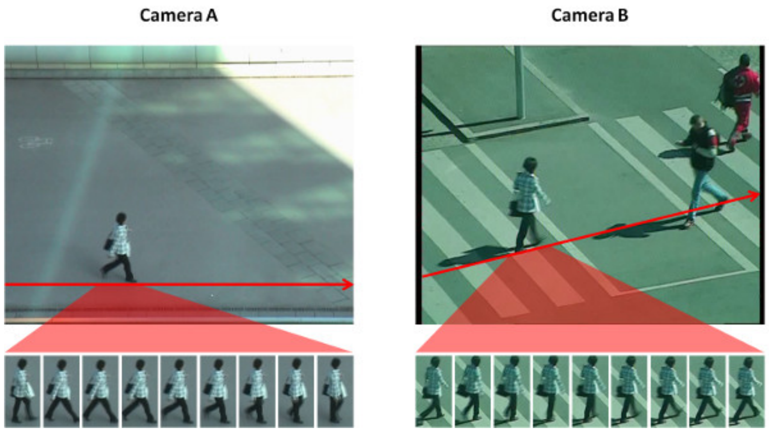
Contribution 4

Theoretical framework	3D object retrieval	Localization	Person re-identification
Theoretical basis for aggregation	Compact representation of 3D objects	Semantic-based descriptor indexing	Query-side aggregation
Performance guarantees in ideal scenarios	Comparison of aggregation schemes	Fast accurate panorama retrieval	Simplification of neural network based methods

Stanford University

Person video re-identification

- Task: associate person video tracks from different cameras



Credit: PRID2011 dataset *[Hirzer et al., '11]*
iLIDS-VID dataset *[Wang et al., '14]*

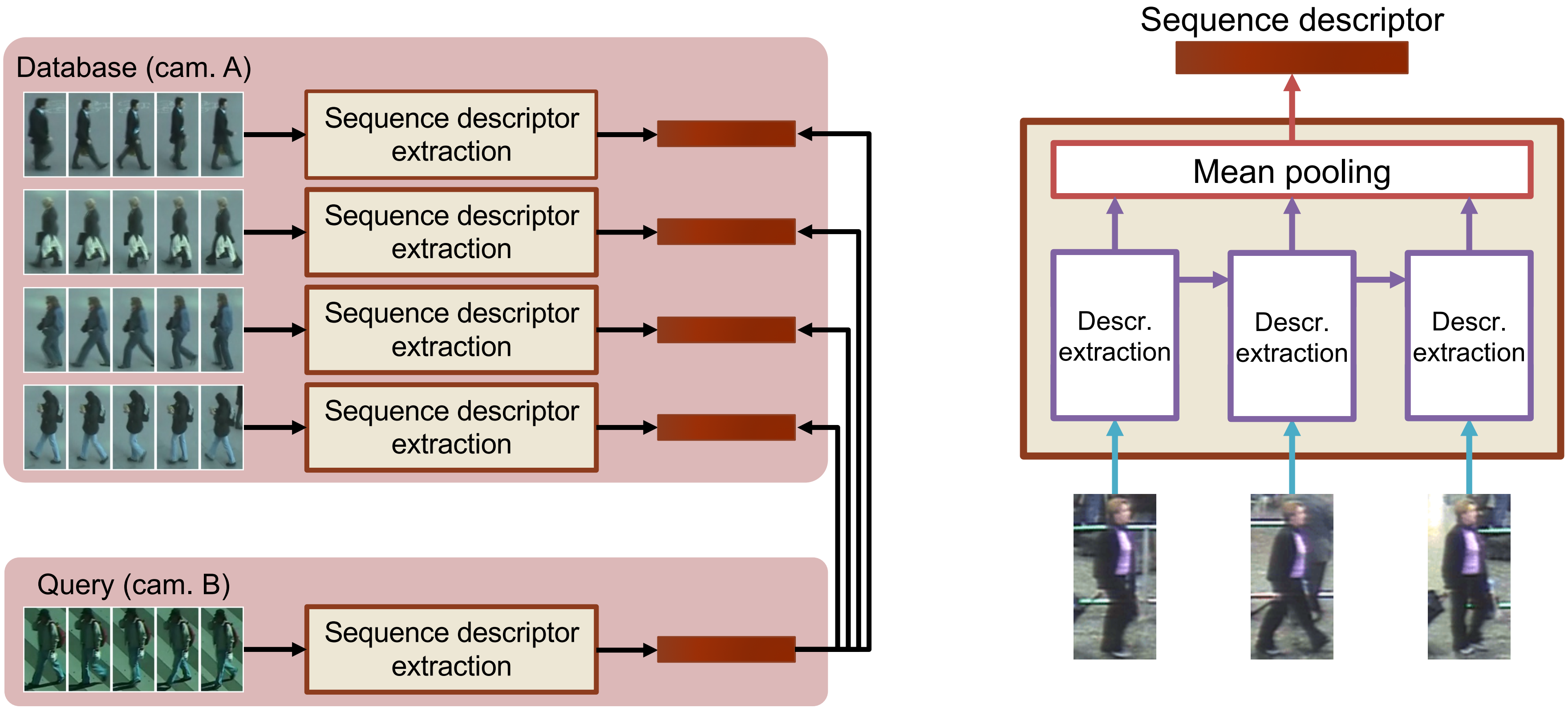
Lighting variations

Viewpoint changes

Clothing similarity

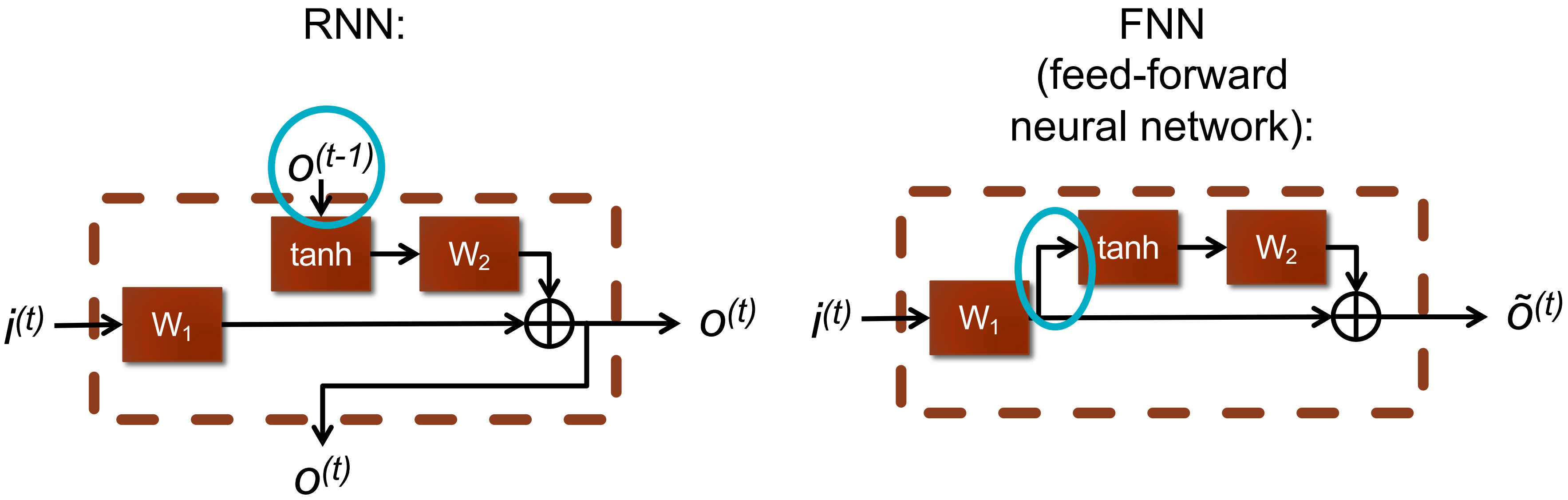
Background clutter and occlusions

Framework: re-identification by retrieval



[McLaughlin et al., '16]

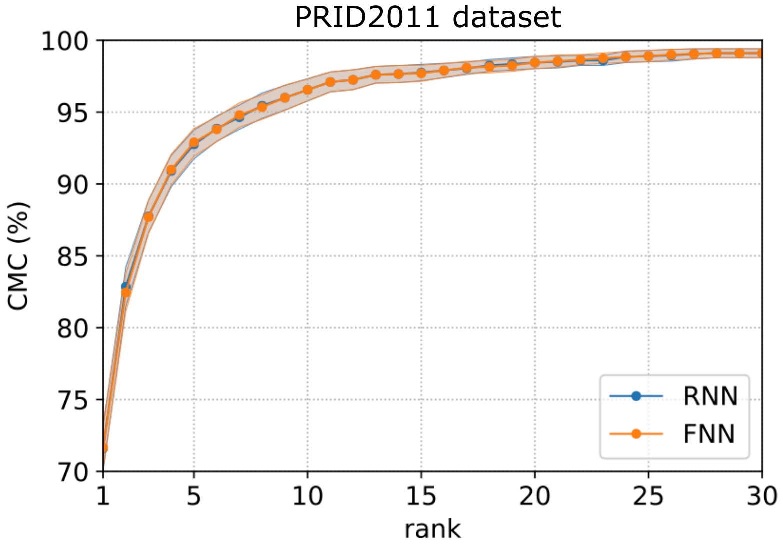
Proposed feed-forward approximation



- Same memory footprint
- Direct mapping between RNN and FNN parameters

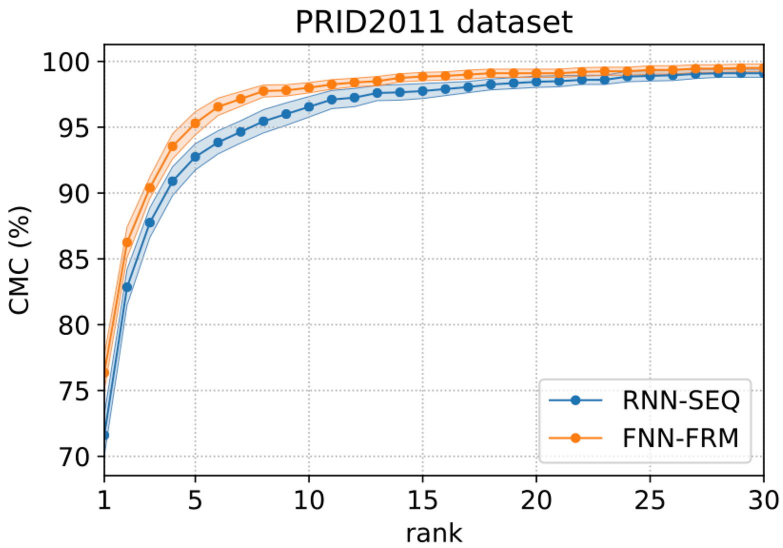
Validation of our approximation

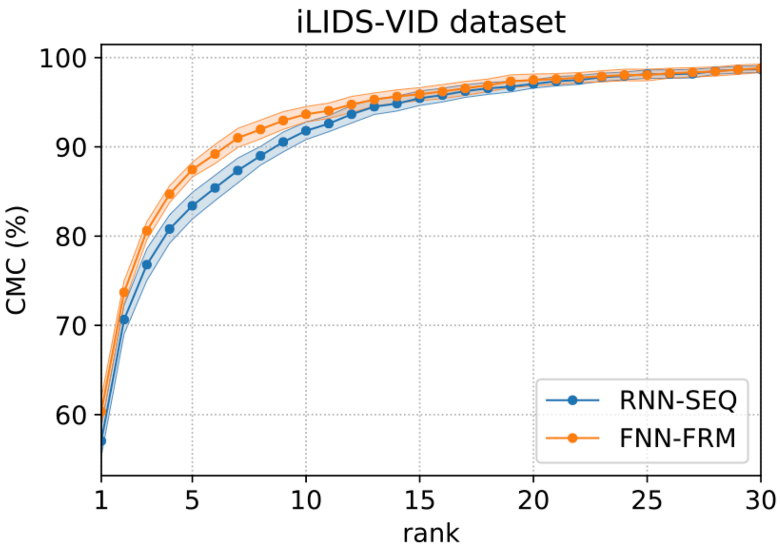
- Train weights on RNN
- Evaluate on RNN and FNN using the weights directly (**no re-training**)
- Same performance is observed



Improved training process

- More flexibility in training
  - SEQ: sequences of consecutive frames
  - FRM: independent frames



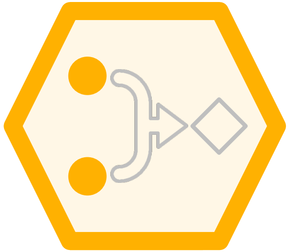


Person re-identification – Take-away

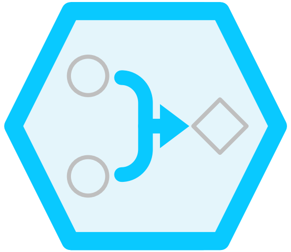
Complex techniques outperformed by simpler and more flexible temporal pooling methods

Code (partially) available on GitHub:
https://github.com/jbboin/action-recognition-revisited

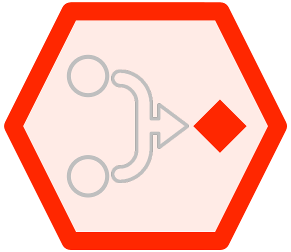
Conclusions



- Within-class aggregation keeps search to higher levels of abstraction
- In a class, aggregating based on similar characteristics is beneficial
- 16x speed increase for 3D object retrieval; 3x for localization



- GMP provides a better representation for a set of descriptors
- Higher performance when aggregating many dissimilar descriptors
- Simple pooling techniques outperform more complex ones



- Theoretical cumulative gains when nesting aggregation levels
- Hierarchical indexing makes coarse-to-fine search possible
- 44x speed increase for localization compared to exhaustive search