Large-Scale Video Retrieval Using Image Queries

André Filgueiras de Araujo

Department of Electrical Engineering

Stanford University

The “Dark Matter” of the Digital Age

85% of data in the form of multimedia

YouTube: 400+ hours of video uploaded per minute

Facebook: 8+ billion video views per day

nest: 100+ hours of video uploaded per minute

<u>Key problem:</u> How can we make sense of these data?

Automatic Visual Recognition

Image classification

- *Is this an urban landscape?*

Object detection

- *Does this image contain a bus? Where?*

Instance recognition (a.k.a. "visual search")

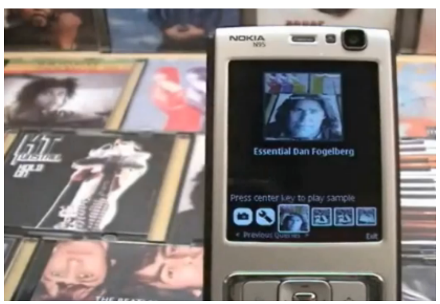
- *Does this image contain the "Wicked" billboard?*

Visual Search

Image query

Retrieval System

Database of images



Product recognition
[Tsai et al., MM'08, MM'10]

Location recognition
[Chen et al., CVPR'11]

Google goggles labs

FIREFLY Amazon

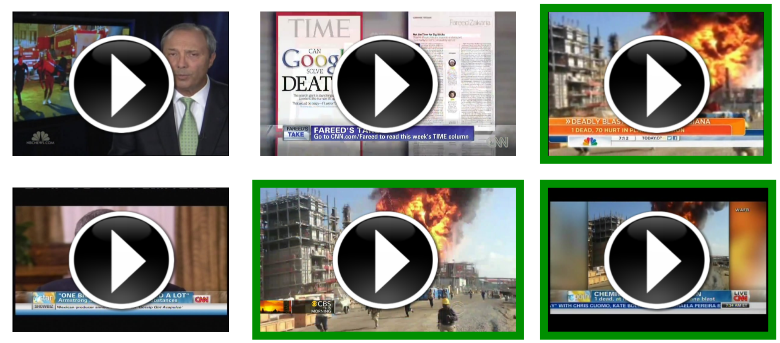
Commercial applications

Video Retrieval Using Image Queries

Image query

Retrieval System

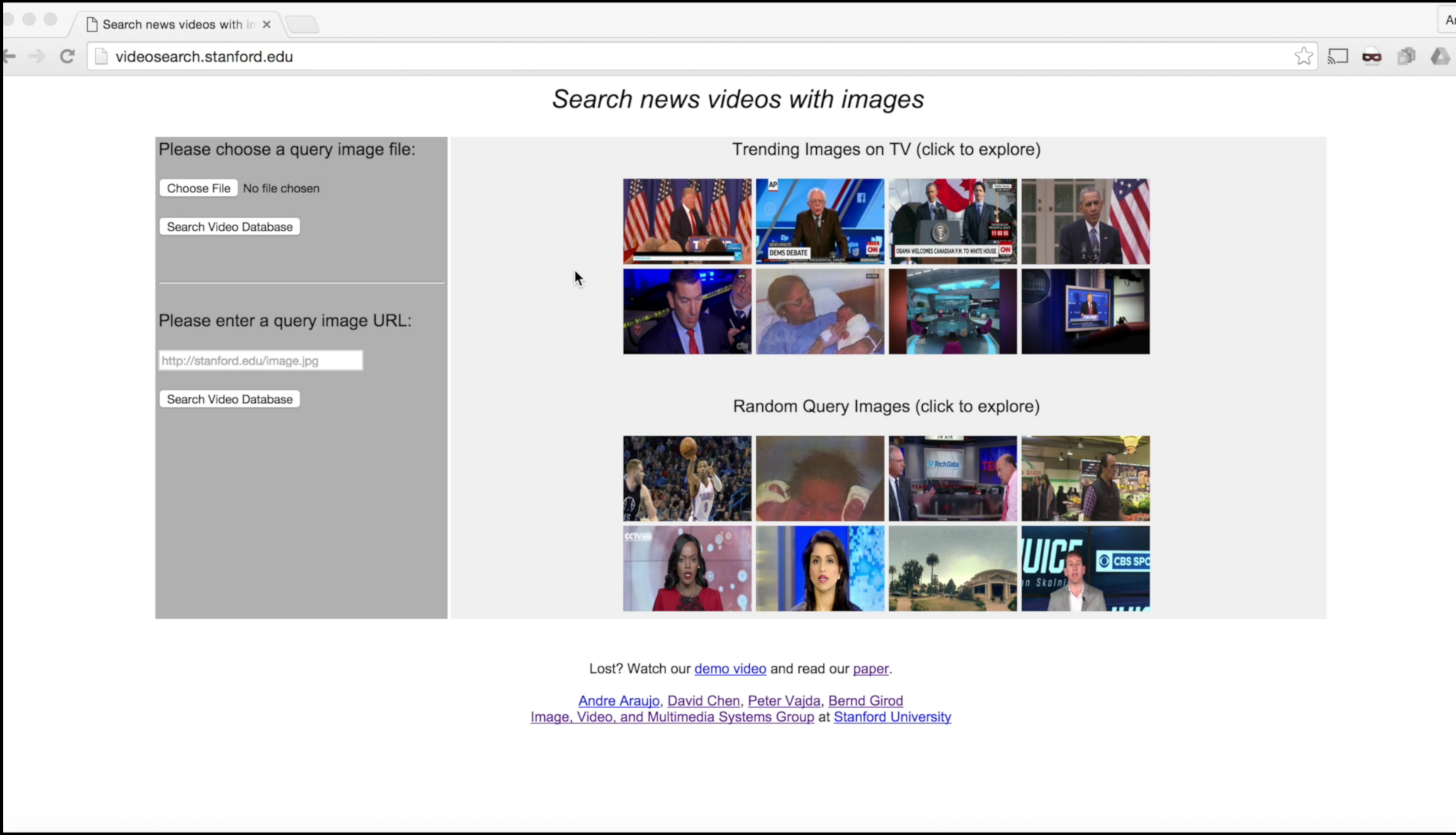
Database of video clips



Applications:

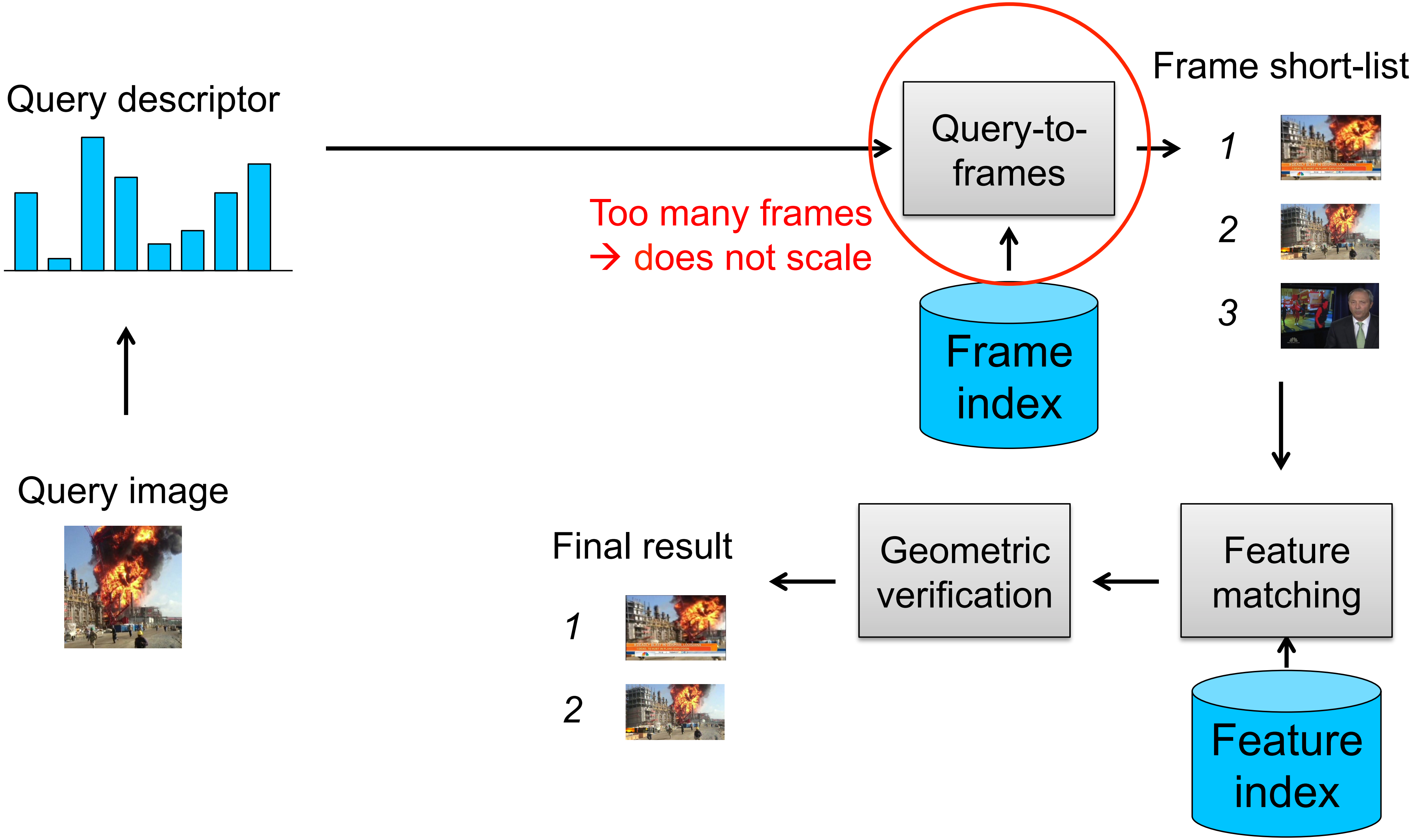
- Brand monitoring: search YouTube using product images
- News videos: search event footage using photos
- Online education: search lectures using slides

Online Prototype



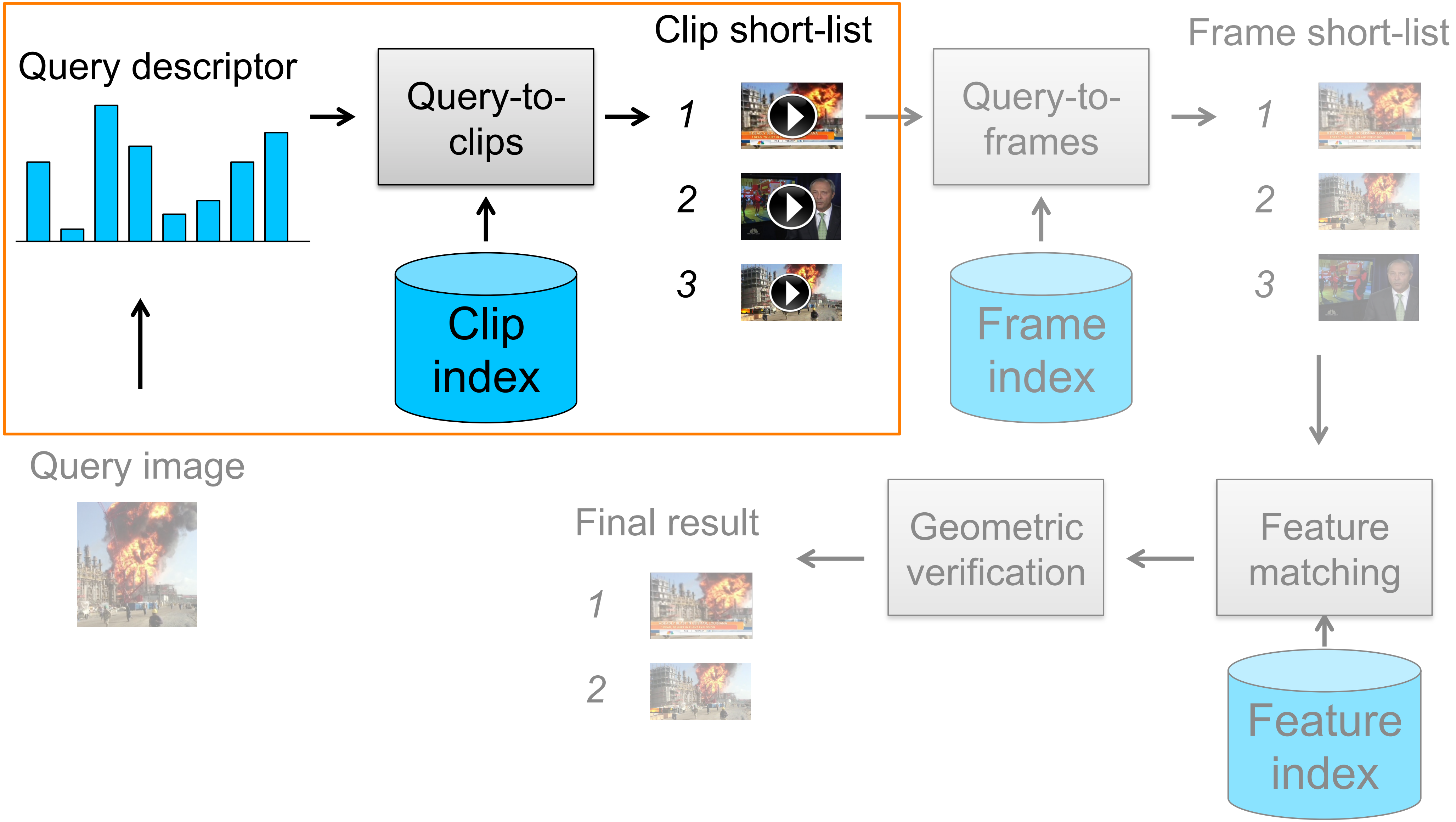
http://videosearch.stanford.edu

Simple Architecture

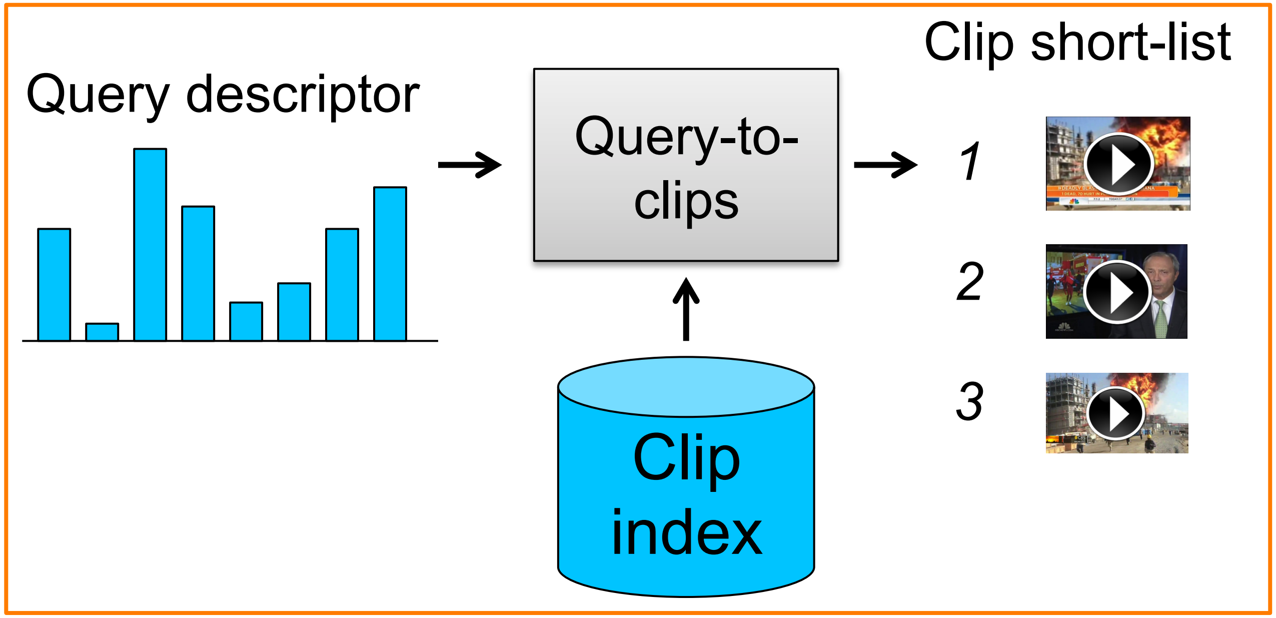


Large-Scale Architecture

Focus of this work



Video Retrieval Using Image Queries



Main challenges:

- Asymmetry: how can we compare images to videos?
- Temporal aggregation: how can we describe a video clip for query-by-image retrieval?

Contributions

Fisher Vector Comparisons

- Asymmetric comparisons for Fisher vectors
- Cluttered query or database images

Fisher Vector Aggregation

- Fisher vector descriptors for video segments
- Compact database for large-scale retrieval

Bloom Filter Aggregation

- Bloom filter descriptors for video segments
- Fast and accurate large-scale retrieval

Related Work: Visual Search

Query

	Images	Videos
Video	***Augmented Reality*** **TCD** *[Makar et al., '12]* **Hybrid Vis. Search** *[Chen et al., '14]*	***Content Tracking*** **Frame Mat. + ST** *[Douze et al., '10]* **TRECVID-CCD** *[Over et al., '12]*
Image	***Traditional Visual Search*** **FV** *[Perronnin et al., '07]* **BoW** *[Sivic et al., '03]* **SIFT** *[Lowe, '04]*	***Video Retrieval by Image*** *Discussed on next slide*

Database

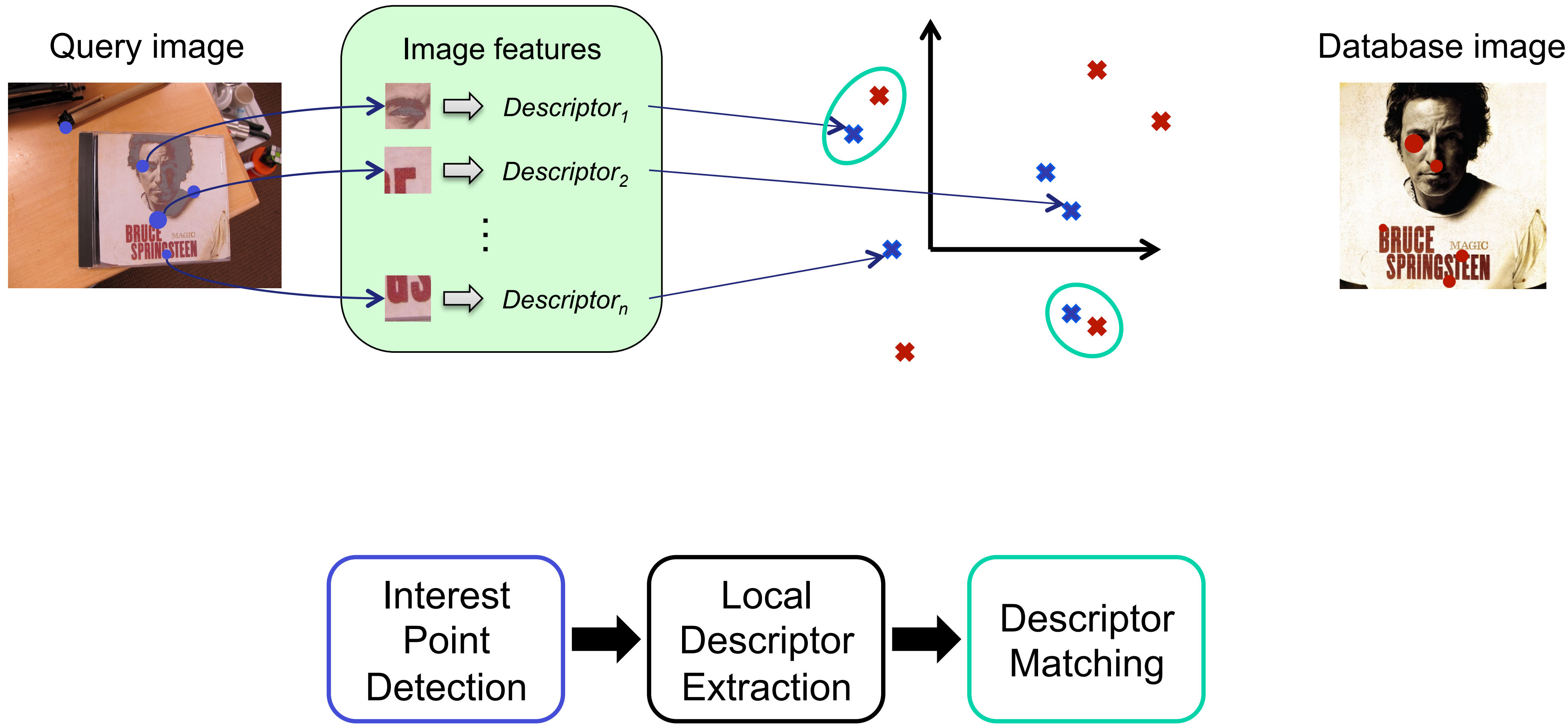
Related Work: Video Retrieval Using Images

- Early work
  - BoW retrieval of movie frames *[Sivic and Zisserman, ICCV'03]*
  - Object-level retrieval of movie shots *[Sivic et al., ECCV'04]*

- TRECVID Instance Search Challenge *[Over et al., TRECVID'10-15]*
  - Frame-based BoW with Color SIFT *[Le et al., '10-11]*
  - Shot-based aggregation using BoW *[Zhu et al., '13] [Ballas et al., '14]*
  - BoW query-adaptive asymmetrical dissimilarities *[Zhu et al., '13]*

- Object localization in videos
  - SURF-based matching per shot *[Apostolidis et al., ICME'13]*
  - Optimal path using dynamic programming *[Meng et al. ICIP'15]*

Background: Pairwise Image Matching



Background: Fisher Vector (FV)

[Perronnin and Dance, CVPR'07]

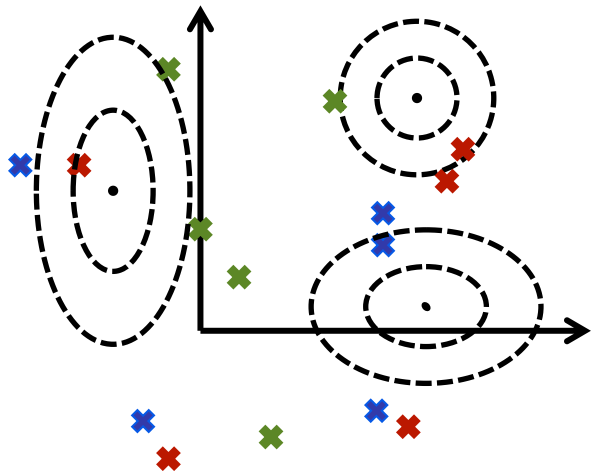
- State-of-the-art technique for large-scale retrieval
- Key property: represent a set of local descriptors by a compact fixed-length vector
  → Two images can be compared by comparing their Fisher vectors
- Construction: describe an image with aggregated Fisher scores of its local descriptors
  - Local descriptor distribution: Gaussian Mixture Model (GMM)
  - Usually only Gaussian means are taken into account
- Extension of Bag-of-Words technique *[Sivic and Zisserman, ICCV'03]*

Background: Fisher Vector (FV)

[Perronnin and Dance, CVPR'07]

Query image

Descriptor space



Database image 1

Database image 2

⋮

Query FV	-0.2	0.2	-0.3	-0.3	-0.3	0.8
DB Im. 1 FV	-0.3	0.3	0.3	-0.6	-0.3	0.3
DB Im. 2 FV	0.5	-0.2	-0.7	0.1	-0.6	0

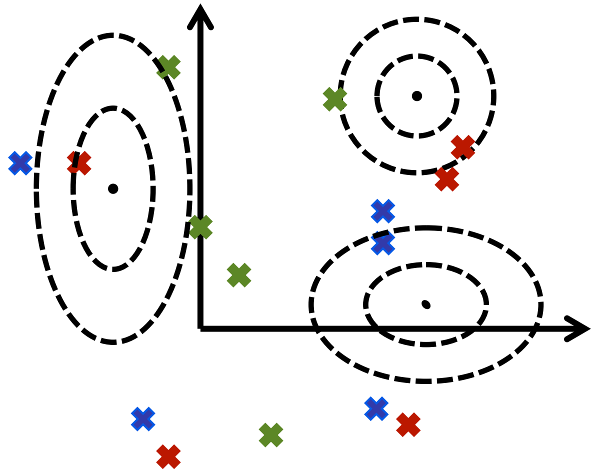
⋮

Background: Binarized Fisher Vector (FV*)

[Perronnin et al., CVPR’10]

Query image

Descriptor space



Database image 1

Database image 2

⋮

*Query FV**	0	0	0	0	0	1
*DB Im. 1 FV**	0	1	1	0	0	1
*DB Im. 2 FV**	1	0	0	1	0	0

⋮

Contribution 1

Fisher Vector Comparisons

- Asymmetric comparisons for Fisher vectors
- Cluttered query or database images

Fisher Vector Aggregation

- Fisher vector descriptors for video segments
- Compact database for large-scale retrieval

Bloom Filter Aggregation

- Bloom filter descriptors for video segments
- Fast and accurate large-scale retrieval

Asymmetric Image Comparison

Query image

Database image

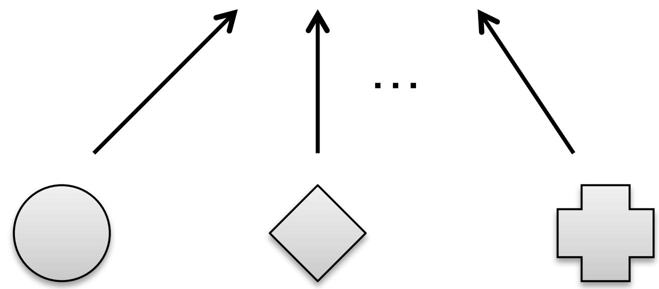
Object retrieval application

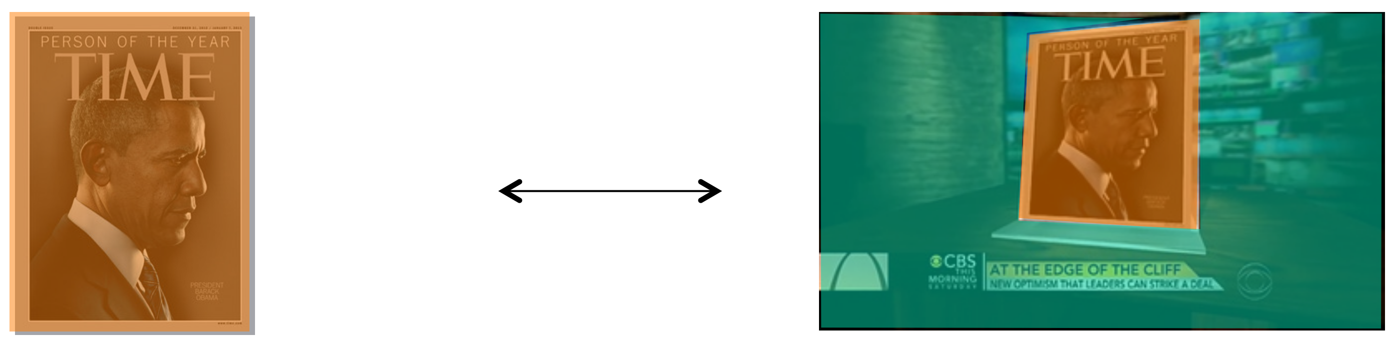
Video bookmarking application

How can we incorporate asymmetry in FV comparisons?

Asymmetric Comparison for FV

Fisher vector = $[v_1, v_2, \ldots, v_K]$

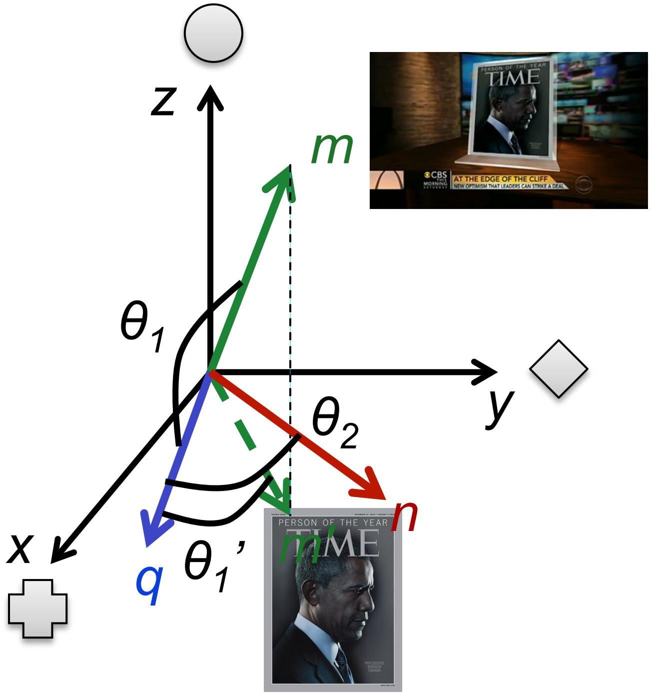




Regions and have different statistics

→ features from are usually not present in

Asymmetric Comparison for FV



q	query
m	correct match in database
n	incorrect match in database
θ_1	= angle(*q*, *m*)
θ_2	= angle(*q*, *n*)
θ_1'	= angle(*q*, *m'*)

- FV comparison metric: cosine similarity
- We want: $\theta_1 < \theta_2$
- Common failure case:

  $\theta_1 > \theta_2$ but $\theta_1' < \theta_2$

- Insight:

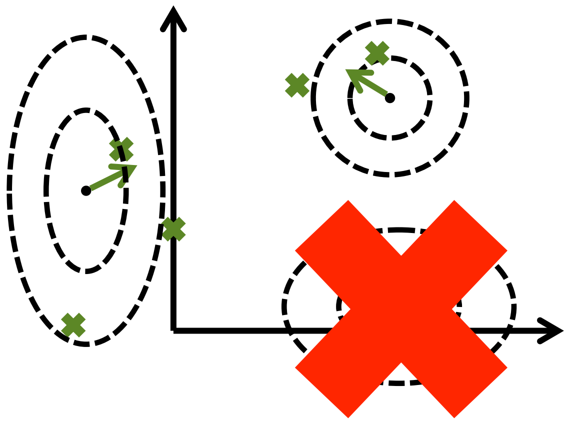
Compare query and database based on their projections to the x-y plane

(i.e., using only Gaussians visited by query)

Asymmetric Comparison for FV

Image

Descriptor space



Gaussian not visited by this image

Original FV

0.7	0.2	-0.5	0.2	-0.2	0.2

Re-norm. Zero

Modified FV

0.8	0.3	-0.5	0.3	0	0

Asymmetric Comparison for FV

- Two retrieval problems
  - Query contained in database

    All database images compared to query based on the same subspace

Query

Database

Query image defines projection

  - Database contained in query

    Problem: each database image is compared to the query based on different subspaces

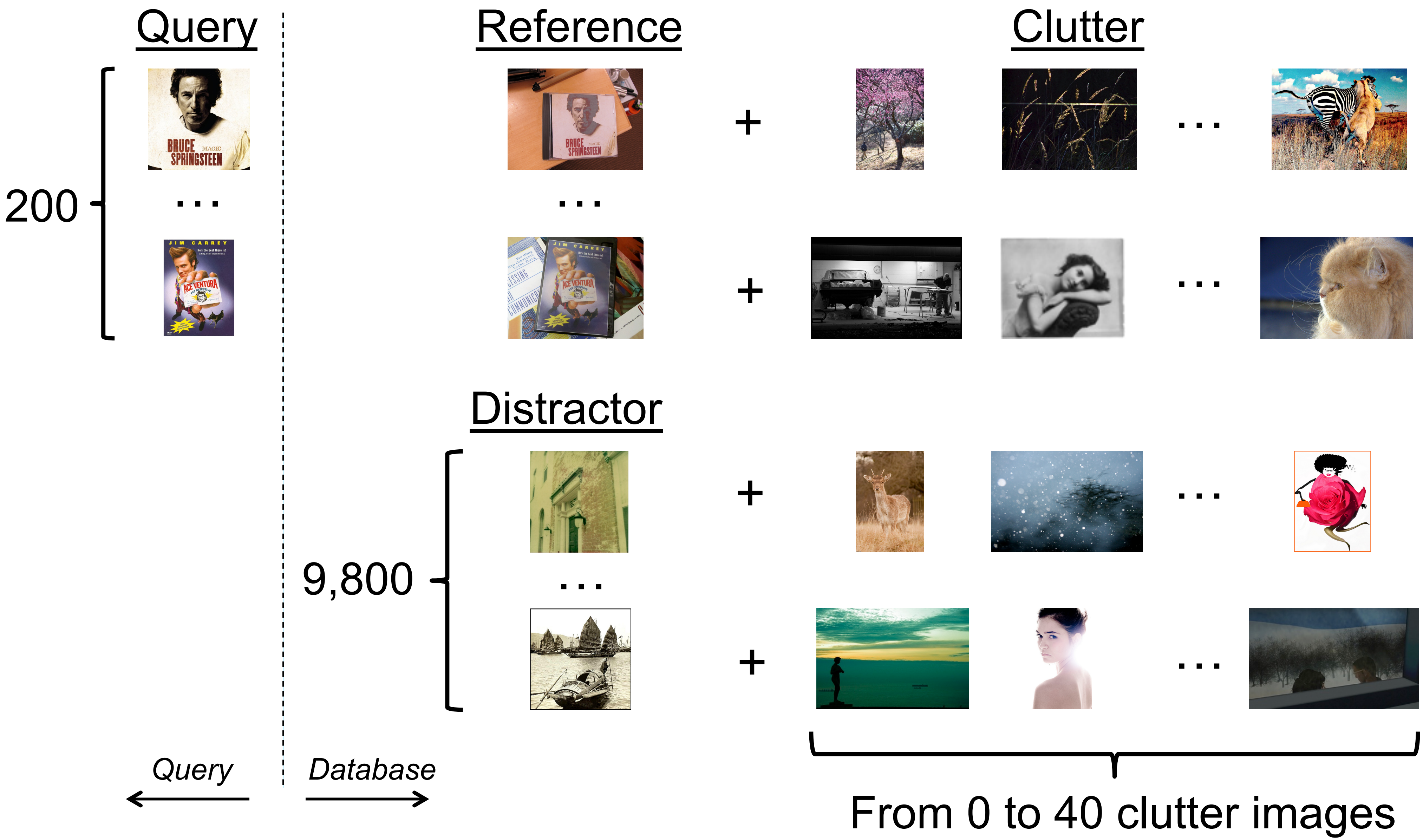
    Solution: introduce weight to favor database images with more visited Gaussians

Query

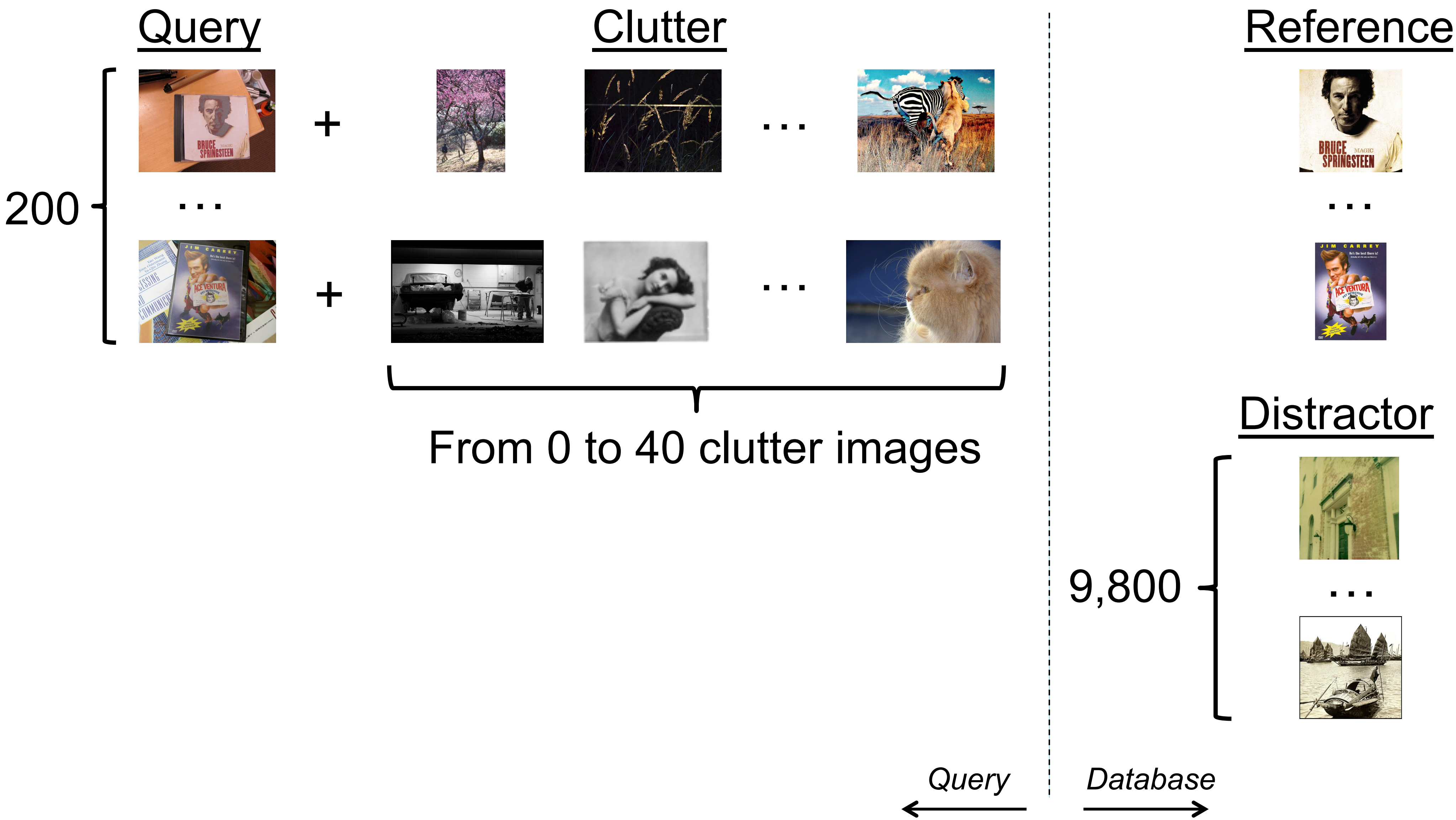
Database

Database image defines projection

Dataset: Query Contained in Database

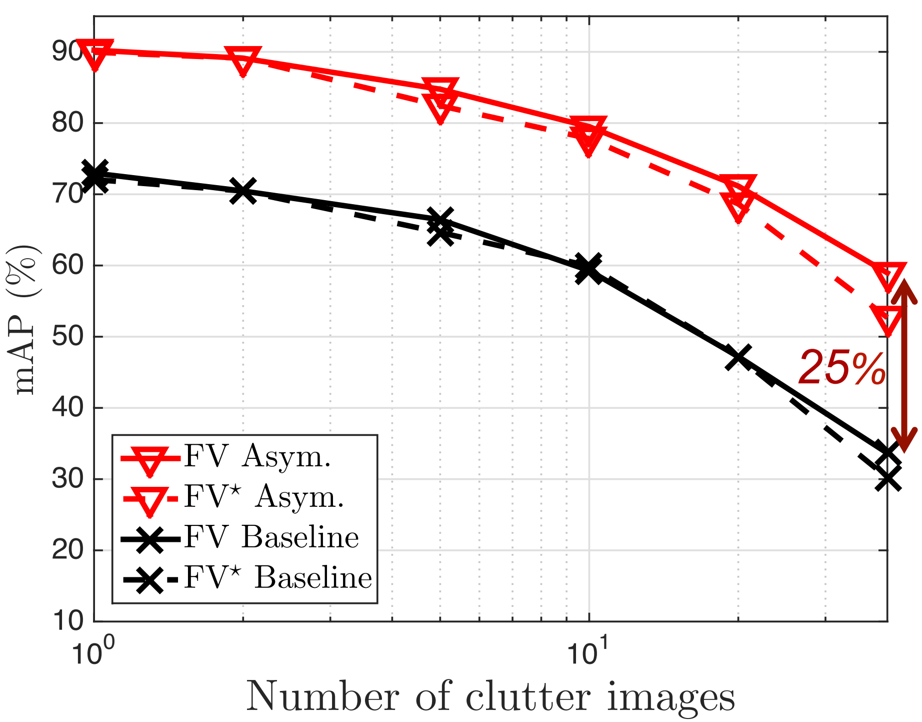


Dataset: Database Contained in Query

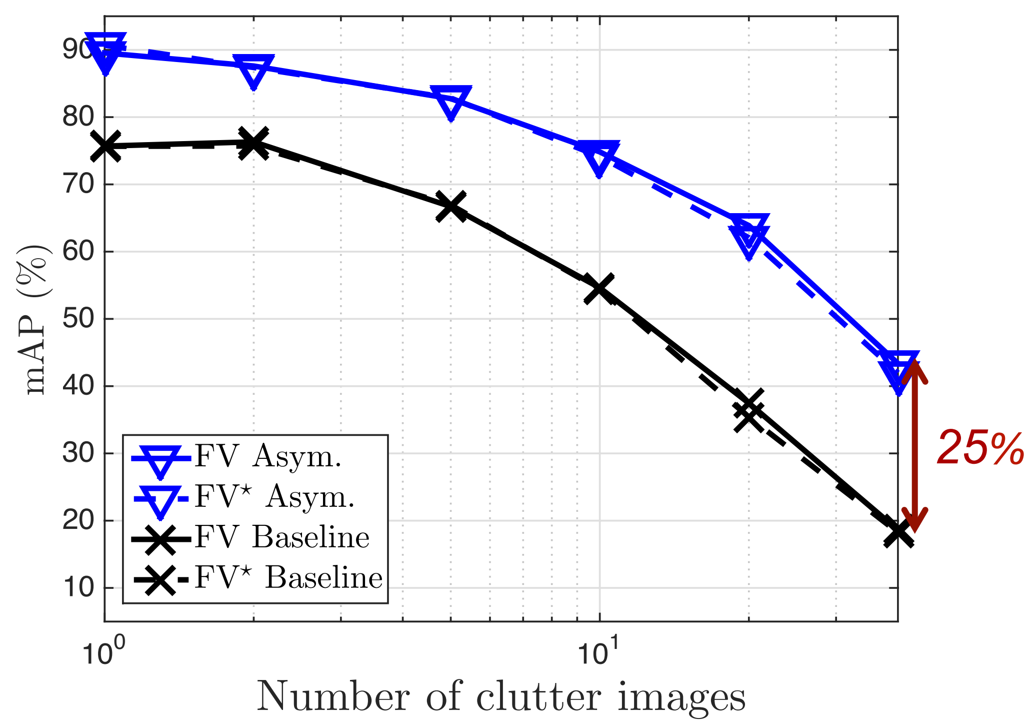


Experiments: Asymmetric FV Comparisons

Query contained in database



Database contained in query



Contribution 2

Fisher Vector Comparisons

- Asymmetric comparisons for Fisher vectors
- Cluttered query or database images

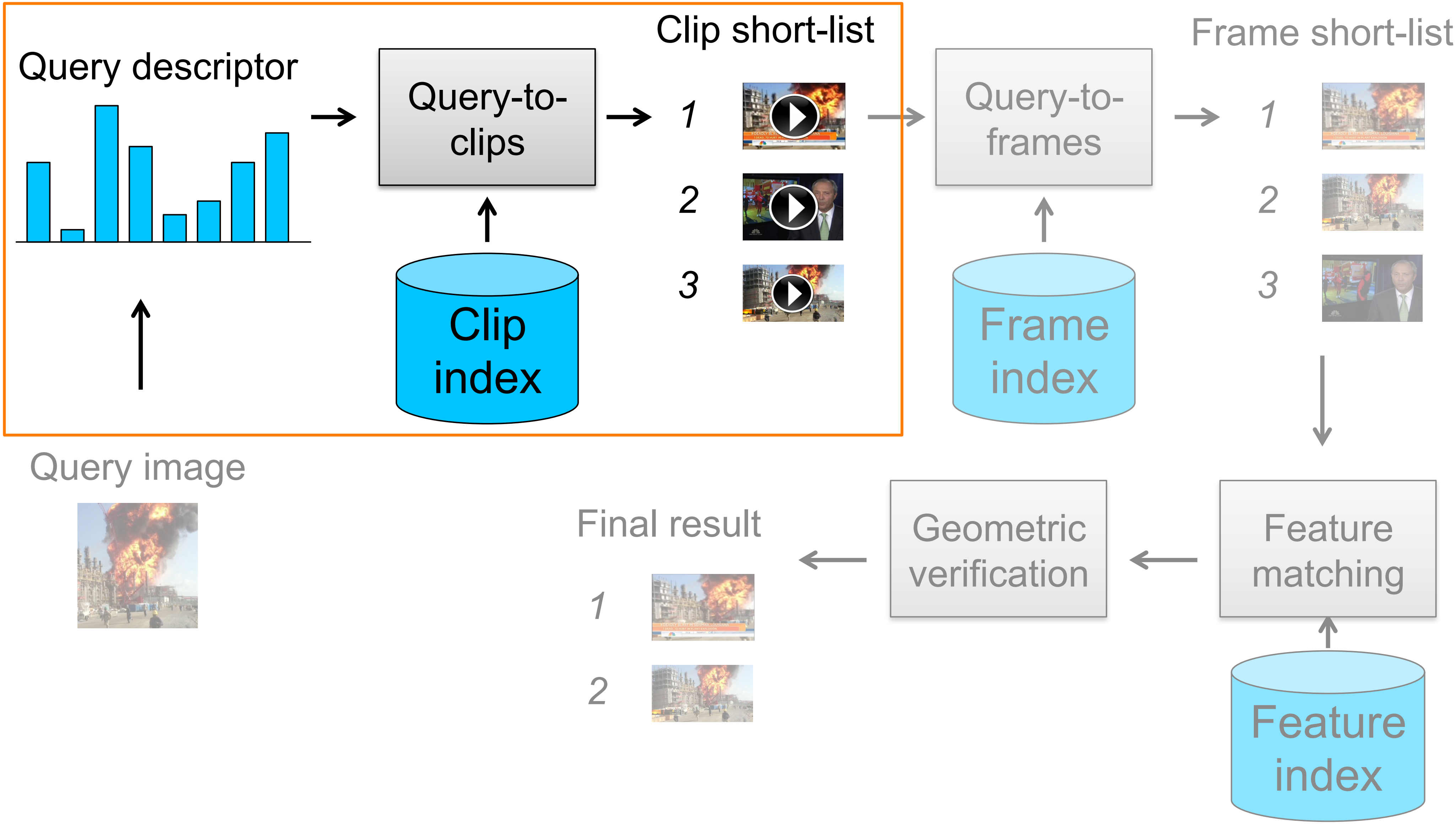
Fisher Vector Aggregation

- Fisher vector descriptors for video segments
- Compact database for large-scale retrieval

Bloom Filter Aggregation

- Bloom filter descriptors for video segments
- Fast and accurate large-scale retrieval

Large-Scale Architecture



Temporal Structure

<u>Frames</u>

1 fps

<u>Shots</u>

Contain similar frames

Length of seconds

<u>Clips</u>

Contain diverse shots

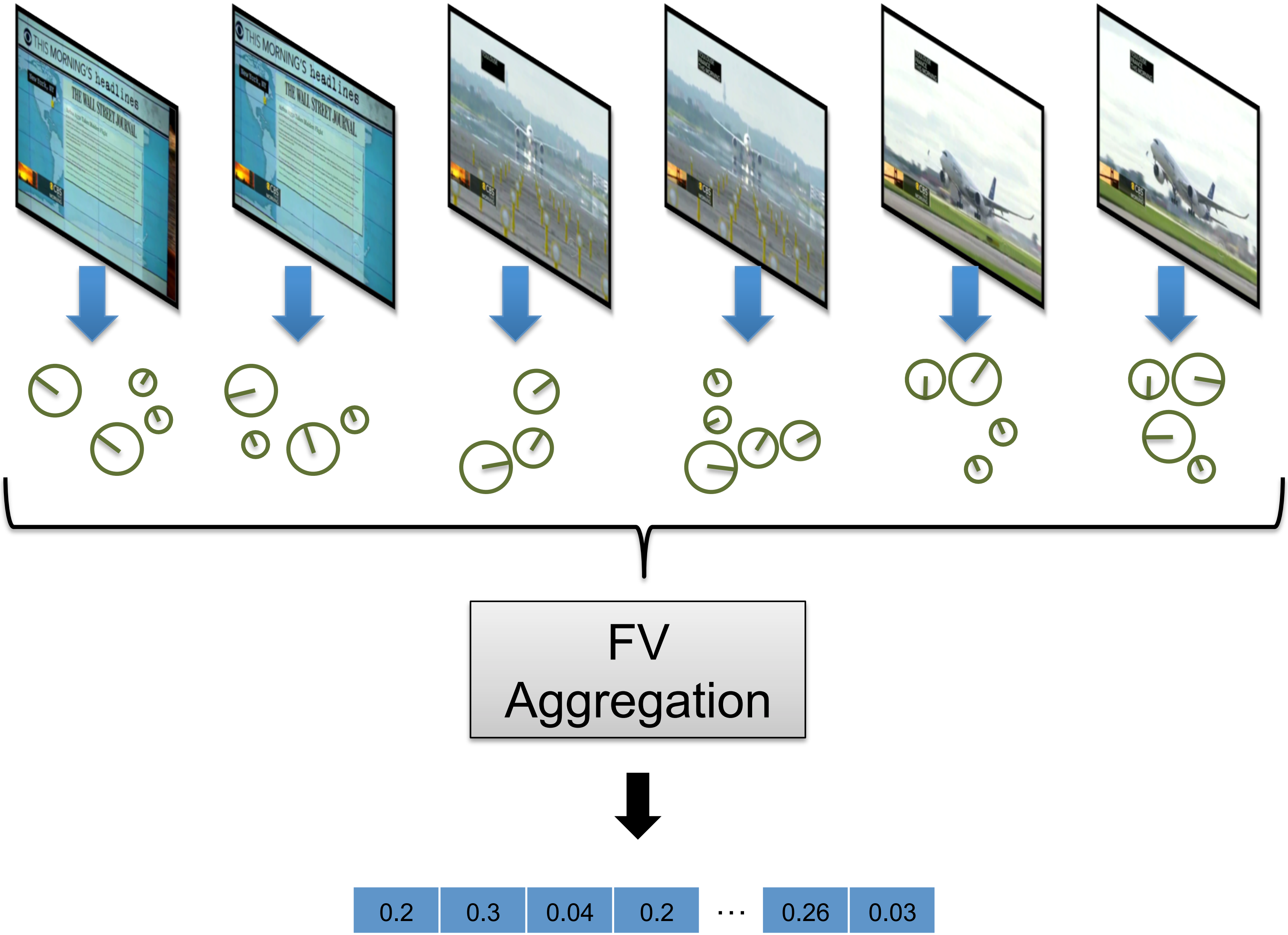
Length of minutes

Clip Fisher Vector

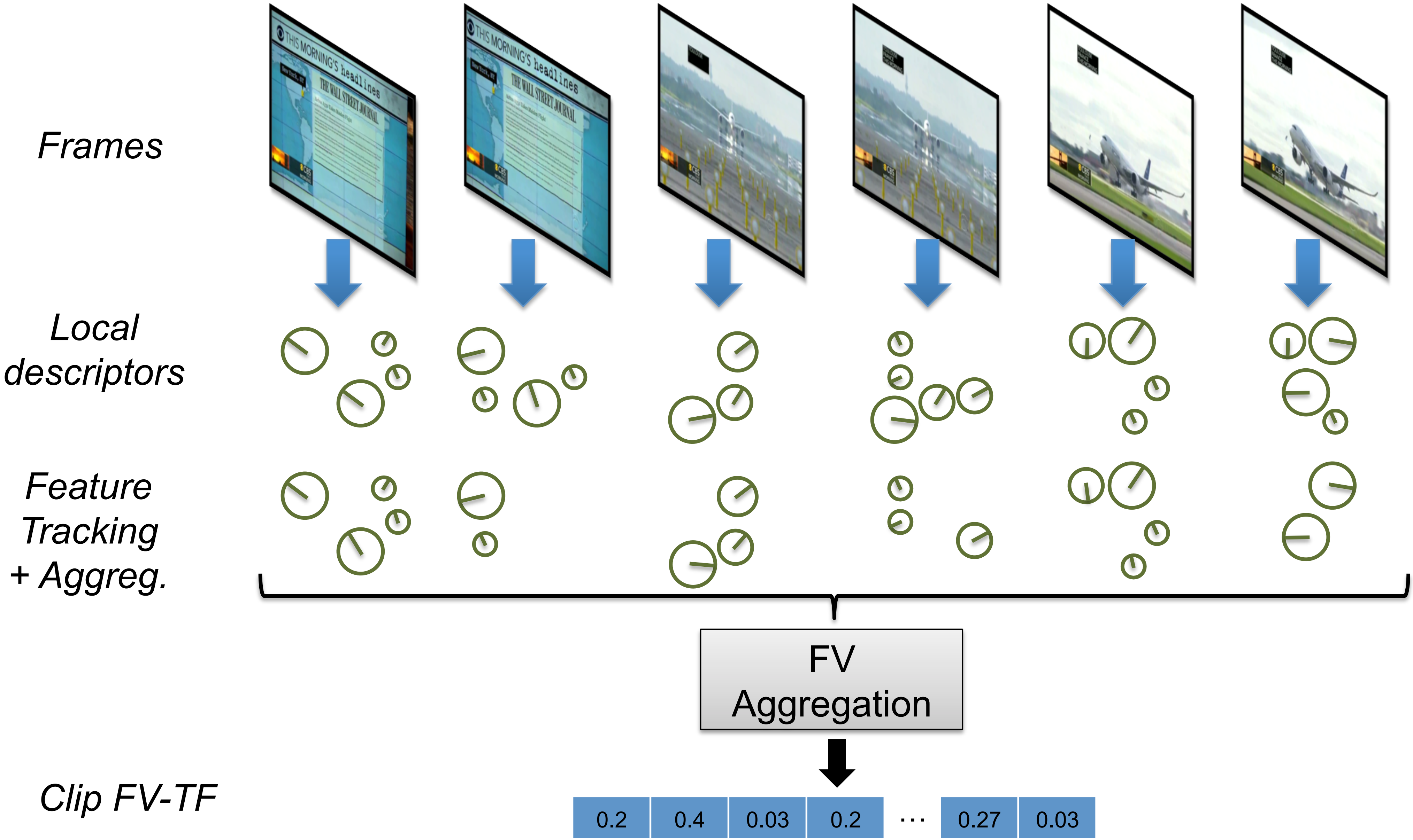
Frames

Local descriptors

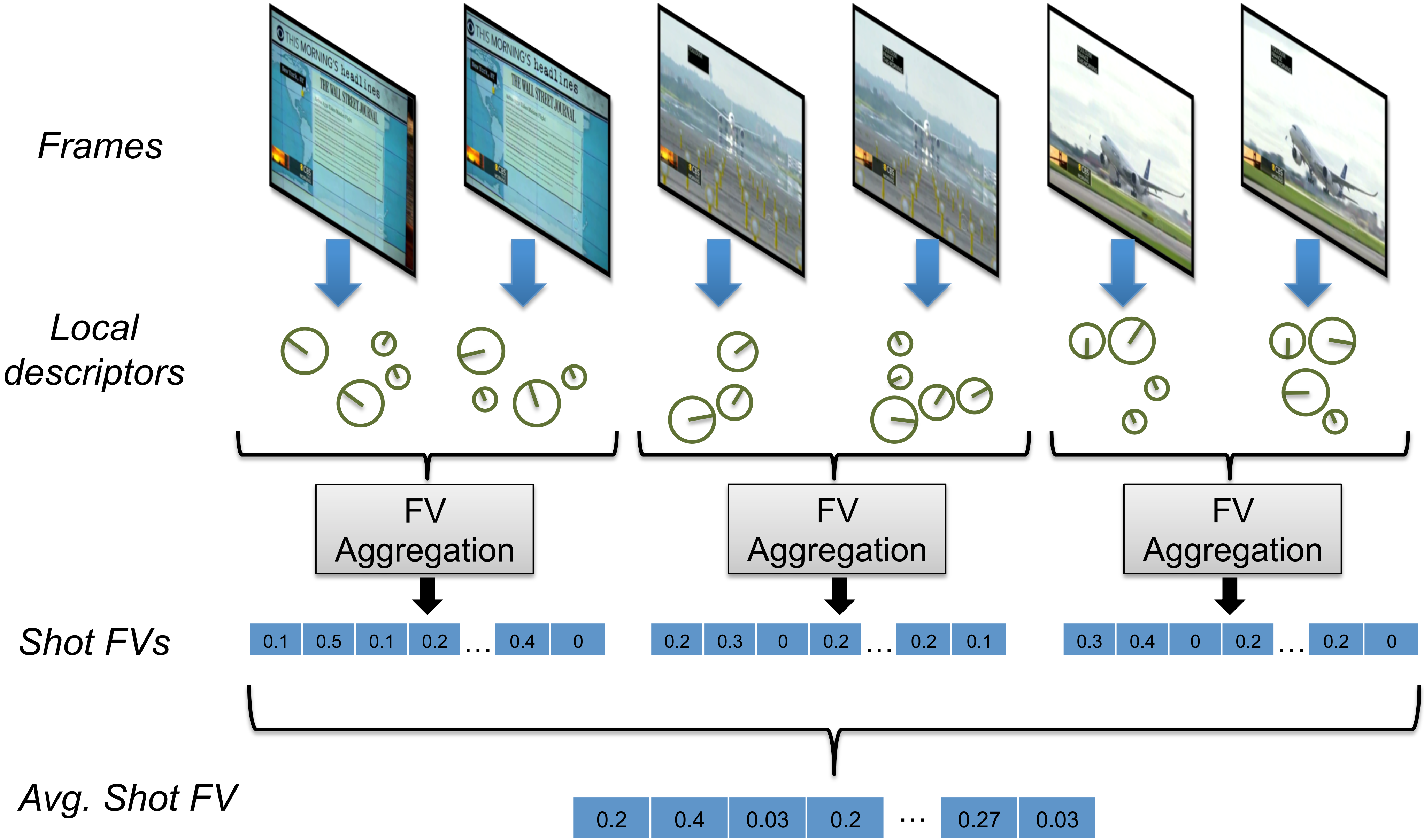
Clip FV



Clip Fisher Vector with Tracked Features



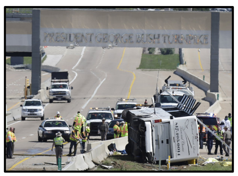
Averaged Shot Fisher Vectors



Datasets

News Videos

Query | *Database*

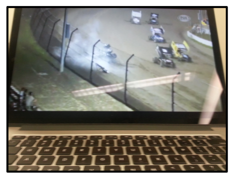
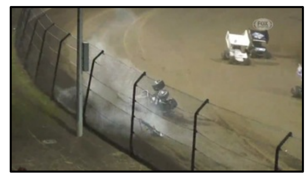



- 229 queries (from web)
- 2.7 minutes/clip
- 50.6 shots/clip
- Versions
  - 600k frames, 164h, 3.4k clips
  - 4M frames, 1,079h, 24.3k clips

Video Bookmarking

Query | *Database*

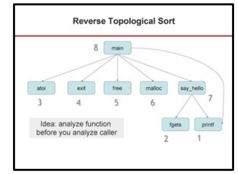


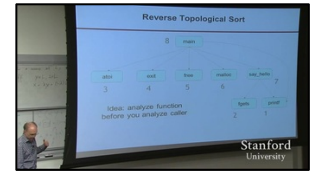




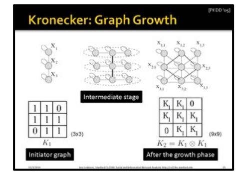
- 282 queries (smartphone pics)
- 2.7 minutes/clip
- 50.6 shots/clip
- Versions
  - 600k frames, 164h, 3.4k clips
  - 4M frames, 1,079h, 24.3k clips

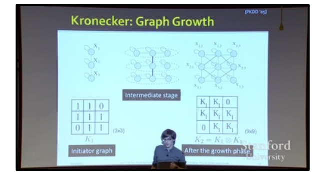
Lecture Videos

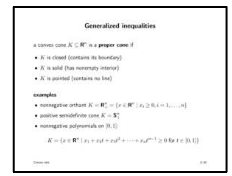
Query | *Database*

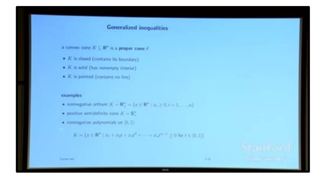






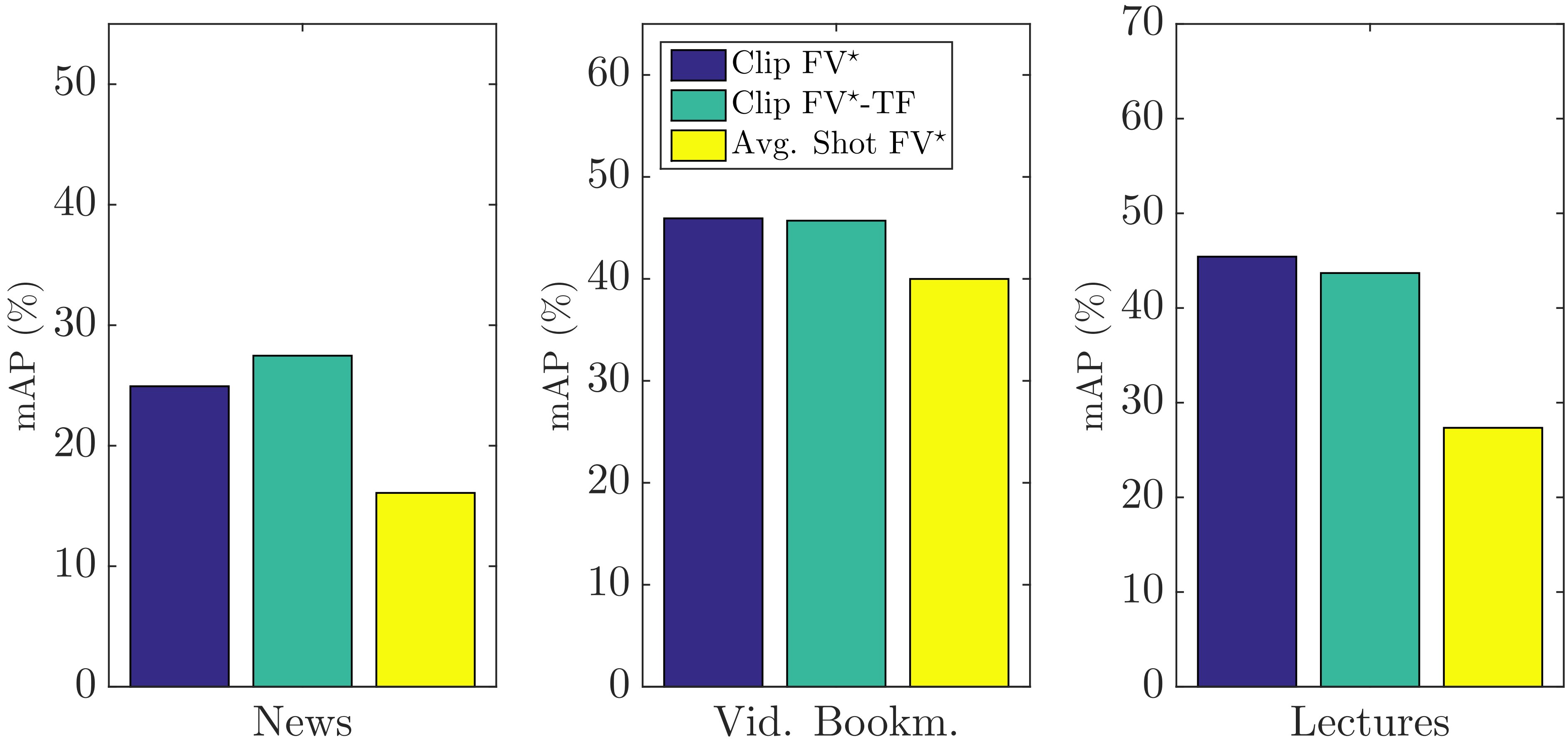




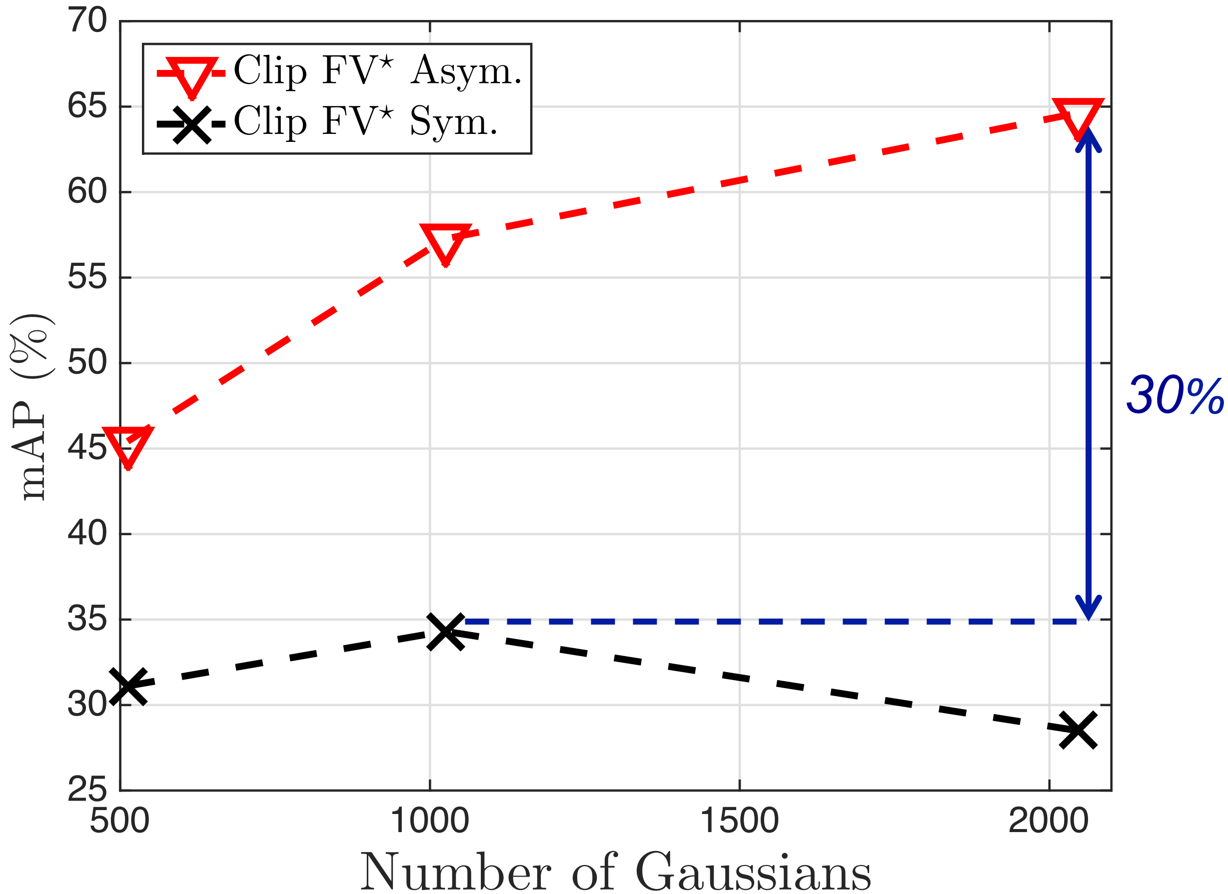


- 258 queries (slides)
- 8.2 minutes/clip
- 58.8 shots/clip
- Versions
  - 600k frames, 169h, 1.1k clips
  - 1.5M frames, 408h, 2.9k clips

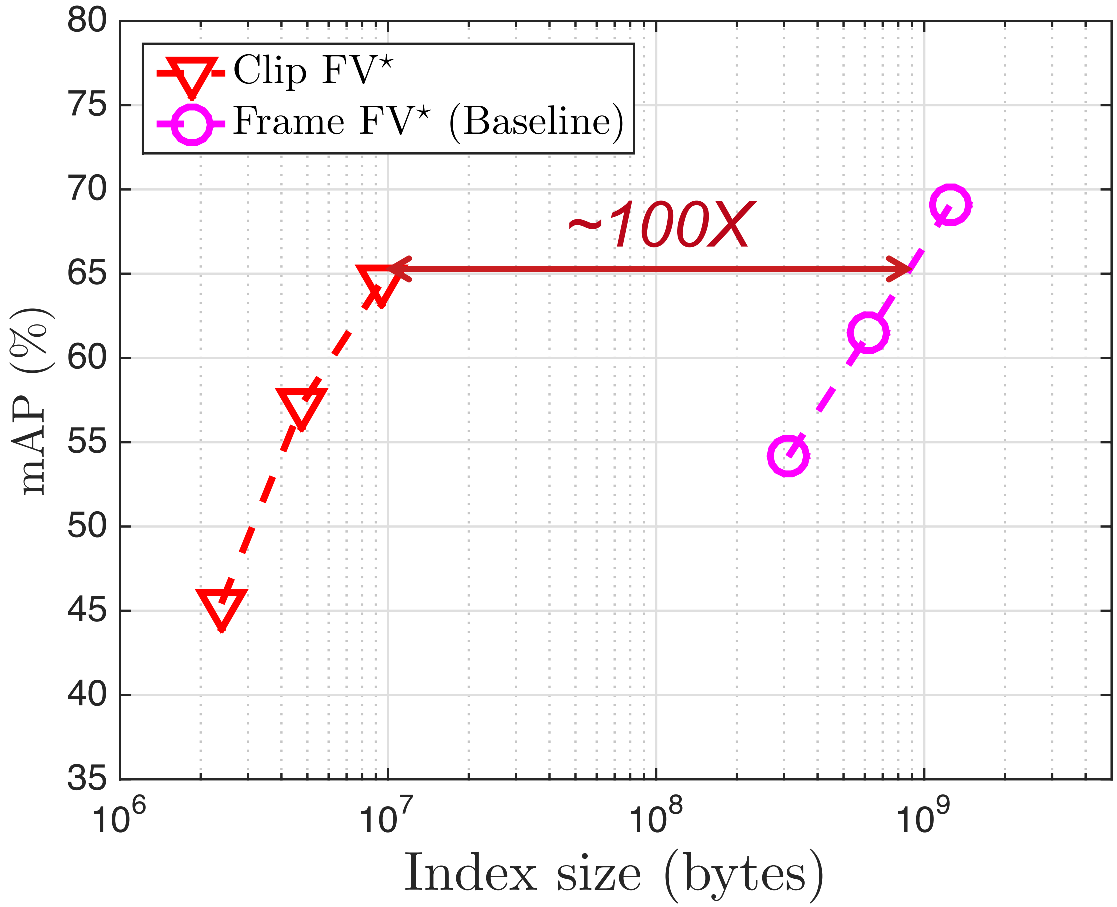
Experiments: Comparison of Techniques



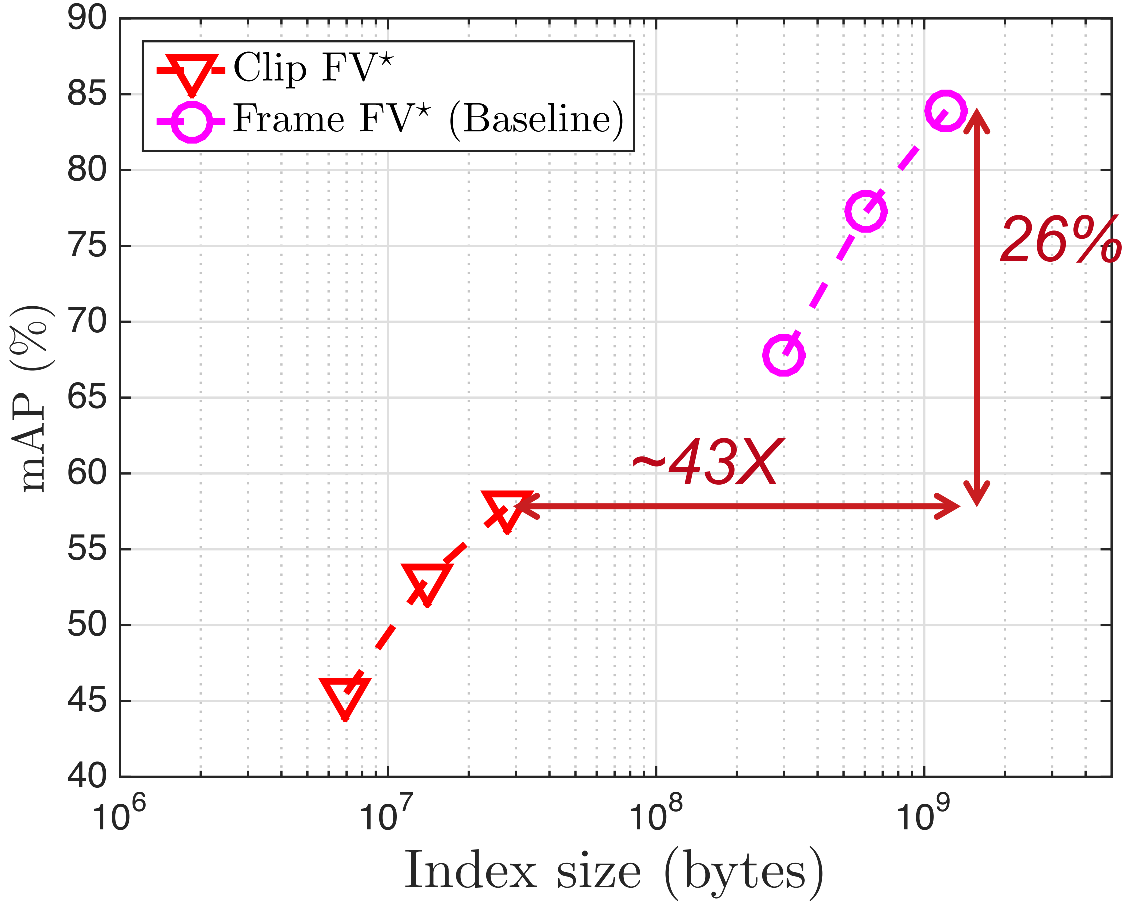
Experiments: Lecture Videos Dataset



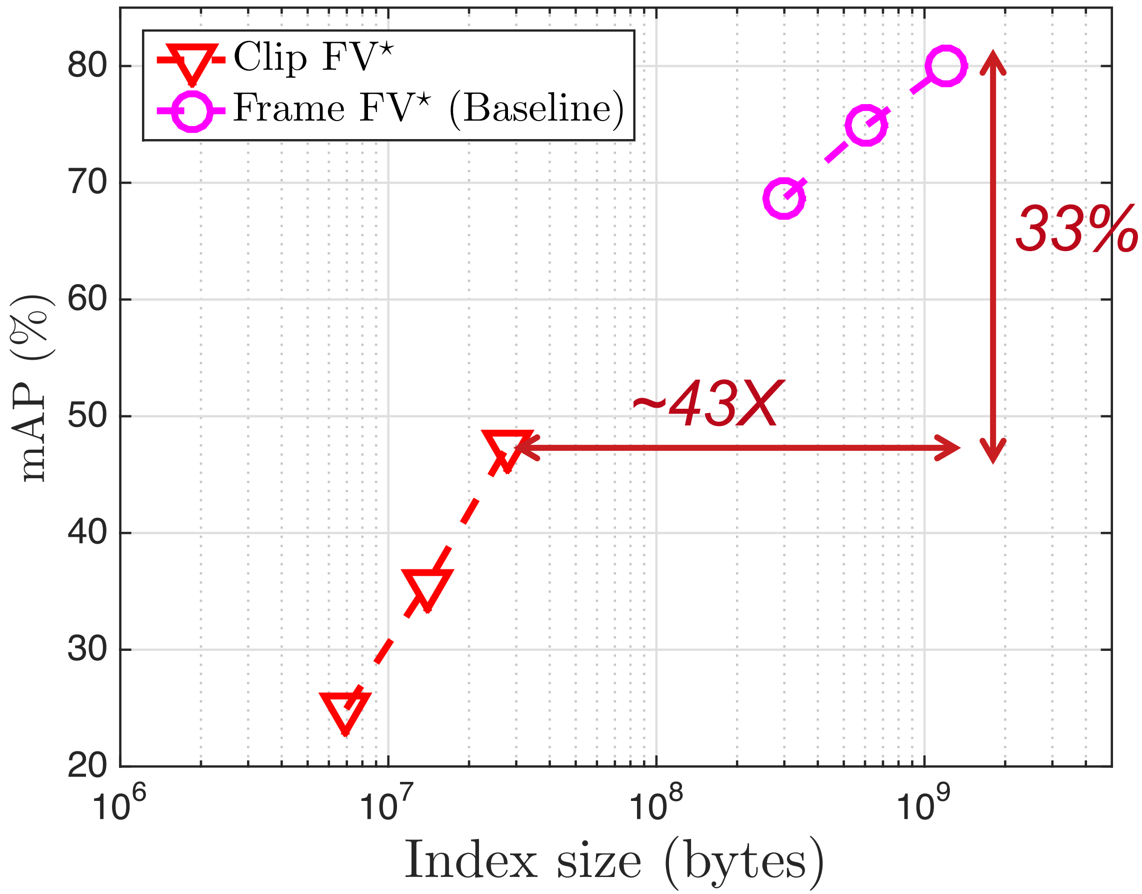
Experiments: Lecture Videos Dataset



Experiments: Video Bookmarking Dataset



Experiments: News Videos Dataset



Contribution 3

Fisher Vector Comparisons

- Asymmetric comparisons for Fisher vectors
- Cluttered query or database images

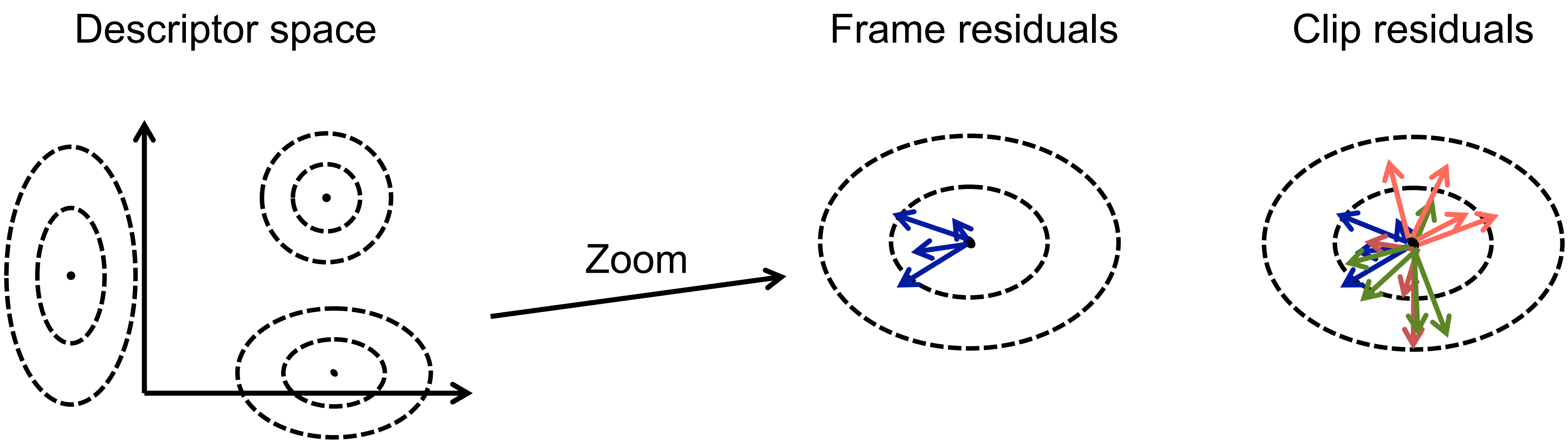
Fisher Vector Aggregation

- Fisher vector descriptors for video segments
- Compact database for large-scale retrieval

Bloom Filter Aggregation

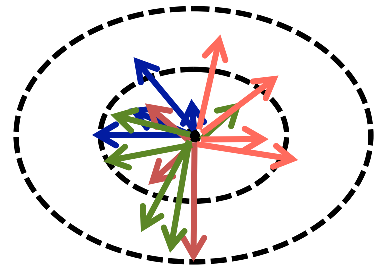
- Bloom filter descriptors for video segments
- Fast and accurate large-scale retrieval

Aggregation Methods

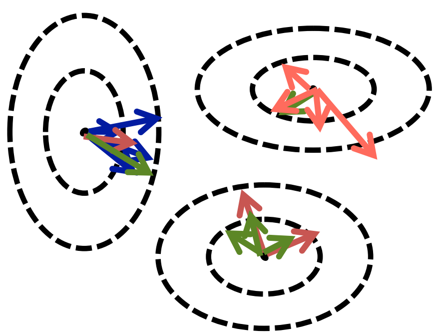


Aggregation Methods

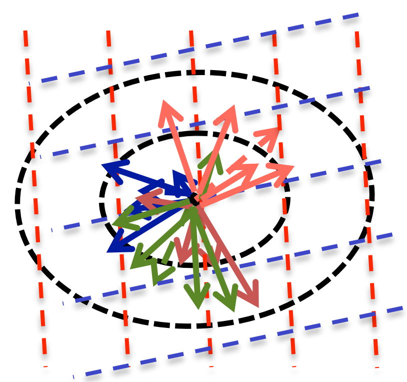
Clip residuals



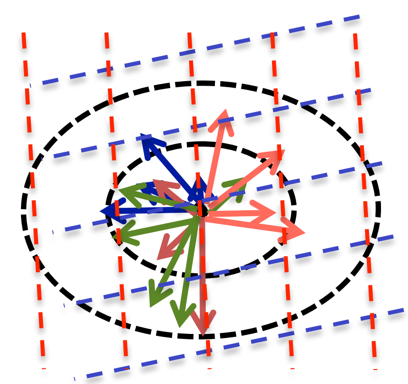
1) Clip FV → increase number of Gaussians



2) Frame FV + temporal aggregation by hashing



3) Spatio-temporal hashing



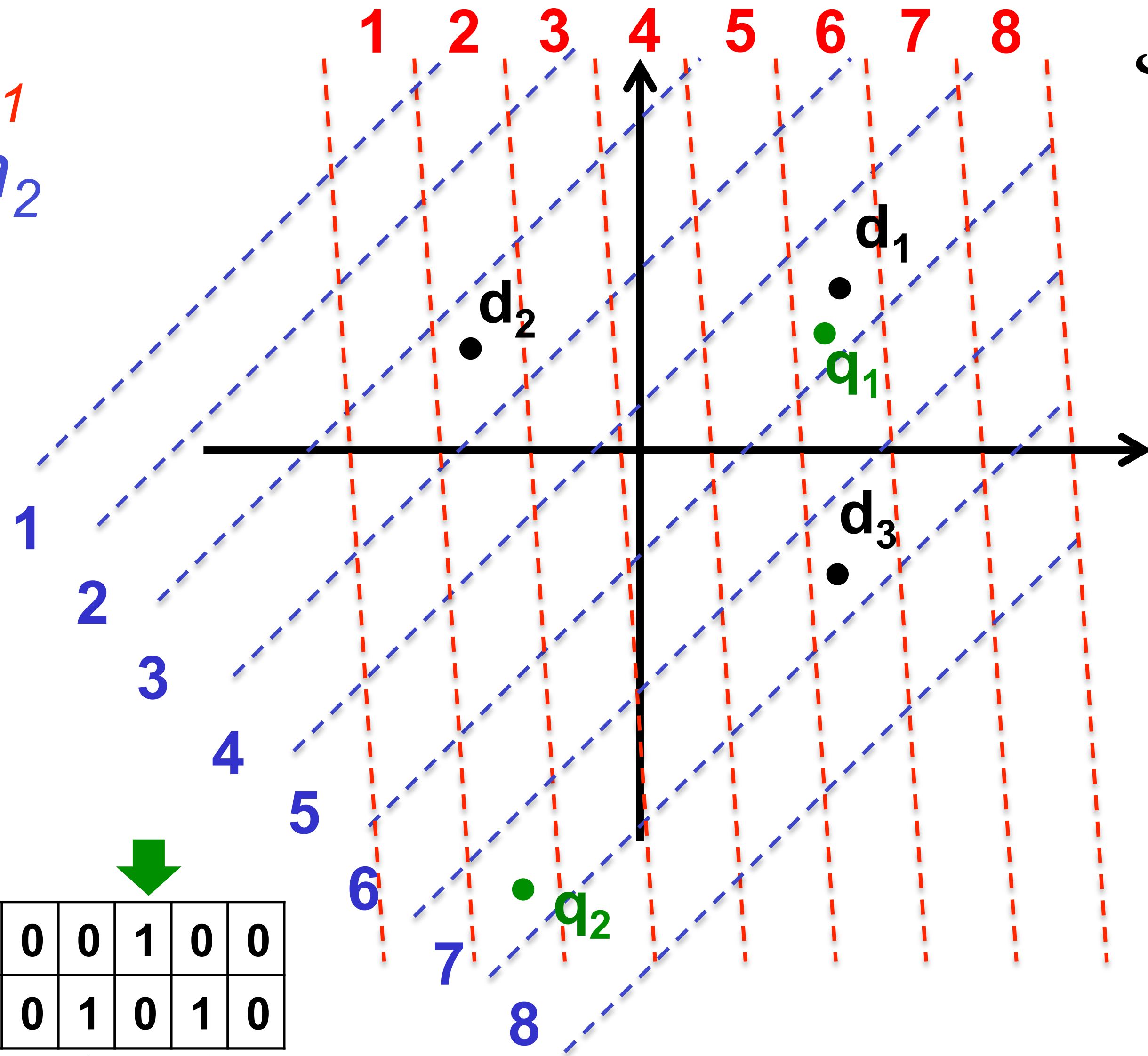
Aggregation using Bloom filters

Bloom Filter (BF)

Red: h_1

Blue: h_2

$\mathcal{S} = \{ d_1 , d_2 , d_3 \}$

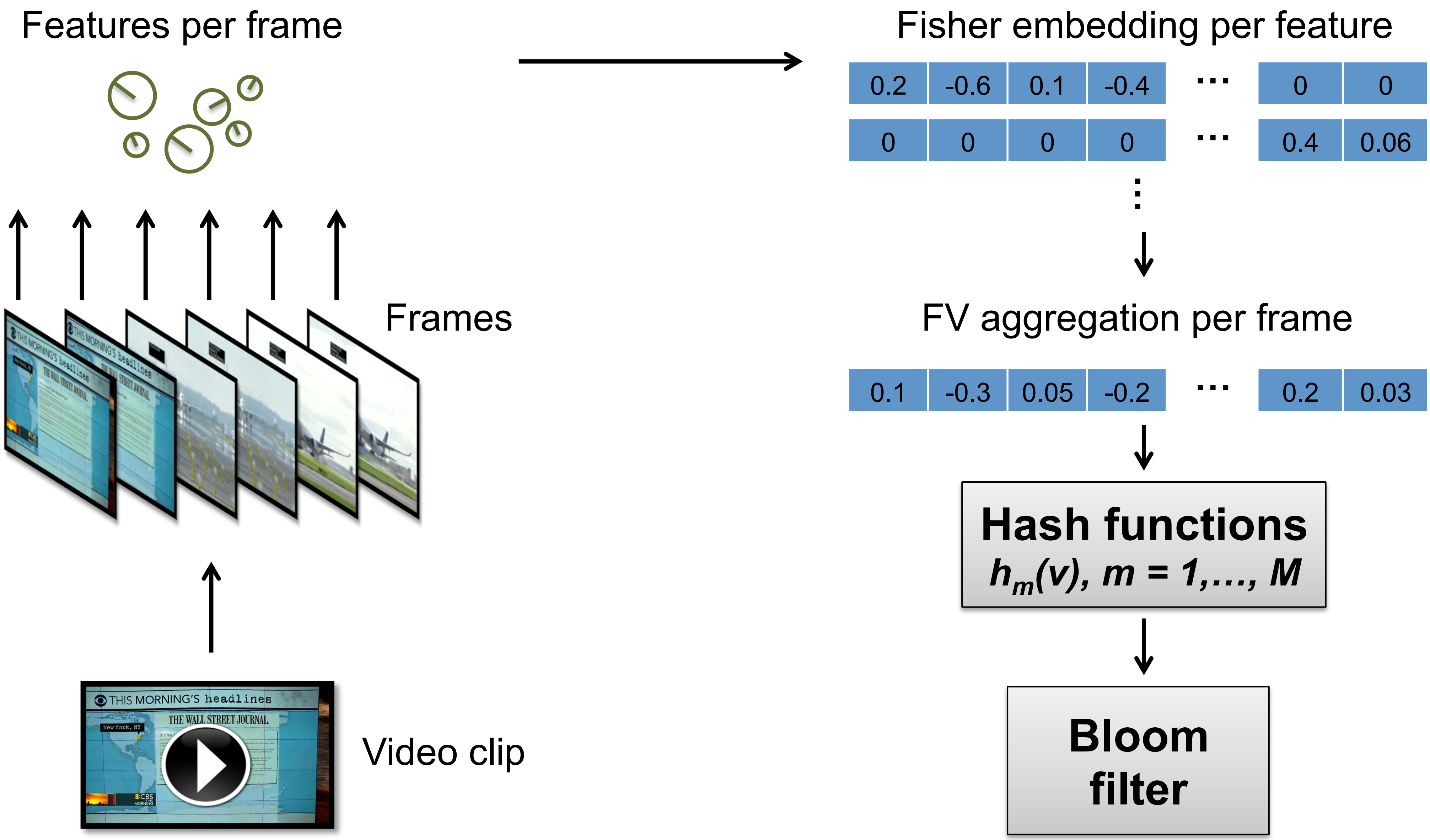


q_1: 2 matches (TP)

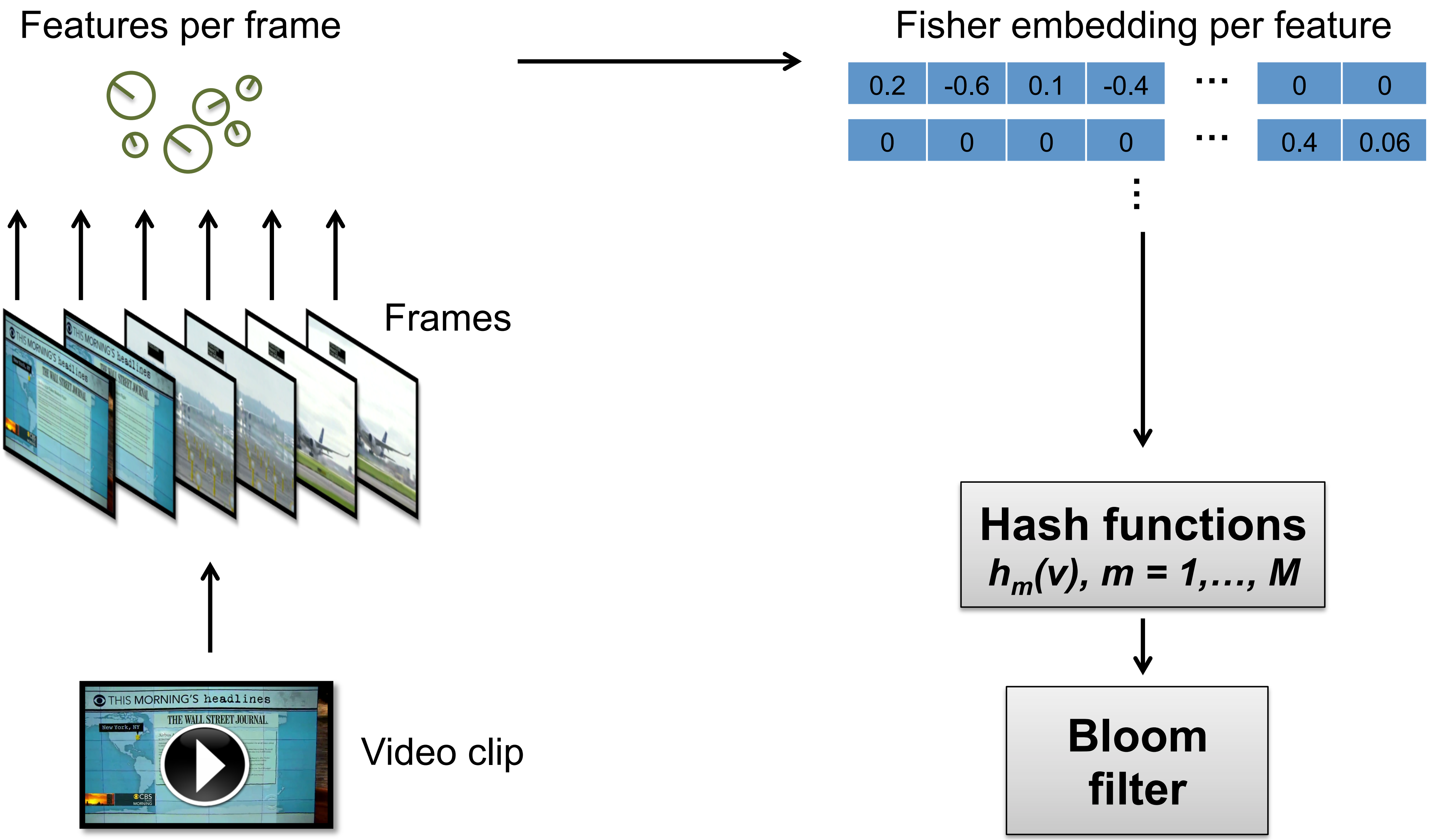
q_2: 2 matches (FP)

b_1	0	1	0	0	0	1	0	0
b_2	0	0	1	0	1	0	1	0

BF using Global Descriptors (BF-GD)

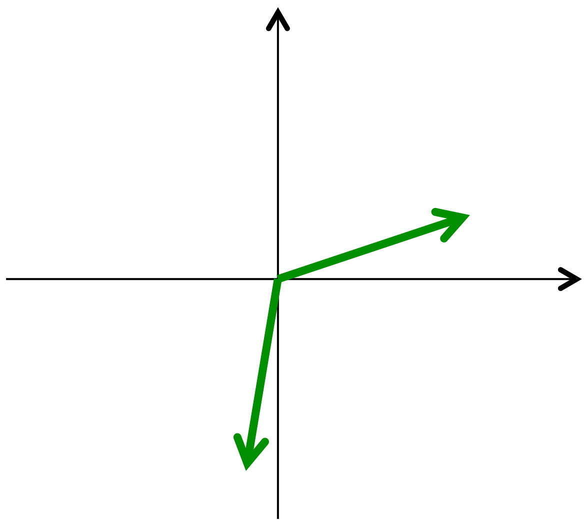


BF using Point-Indexed Descriptors (BF-PI)



Hash Functions

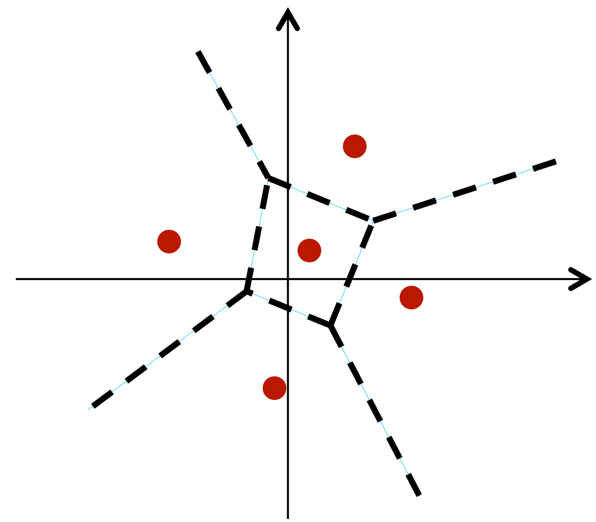
Locality-Sensitive Hashing (LSH)



Random hyperplanes

One hyperplane per bit

Vector Quantization (VQ)

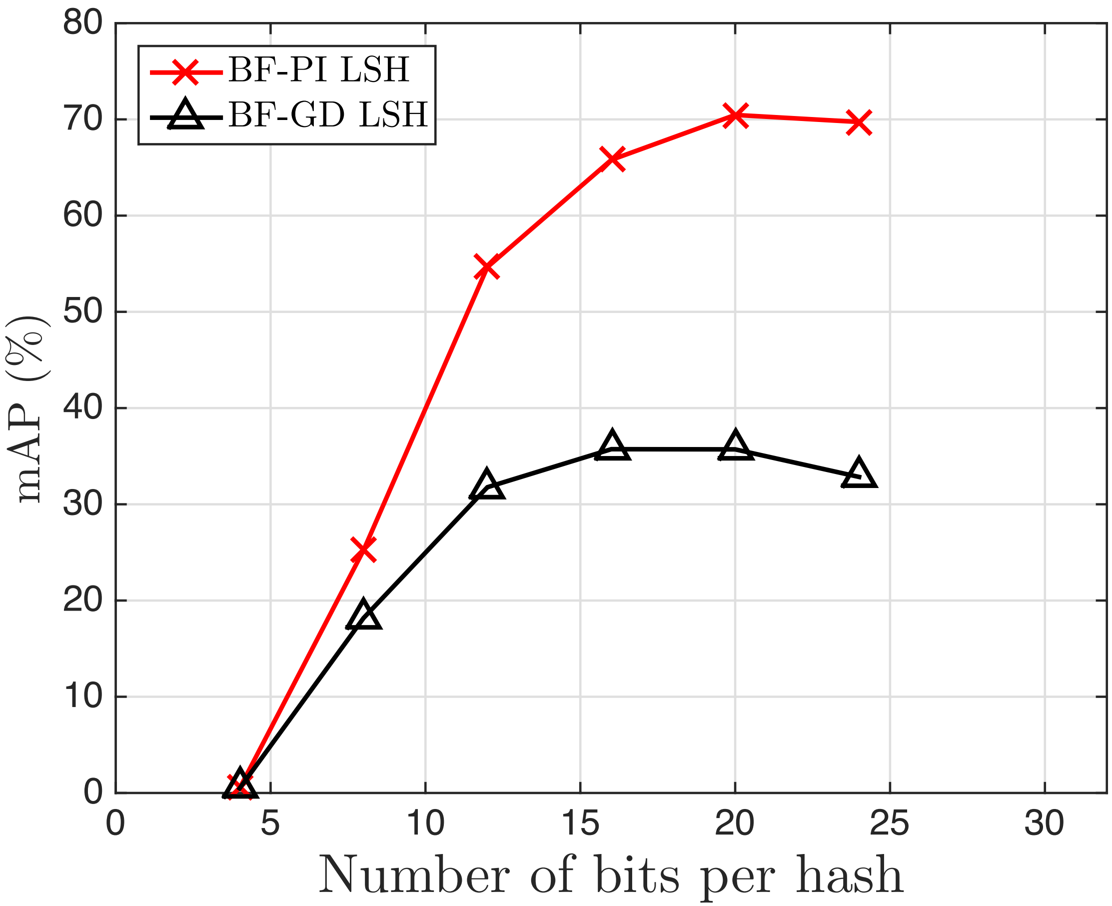


Trained using
Approximate K-Means

bits = $\log_2$(# centroids)

Experiments: BF-GD vs BF-PI

Dataset: News Videos – 600k

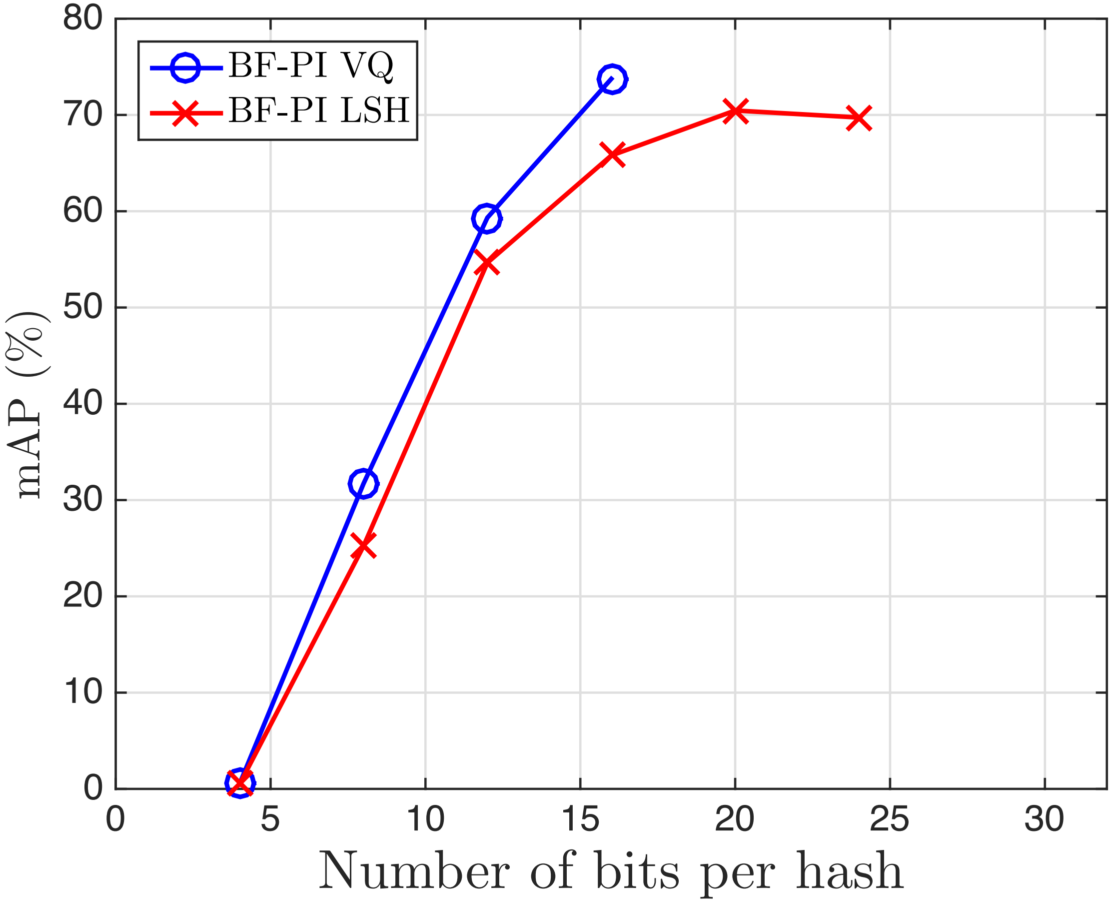


Visual Invariance

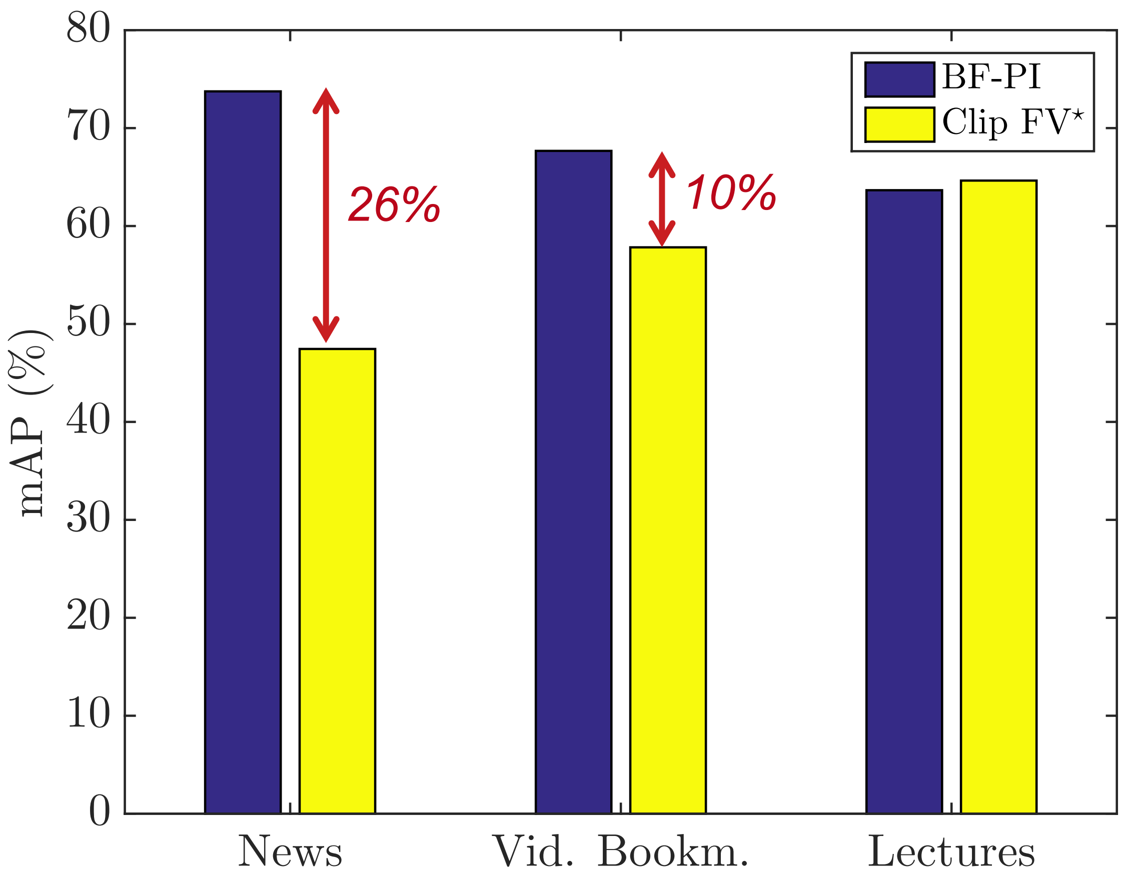
Visual Discriminativeness

Experiments: BF-PI with Different Hashes

<u>Dataset</u>: News Videos – 600k

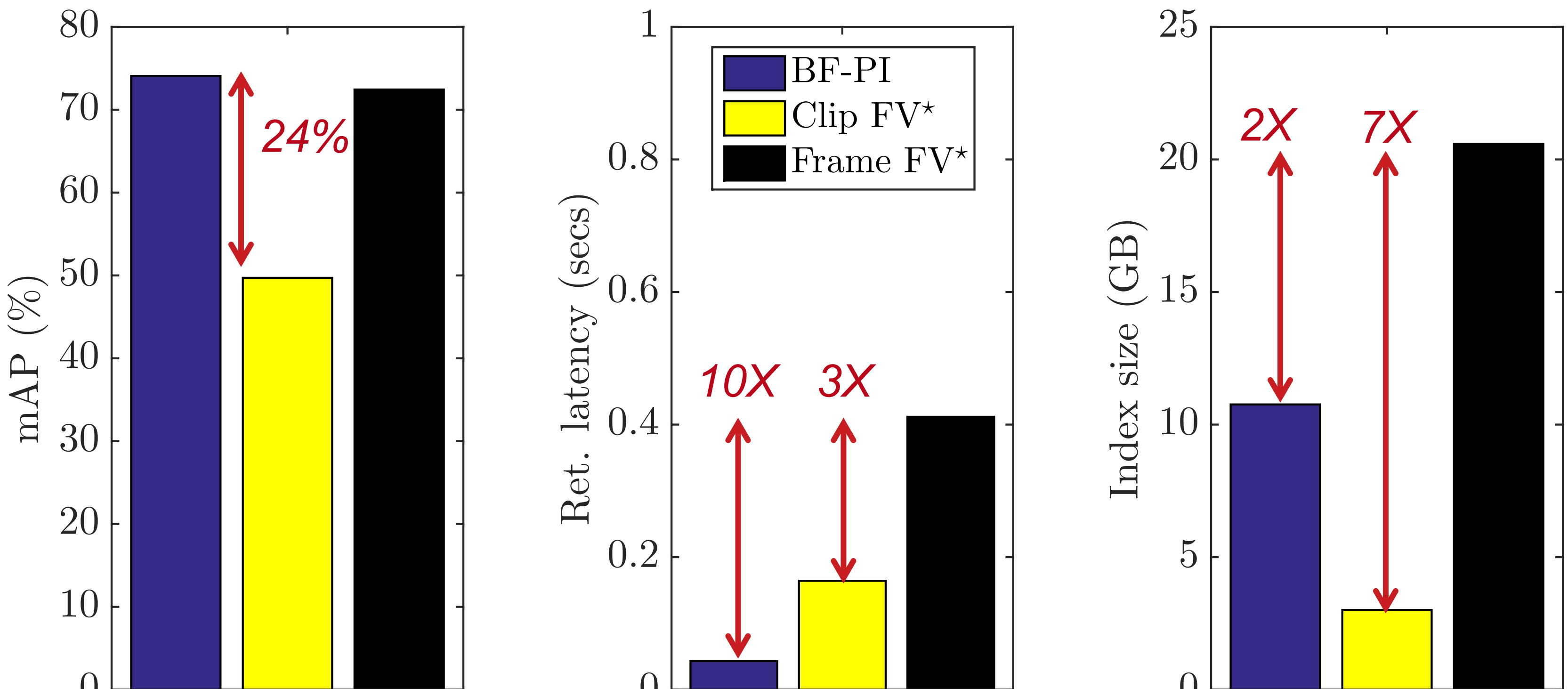


Experiments: Results on 600k Datasets



Experiments: Large-Scale with Re-Ranking

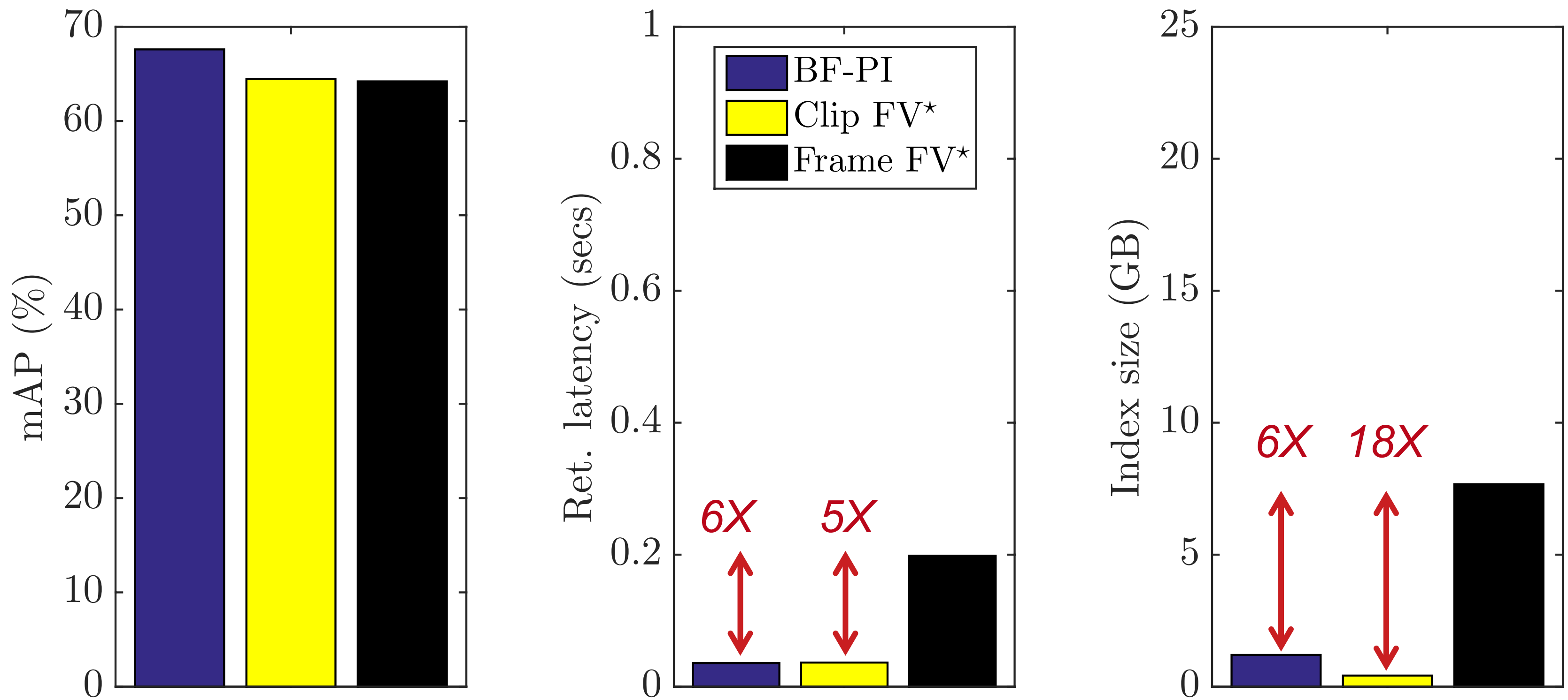
<u>Dataset</u>: News Videos – 4M



BF-PI and *Clip FV** results are re-ranked using *Shot-FV** descriptors

Experiments: Large-Scale with Re-Ranking

<u>Dataset</u>: Lecture Videos – 1.5M



BF-PI and *Clip FV** results are re-ranked using *Shot-FV** descriptors

Conclusions

Fisher Vector Comparisons

- Asymmetric comparisons by projecting cluttered FVs
- Studied two asymmetric retrieval problems
- Large retrieval gains (up to 25% mAP) in both cases

Fisher Vector Aggregation

- Fisher vector aggregation over video segments
- Simple aggregation outperforms other techniques
- Effective retrieval with 100X compression for lectures dataset

Bloom Filter Aggregation

- Bloom filter aggregation over video segments
- Studied hash functions and spatio-temporal aggregation schemes
- Lighter, faster than frame-based schemes, with similar accuracy