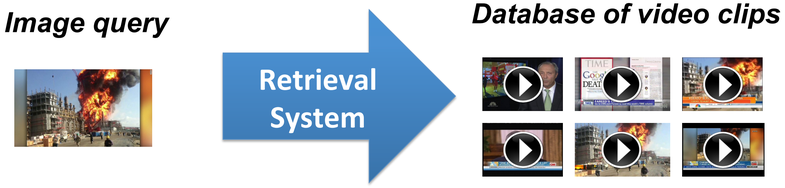
Temporal Aggregation for Large-Scale Query-by-Image Video Retrieval

André Araujo, Jason Chaves, Roland Angst and Bernd Girod
Department of Electrical Engineering, Stanford University, USA
http://stanford.edu/~afaraujo/

Query-by-Image Video Retrieval



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

Challenges → Our contributions

Temporal redundancy → Shot aggregation
Large database (scalability) → Scene aggregation
Query-database asymmetry → Asymmetric comparisons

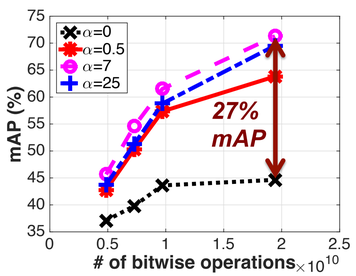
Result: 10X faster retrieval with similar mAP compared to state-of-the-art baseline

Experiments

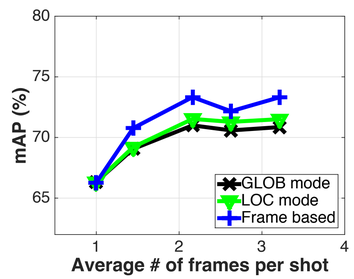
- SIFT local features + SCFV global descriptors
- Stanford I2V dataset
  - Light version: *78* queries and *1,035*h of video in database
  - Full version: *229* queries and *3,801*h of video in database

Small-scale experiments (light dataset)

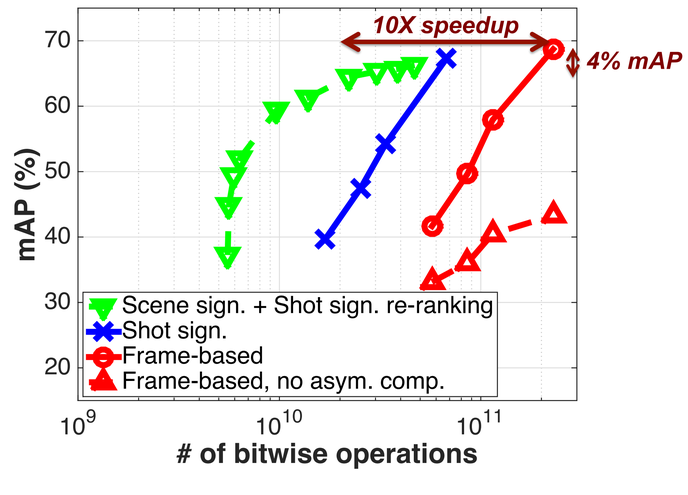
Asymmetric comparisons
(shot signatures w/ LOC mode)



Shot aggregation modes

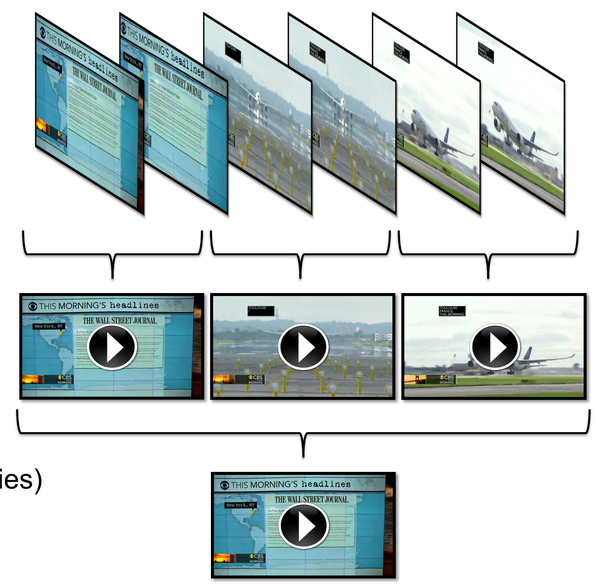


Large-scale experiments (full dataset)



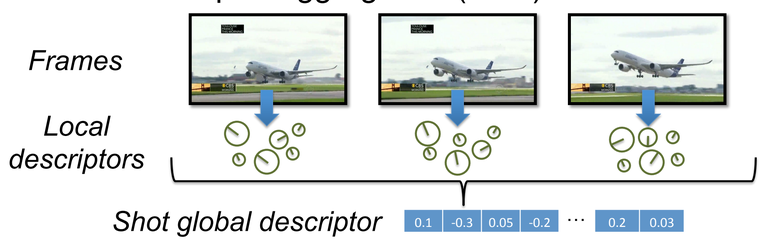
Temporal Aggregation

Temporal fragments

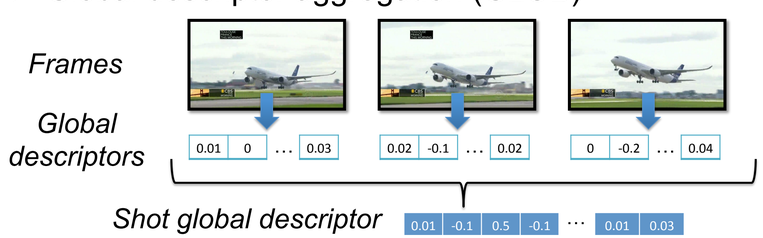


Shot/Scene aggregation modes

- Local descriptor aggregation (LOC)



- Global descriptor aggregation (GLOB)



- Other modes also described in the paper

Asymmetric Comparisons

Examples: query images and database frames

- Asymmetry even more pronounced when using temporally-aggregated signatures in database

Solution for Fisher vector-like signatures

d — Dimensionality of local descriptors after PCA

$\mathcal{G}_i$ — Query FV's d-dimensional residual for Gaussian i

Rule:
Ignore the d components of $\mathcal{G}_i$ in score computation if:

$$\|\mathcal{G}_i\|_1 \leq \alpha$$

(α: threshold)