EFFECTIVE FISHER VECTOR AGGREGATION FOR 3D OBJECT RETRIEVAL

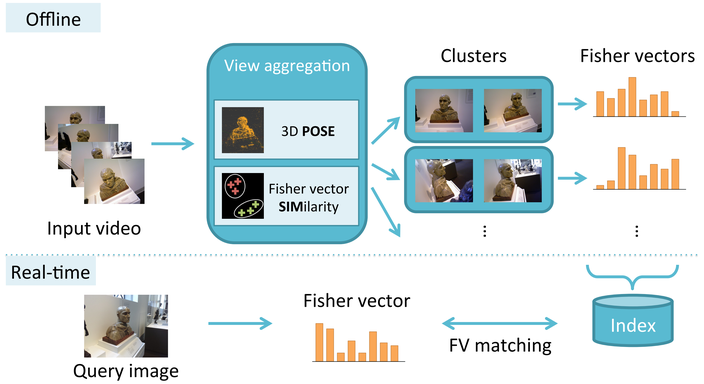
Jean-Baptiste Boin* (jboin@stanford.edu), André Araujo*, Lamberto Ballan†*, Bernd Girod*

*Department of Electrical Engineering, Stanford University, USA, †Media Integration and Communication Center, University of Florence, Italy

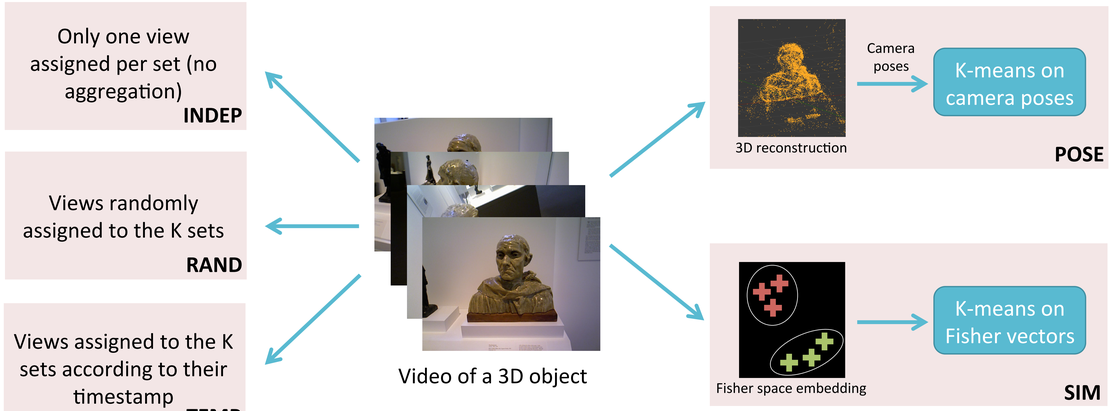
In a nutshell...

- 3D object retrieval
  - Database: **videos** of objects captured manually from different viewpoints
  - Query: single **image**.
- Videos are represented as K **Fisher vectors** (FV) aggregating information from **different viewpoints**.
- Two proposed frame **aggregation methods** are shown to considerably **outperform** standard techniques (with or without frame aggregation) on **large-scale experiments** using an existing video dataset.

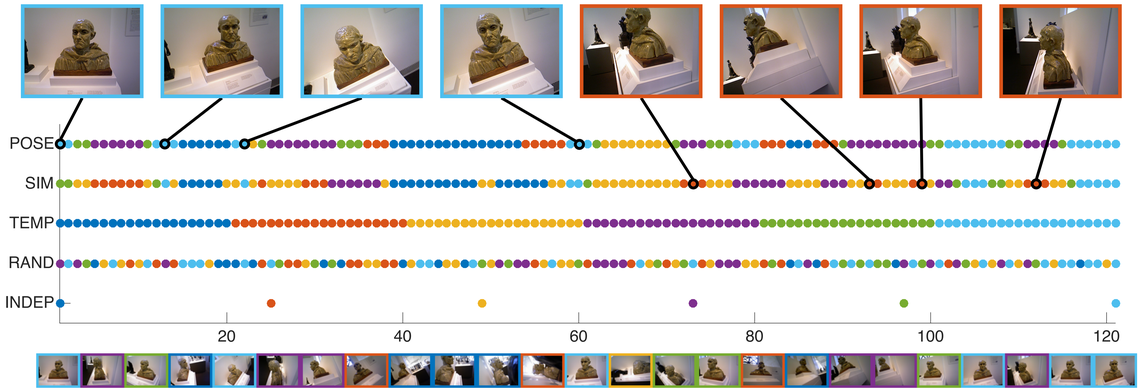
Overview of the system



View aggregation methods

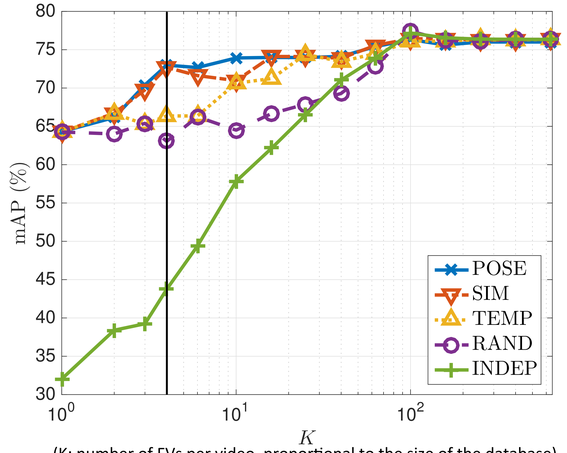


Results of the view aggregation on one video sequence



- The frames are visually similar within a cluster for POSE and SIM
- The distribution is more noisy for SIM

Evaluation



(K: number of FVs per video, proportional to the size of the database)
FVs are generated with 512 Gaussians

Method	128 Gaussians	256 Gaussians	512 Gaussians
INDEP	37.79%	38.75%	43.75%
RAND	56.68%	62.51%	63.04%
TEMP	58.86%	64.34%	66.36%
SIM (ours)	**66.58%**	**70.65%**	72.69%
POSE (ours)	64.63%	69.40%	**72.98%**

mAP for a fixed value of K (K=4)

- For K = 4 (fixed database size), the boost in the mAP is significant when using SIM or POSE compared to naïve aggregation schemes.
- With 512 Gaussians, 71+% mAP target achieved for:
  - K = 4 for POSE and SIM
  - K = 40 for INDEP (**10X** compression)
  - K = 16 for TEMP (**4X** compression).

References

[1] S. Choi, Q.-Y. Zhou, S. Miller, and V. Koltun, "A large dataset of object scans," arXiv:1602.02481, 2016.

[2] A. Araujo, J. Chaves, R. Angst, and B. Girod, "Temporal Aggregation for Large-Scale Query-by-Image Video Retrieval," in Proc. ICIP, 2015.