Efficient Video Search Using Image Queries

A. Araujo[1], M. Makar[2], V. Chandrasekhar[3], D. Chen[1], S. Tsai[1], H. Chen[1], R. Angst[1] and B. Girod[1]

[1]Stanford University, USA [2]Qualcomm Inc., USA [3]Institute for Infocomm Research, Singapore

Image-based Search of Videos

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

Challenges:

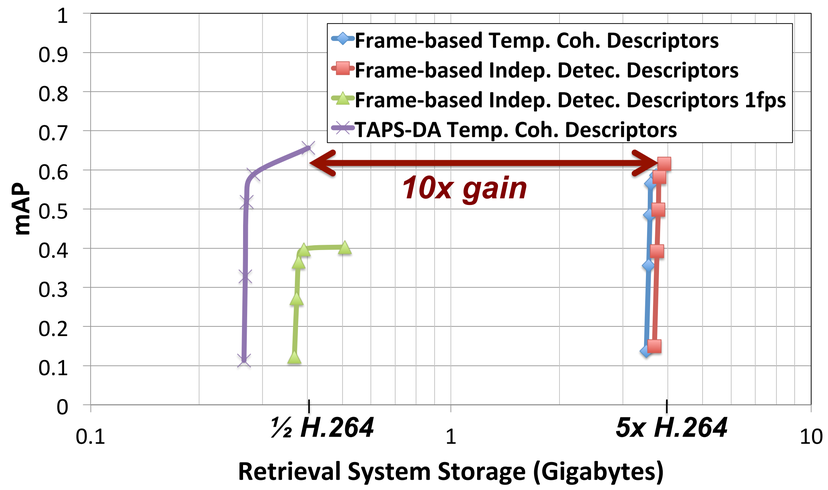
- High index memory requirements
- Can we use temporal information to improve retrieval?

Contribution:

One order of magnitude database storage reduction, with moderate boost of mean average precision (mAP)

Experiments

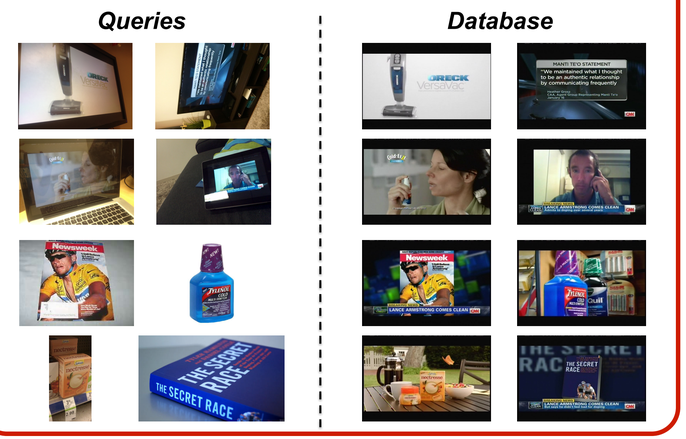
- SIFT features
- Vocabulary tree with varying size
- Geometric Verification using RANSAC



New Dataset: CNN2h

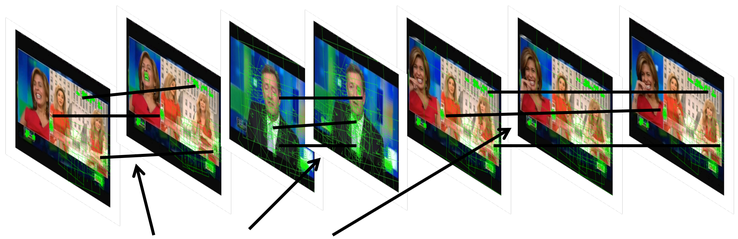
Available online!

- 2 hours of CNN video at 10fps
- 139 queries, 72,000 database frames
- 2,951 true matching image pairs
- 21,412 false matching image pairs



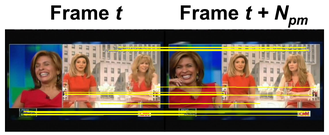
Temporally Aggregated Patch Set (TAPS)

Constructing TAPS: Keypoint Detection and Tracking



Temporally Coherent Keypoint Detection (*Makar et al, 2012*)

Long-range tracking within a sliding window of N_{pm} frames by feature matching (*Sivic & Zisserman, 2006*)



This finds tracks across shots or occlusions

TAPS Description: Descriptor Average (DA)

- Image Pairwise Matching

$$E_{\mathbf{u}}[\|\mathbf{q}-\mathbf{u}\|^2] = \|\mathbf{q}-\boldsymbol{\mu}\|^2 + \sum_i \sigma_i^2$$

Mean sq. distance between query **q** and database descriptor **u** from a given TAPS

Sq. distance between query **q** and mean database descriptor ***μ*** from a given TAPS

Appearance variance (intra-TAPS variance)

- Image Retrieval using Bag-of-Words

$$\sum_m dist(\mathbf{u}_m, \hat{\mathbf{x}}_{c_m}) = \sum_m E_{\mathbf{u}_m}[\|\mathbf{u}_m - \hat{\mathbf{x}}_{c_m}\|^2]$$

Using mean sq. distance as desired distortion

centroids

$$= \sum_m \|\boldsymbol{\mu}_m - \hat{\mathbf{x}}_{c_m}\|^2 + \sum_{m,i} \sigma_{mi}^2$$

Equivalent to using K-means with each TAPS's means!

TAPS Encoding

- Represent all patches from a TAPS using a single *TAPS descriptor*: the average descriptor
- Storage of descriptors replaced by storage of *TAPS descriptors* + *TAPS numbers*
- Encode *TAPS numbers* using *Predictive Coding* based on *TAPS numbers* from previous frames