METU

Exploiting Local Indexing and Deep Feature Confidence Scores for Fast Image-to-Video Search

Savas Ozkan and Gozde Bozdagi Akar

Middle East Technical University, Department of Electrical/Electronics Engineering, 06800, Ankara, Turkey

ICPR 2020 — 25th INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION — Milan, Italy 10 | 15 January 2021

Motivation

We have two main motivations in our paper:

1) To boost the visual search for severe visual challenges, individual decisions of local and global descriptors are exploited at query time.

Local descriptors represent duplicated scenes with geometric deformations.

Global descriptors are more practical for near-duplicate and semantic searches.

2) Is it enough to obtain the highest or fastest accuracy to deploy a complete visual retrieval systems?

A plausible solution must consider hardware limitations before querying to decrease offline step complexity.

Propose Method

An image-based framework is imposed where keyframes are uniformly sampled from a sequence of video. Three main steps are utilized in our model.

- Local Visual Content Representation
- Global Visual Content Representation
- Late Fusion

Local Visual Content Representation

1) Root SIFT and Hessian Laplacian are used for local representation.

2) A feature vector is converted into two interrelated hash codes (original and its residual vector) for a reasonable computation effort as:

$$q_b(f_h) = \min_i \| f_h - c_i \|_2, c_i \in C_{bow},$$

$$q^k_{pq}(r) = \min_i \| r_k - c_i \|_2, c_i \in C^k_{pq}, \forall k.$$

3) A two-fold approach is used for the voting scheme:

Hash codes must be the same, and residual similarities must be in an error tolerance

Matches must obey the geometric model between the query and reference.

$$w_{pq}(h^r_{pq}, h^q_{pq}) = \frac{1}{m}\sum_{k=1}^{m}\left(1 - \frac{1}{d_k} \| q^{r,k}_{pq} - q^{q,k}_{pq} \|_2\right) \qquad \begin{pmatrix} x^q \\ y^q \\ 1 \end{pmatrix} = \begin{bmatrix} \tilde{s}\cos\tilde{\theta} & -\tilde{s}\sin\tilde{\theta} & t_x \\ \tilde{s}\sin\tilde{\theta} & \tilde{s}\cos\tilde{\theta} & t_y \\ 0 & 0 & 1 \end{bmatrix} \begin{pmatrix} x^r \\ y^r \\ 1 \end{pmatrix}$$

Local Descriptor			
Keypoint	Descriptor	Indexing	Total
0.223	0.410	1.331	1.996
Global Descriptor			
Descriptor	PCA	Fisher Kernel	Total
1.163	0.005	0.193	1.187

Table 1. Approximate time spent on representation computation per frame on a single CPU core.

Global Visual Content Representation

1) Densely sampled pre-trained deep convolutional features are obtained from Alexnet-conv3 layer.

2) Densely sampled features are mapped to a 64-dimensional space by PCA for two reasons:

Degrading the sparsity of features.

Providing time advantage in computations.

3) Deep features are aggregated with first-order Fisher Kernel and converted into binary representations.

4) Standard brute-force binary search is replaced with an approximate nearest neighboring in Hamming space.

$$w_b(b^r, b^q) = \begin{cases} g_h(b^r, b^q), & \text{if } b^r \text{ is in KNN of } b^q \\ 0, & \text{otherwise} \end{cases}$$

Late Fusion

1) The idea is to search two databases for local and global representations by depicting the same visual content. The similar scenes are retrieved from these databases.

2) A settling point is determined from each list to normalize these scores. First-order score derivatives are computed between all two consecutive confidence scores, and the gradient converges to a minimal number after a period.

3) Normalized local and global scores are merged.

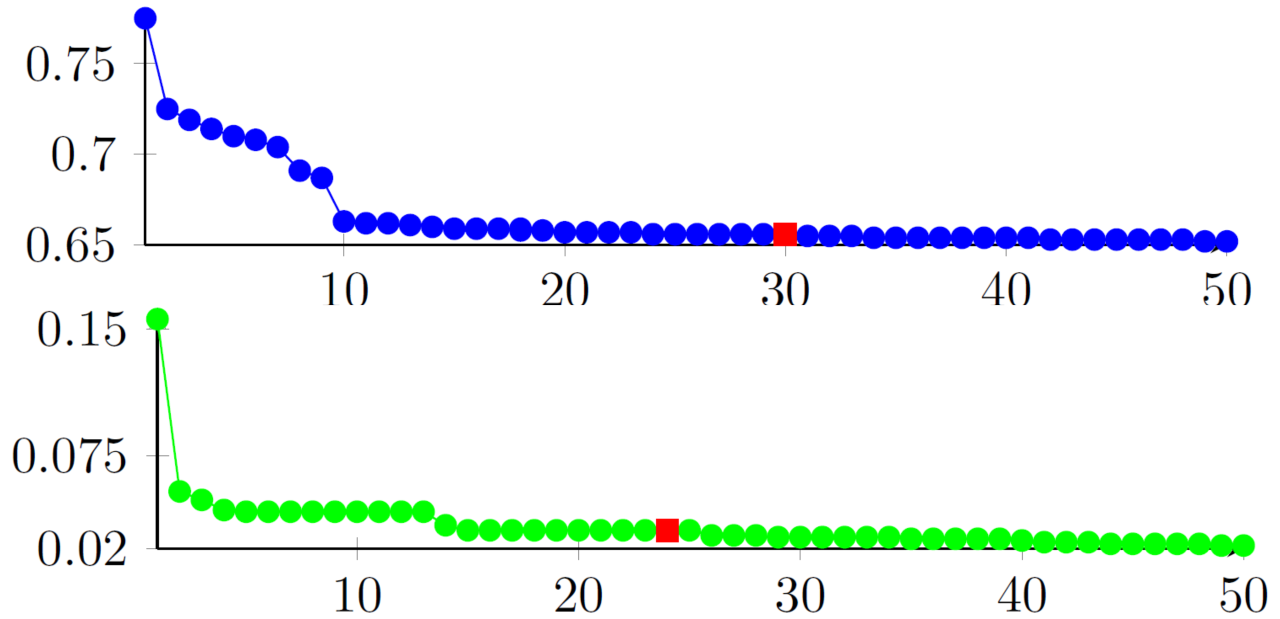


Figure 1. Top confidence scores for two ranked lists.

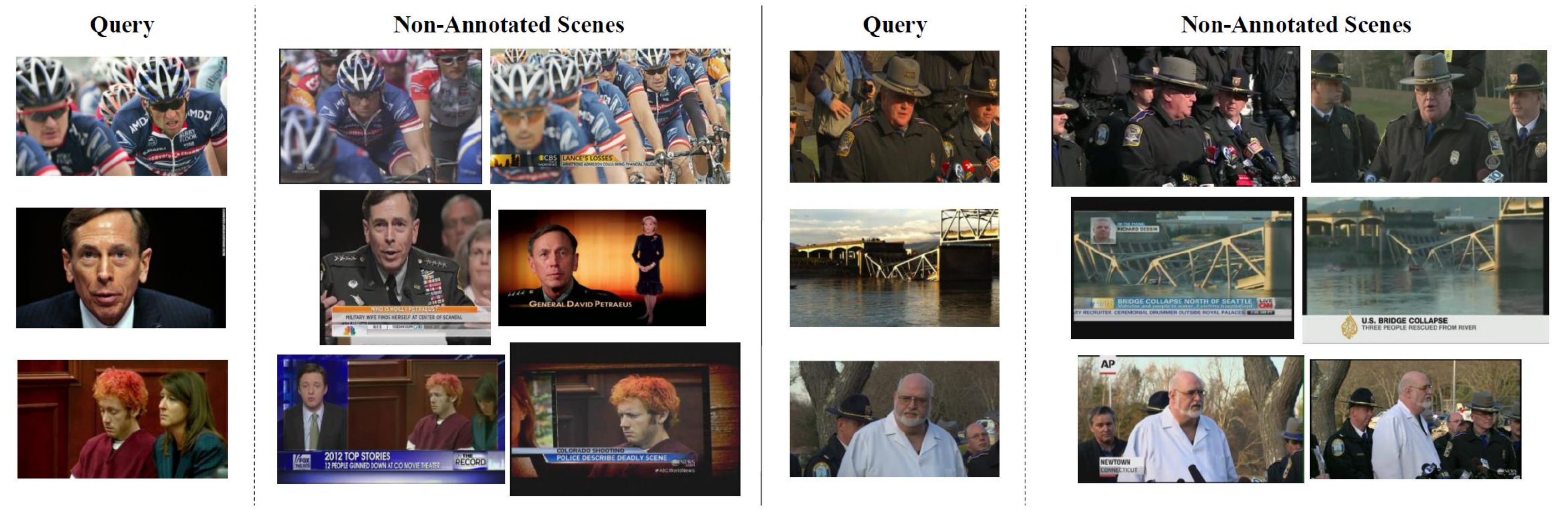


Figure 2. Non-annotated scene samples are unveiled by our retrieval results on Stanford I2V dataset

Experiments

The experiments are conducted on Stanford I2V. The full and light versions consist of 3801 and 1035 hours of videos.

Comparison with baseline

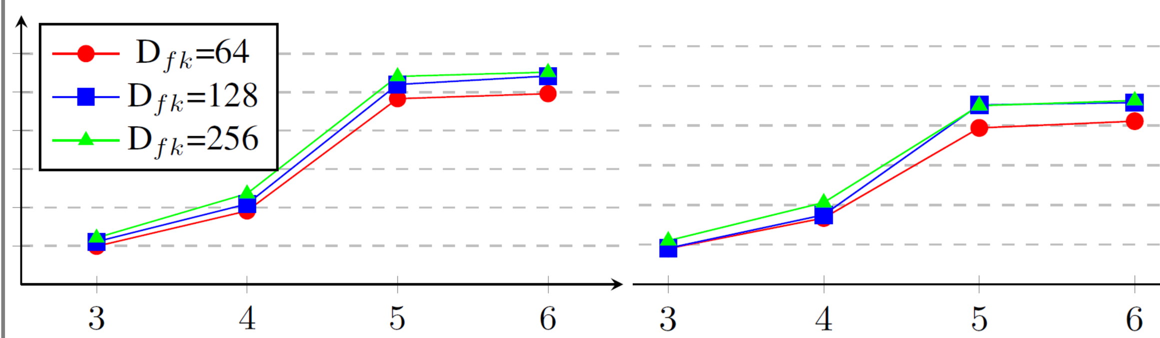
$[D_{bow}\text{-}D_{fk}]$	Light Dataset mAP	Light Dataset mAP@1	Full Dataset mAP	Full Dataset mAP@1	Latency Per 1000h
EH [24]	-	-	0.15	0.37	-
PHOG [24]	-	-	0.22	0.45	-
SCFV [19]	0.46	0.73	0.43	0.64	12.75 sec
BF-PI [21]	≈0.68	-	≈0.65	-	≈ 4.3 sec
RMAC [25]	-	-	≈0.66	-	-
ours[5K - 64]	0.667	0.769	0.582	0.716	17.11 sec
ours[5K - 128]	0.695	**0.794**	0.601	0.755	18.237 sec
ours[5K - 256]	**0.707**	0.782	0.622	0.755	19.253 sec
ours[10K - 64]	0.668	0.769	0.644	0.764	8.675 sec
ours[10K - 128]	0.679	0.782	0.663	**0.786**	9.802 sec
ours[10K - 256]	0.700	0.782	**0.670**	0.777	10.809 sec

Updating Ground Truth Annotations

$[D_{bow}\text{-}D_{fk}]$	Light Dataset mAP	Light Dataset mAP@1	Full Dataset mAP	Full Dataset mAP@1
[5K - 64]	0.697	0.794	0.577	0.720
[5K - 128]	0.735	0.833	0.607	0.755
[5K - 256]	**0.755**	**0.846**	0.624	0.764
[10K - 64]	0.708	0.807	0.648	0.768
[10K - 128]	0.729	0.820	0.667	0.786
[10K - 256]	**0.755**	0.833	**0.681**	**0.790**
EH [24]	-	-	0.19	0.42
SCFV [19]	0.48	0.76	0.44	0.68

The ground truth annotations are updated for SI2V dataset. The annotation list is unveiled with our retrieval results, and it is accessible https://github.com/savasozkan/i2v.

Impact of k values for binary NN voting



Impact of the local threshold for top 100 retrieved scenes