Exploiting Redundancy in Binary Features for Image Retrieval in Large-Scale Video Collections

Aston University
Birmingham

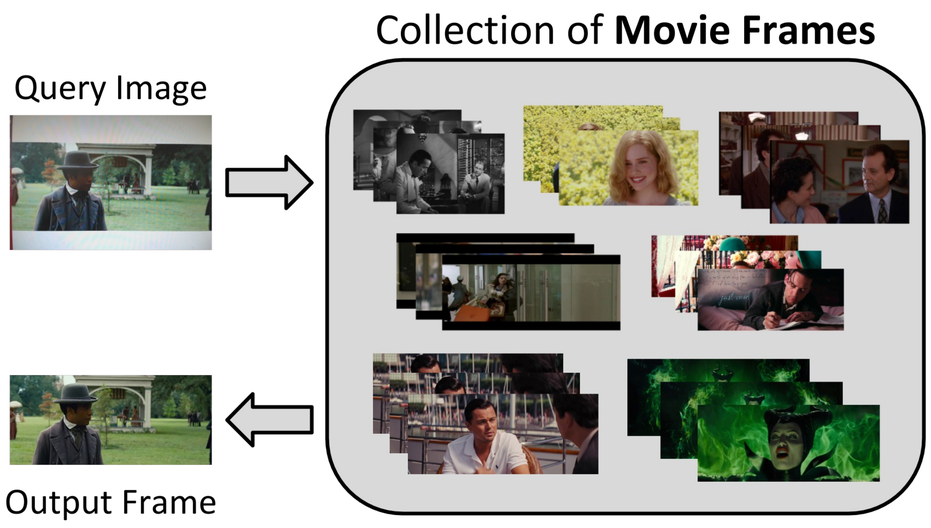
Noa Garcia, George Vogiatzis
{garciadn, g.vogiatzis}@aston.ac.uk

Introduction

Context
Retrieval of video frames from photographs taken during video playback.

Problem
The number of frames in video collections scales very fast. A standard movie (2 hours duration) may contain around 200k frames.

Objective
To reduce the amount of data to be processed in scene retrieval by exploiting the temporal redundancy between frames.

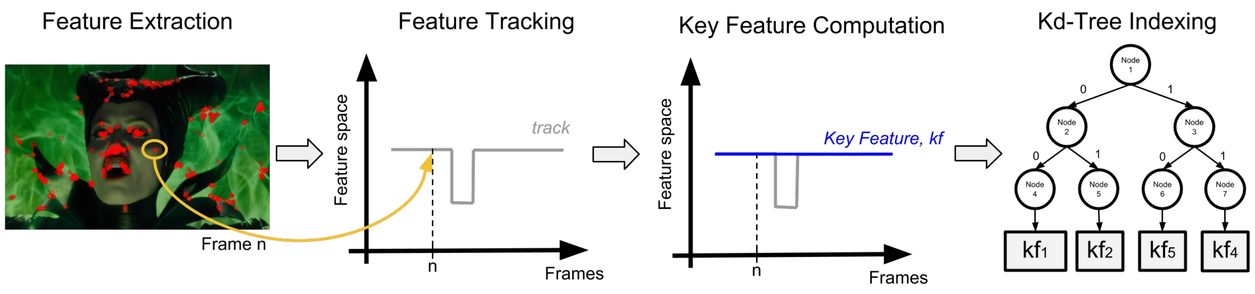


As a result, videos would be more interactive and possibly augmented with extra content that can enhance viewer experience, such as director's commentary, educational information or purchasable items present in the scene.

Methodology

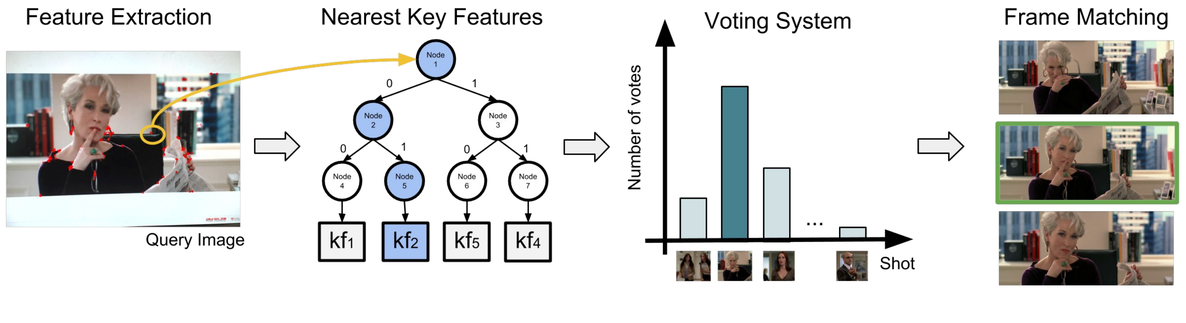
Training phase

1. Extract BRIEF **binary** features [1] from every frame.
2. **Track** features along time during the video.
3. Split frames into **shots** when consecutive frames do not share any track.
4. Aggregate features in the same track into a **key feature** vector by calculating the majority bit for each position.
5. Index key features in a **kd-tree**.



Query phase

1. Extract BRIEF **binary** features [1] from query image.
2. Find the **nearest key features** by using the kd-tree.
3. Key features vote for the **shot** they belong to.
4. Frames contained in the most voted shot are compared against input image by applying **Hamming distance** between their features.



Experiments

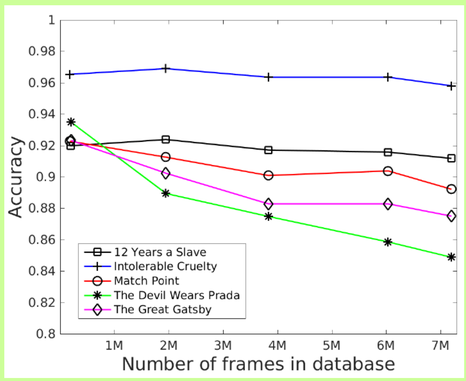
Goal: To compare our work against other systems.
Description: Three systems are implemented: a brute force (BF), the system described in [2] (KT) and the key frame extractor from [3] (KF).
Data: 1 movie (200k frames), 611 query images.
Results:

	BF	KT	KF	Ours
Memory	2.53GB	2.53GB	762MB	**61MB**
Accuracy	**0.98**	0.96	0.93	0.94

Although the performance is similar, the memory is drastically reduced with our method.

Goal: To explore the scalability of our approach.
Description: The collection is incresead up to 40 movies, 7 million frames and 80 hours of video.
Data: 40 movies (7M frames), 25k query images.
Results: The total accuracy over the 40 videos is 0.87, reaching values of 0.98 and 0.97 in some films.

This figure shows how the performance is affected by scaling. In the worst case scenario the loss in accuracy is less than a 8.5%.



Conclusions

By summarizing the content of similar frames in a feature structure known as **key feature**, the redundancy in consecutive frames is exploited, resulting in a significant reduction of the amount of data to be processed for scene retrieval. This system would make videos more interactive and possibly augment them with extra content that can enhance viewer experience.

References

[1] M. Calonder, V. Lepetit, C. Strecha, and P. Fua. BRIEF : Binary Robust Independent Elementary Features. Proc ECCV, 778–792, 2010.
[2] M. Aly, M. Munich, and P. Perona. Distributed kd-trees for retrieval from very large image collections. Proc BMVC, pages 1–11, 2011.
[3] Z. Sun, K. Jia, and H. Chen. Video key frame extraction based on spatial-temporal color distribution. In IIHMSP, pages 196–199, 2008.