Scene segmentation using Temporal Clustering for Accessing and Re-using Broadcast Video

Lorenzo Baraldi, Costantino Grana, Rita Cucchiara

Imagelab, Department of Engineering, University of Modena and Reggio Emilia

Overview

Goal: segment a broadcast video into coherent and story-telling parts (scenes).

- Our model is based on a novel combination of **local image descriptors** and **temporal clustering** techniques.
- The problem of evaluating scene detection results is also addressed, with an **improved performance measure** that reduces the gap between numerical evaluation and expected results.

Contact Information

- Web: http://imagelab.ing.unimore.it
- Email: name.surname@unimore.it

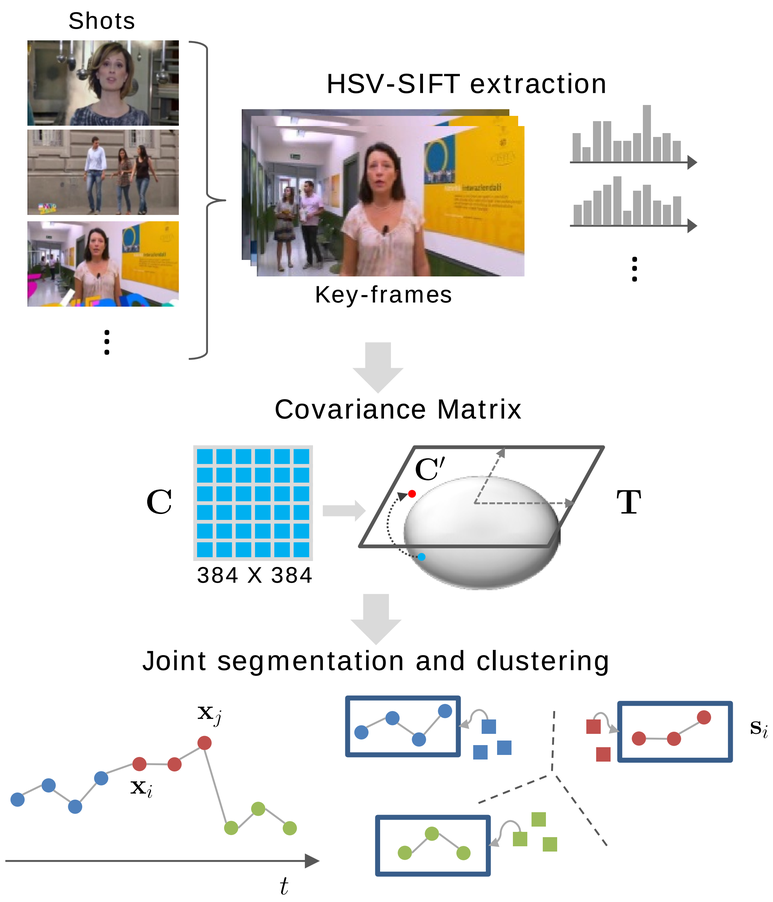

Figure 1: Summary of our approach.

Introduction

The large availability of videos has led to a strong interest in the re-use of video content coming from major broadcasting networks. Unfortunately, re-using videos in ones own presentations or video aided lectures is not an easy task, and requires video editing skills and tools.

Scene detection has been recognized as a tool which effectively may help in this situation, going beyond frames and even beyond simple editing units, such as shots. The task is to identify coherent sequences (scenes) in videos, without any help from the editor or publisher.

- Broadcast videos are sequences of shots, taken without interruption by a single camera.
- Since each shot conveys a uniform content, scene detection can be seen as a temporal clustering problem at the shot level.

Representing shots

- **Keyframes selection**: frames from each shot are described with color histograms and then clustered with Spectral Clustering. Medoids of each cluster are selected as keyframes.
- **Local features extraction**: To encode visual shot similarity, dense HSV-SIFT descriptors are extracted from each keyframe using the Harris-Laplace detector.
- **Feature encoding**: We summarize HSV-SIFT descriptors in a fixed-size representation by computing their covariance matrix $\mathbf{C}$.
  Since covariances belong to the Riemannian manifold of symmetric positive semi-definite matrices, Euclidean operations cannot be computed among them. Therefore, we project $\mathbf{C}$ to an Euclidean tangent space, via:

$$\mathbf{C}' = \log_{\mathbf{T}}(\mathbf{C}) = \mathbf{T}^{\frac{1}{2}} \log\left(\mathbf{T}^{-\frac{1}{2}}\mathbf{C}\mathbf{T}^{-\frac{1}{2}}\right)\mathbf{T}^{\frac{1}{2}}$$

the final descriptor for a shot, $\psi(\mathbf{x}_i)$, is the upper triangular part of $\mathbf{C}'$.

Figure 2: Two consecutives scenes from the RAI dataset.

Temporal segmentation and clustering

Given a sequence of shots $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n]$, we decompose it into a set $\mathbf{s}$ of m scenes (m not known a priori) and jointly assign them to one of k clusters. We minimize the following function [1]:

$$J(\mathbf{G}, \mathbf{s}) = \sum_{c=1}^{k}\sum_{i=1}^{m} g_{ci}\|\phi(\mathbf{s}_i) - \mathbf{m}_c\|^2 \qquad (1)$$

where $\mathbf{s}_i$ is a scene that begins at shot $\mathbf{x}_{s_i}$ and ends at shot $\mathbf{x}_{s_{i+1}}$.

Distances $\|\phi(\mathbf{s}_i) - \mathbf{m}_c\|^2$ between scenes and cluster centers can be implicitly computed without knowing $\mathbf{m}_c$:

$$\|\phi(\mathbf{s}_i) - \mathbf{m}_c\|^2 = \phi(\mathbf{s}_i)^T\phi(\mathbf{s}_i) - \frac{2}{n_c}\sum_{j=1}^{m} g_{cj}\phi(\mathbf{s}_i)^T\phi(\mathbf{s}_j) + \frac{1}{n_c^2}\sum_{j_1,j_2=1}^{m} g_{cj_1}g_{cj_2}\phi(\mathbf{s}_{j_1})^T\phi(\mathbf{s}_{j_2}) \qquad (2)$$

The mapping $\phi(\cdot)$ determines the similarity between two scenes. We define it as the average similarity between the shots belonging to the two scenes:

$$\phi(\mathbf{s}_i)^T\phi(\mathbf{s}_j) = \frac{\sum_{h=s_i}^{s_{i+1}}\sum_{k=s_j}^{s_{j+1}}\kappa_{hk}}{(s_{i+1} - s_i + 1)(s_{j+1} - s_j + 1)} \qquad (3)$$

where κ_{ij} is the similarity of shots $\mathbf{x}_i$ and $\mathbf{x}_j$, that we compute applying a Gaussian kernel to feature vectors $\psi(\mathbf{x}_i)$ and $\psi(\mathbf{x}_j)$:

$$\kappa_{ij} = \exp\left(-\frac{\|\psi(\mathbf{x}_i) - \psi(\mathbf{x}_j)\|^2}{2\sigma^2}\right). \qquad (4)$$

The number of clusters, k, is selected using the maximum eigen-gap criterion on the Normalized Laplacian of matrix κ_{ij}.

Method	F_{CO}	F^*_{CO}
Chasanis *et al.* [2]	0.44	0.44
Sidiropoulos *et al.* [3]	0.54	0.43
Our method	**0.58**	**0.58**

Table 1: Comparison with the state of the art on the RAI dataset

Evaluation

We adopt the Coverage and Overflow schema [4]. For each ground truth scene $\tilde{\mathbf{s}}_i$:

$$\mathcal{C}_t = \frac{\max_{i=1...,m} \#(\mathbf{s}_i \cap \tilde{\mathbf{s}}_t)}{\#(\tilde{\mathbf{s}}_t)}$$

$$\mathcal{O}_t = \frac{\sum_{i=1}^{m} \#(\mathbf{s}_i \setminus \tilde{\mathbf{s}}_t) \cdot \min(1, \#(\mathbf{s}_i \cap \tilde{\mathbf{s}}_t))}{\#(\tilde{\mathbf{s}}_{t-1}) + \#(\tilde{\mathbf{s}}_{t+1})}$$

where $\mathbf{s}_i$ is a detected scenes, and scenes are represented as sets of shot indexes.

We identify two inconveniences and propose an improved definition ($\mathcal{C}^*$, $\mathcal{O}^*$):

- An error on a short shot is given the same importance of an error on a very long shot. Compute $\mathcal{C}$ and $\mathcal{O}$ at frame level.
- The amount of error due to overflowing should be related to the current scene length, instead of its two neighbors. Normalize $\mathcal{O}_t$ with respect to the length of $\tilde{\mathbf{s}}_t$.

We evaluate our approach on a collection of ten randomly selected broadcasting videos from the Rai Scuola video archive. Our dataset and the corresponding annotations are available for download.

References

[1] F. Zhou, F. De la Torre, and J. K. Hodgins,
"Hierarchical aligned cluster analysis for temporal clustering of human motion,"
IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 3, pp. 582–596, 2013.

[2] V. T. Chasanis, C. Likas, and N. P. Galatsanos,
"Scene detection in videos using shot clustering and sequence alignment,"
IEEE Trans. Multimedia, vol. 11, no. 1, pp. 89–100, 2009.

[3] P. Sidiropoulos, V. Mezaris, I. Kompatsiaris, H. Meinedo, M. Bugalho, and I. Trancoso,
"Temporal video segmentation to scenes using high-level audiovisual features,"
IEEE Trans. Circuits Syst. Video Technol., vol. 21, no. 8, pp. 1163–1177, 2011.

[4] J. Vendrig and M. Worring,
"Systematic evaluation of logical story unit segmentation,"
IEEE Trans. Multimedia, vol. 4, no. 4, pp. 492–499, 2002.