Scene-driven Retrieval in Edited Videos using Aesthetic and Semantic Deep Features

Lorenzo Baraldi, Costantino Grana, Rita Cucchiara

Imagelab - University of Modena and Reggio Emilia, Italy
Outline

• Introduction
• Scene-driven retrieval with thumbnail selection
  • Scene detection
  • Semantic concept detection
  • Aesthetic ranking
• Experimental results
Introduction

Novelties of our work

- Focus on collections of long broadcast videos
- Query-dependent, semantically and aesthetically valuable thumbnails
- Search with finer granularity level
- No manually provided annotations (description, tags)

Variety of topics: all should be searchable
Related works

Thumbnails are *surrogates* for videos [Craggs et al., 2014]

- Thumbnails create an *intention gap* if not relevant

- C. Liu et al. (ICIP 2011) reinforcement + relevance model to compute query/thumbnail similarity

- W. Liu et al. (CVPR 2015) deeply-learned latent space to compute similarity between query and thumbnail
Outline

• Introduction

• **Scene-driven retrieval with thumbnail selection**
  • Scene detection
  • Semantic concept detection
  • Aesthetic ranking

• Experimental results
Overview

- Broadcast videos can be decomposed at three different levels:
  - **Frames**
  - **Shots**
    - Taken by a single camera
  - **Scenes**
    - Uniform semantic content
Overview

- Scene-based retrieval

- Given a query, return a ranked list of:
  
  \[(\text{video, scene, thumbnail})\]
  
  Belongs to retrieved video
  Should be relevant for query

  \[(\text{video, scene, thumbnail})\]
  
  Belongs to retrieved scene
  Should be relevant for query and aesthetically remarkable

  ...
Scene detection

Group adjacent shots according to semantic coherence

Can be identified with visual features only

Can not be identified with visual features

Need of multi-modal features!
Scene detection

Shot clustering according to learned similarity metric.

A Siamese DN is trained to predict whether two shots should belong to the same scene.

At test phase: spectral clustering

[Baraldi, Grana, Cucchiara; ACM MM 2015]
Scene detection

Training on ILSVRC12 (1.2 million images, 978 object classes)
Fine-tuning with Places Dataset (2.5 million images, 205 scene categories)
Fine-tuning with BBC Planet Earth (12 million shot pairs)

[Baraldi, Grana, Cucchiara; ACM MM 2015]
Outline

• Introduction
• Scene-driven retrieval with thumbnail selection
  • Scene detection
  • Semantic concept detection
  • Aesthetic ranking
• Experimental results
Semantic Concept Detection

Hypothesis:

- In broadcast videos (documentaries, news, educational footage, ...) speaker describes what the video shows

Video transcript suggests semantic concepts

- No ontologies
- No pre-defined concept list
- No user queries

“Semantic Concept Detection” is the task of extracting concepts defined in the transcript in frames → shots → scenes
Every year three million caribou migrate across the arctic tundra. The immensity of the herd can only be properly appreciated from the air.

Video transcript

Probabilistic SVM classifier on CNN features

$P(s, u)$ Probability that shot $s$ contains “caribou”

Most similar category in semantic space (caribou)

40,000 categories (1000 images per class on avg)
Imagine our world without sun. Male emperor penguins are facing the nearest that exists on Plane earth, winter in Antarctica.

Video transcript

Most similar category
In semantic space
(polar, glacier)

40,000 categories
(1,000 images per class on avg)

Probability that shot s contains “polar/glacier”
Semantic Concept Detection

Visual concepts
plain · steppe · grassland · reach · forest

Visual concepts
cloud · plateau · mountain · side · sun

Visual concepts
animal · elephant · water · grass · hole

Visual concepts
glass · mammal · hog · pig · work
Outline

• Introduction
• Scene-driven retrieval with thumbnail selection
  • Scene detection
  • Semantic concept detection
  • Aesthetic ranking
• Experimental results
Aesthetic Ranking

[Simonyan, Zisserman; 2015]

Input image

conv1_1: 64 filters
conv1_2: 64 filters
conv2_1: 128 filters
conv2_2: 128 filters
conv3_1: 256 filters
conv3_2: 256 filters
conv3_3: 256 filters
conv4_1: 512 filters
conv4_2: 512 filters
conv4_3: 512 filters
conv5_1: 512 filters
conv5_2: 512 filters
conv5_3: 512 filters

Edges
Textures
Part of objects
Objects
Aesthetic ranking

Learn a (linear) ranking model for aesthetics, with a small training set

- Pre-trained CNN
- Hypercolumn features [Hariharan et. al, 2015]
  - From ~4000 activation maps to 5!
  - Disregard information on classes, focus on level (edges vs patterns vs objects) and position

Input image

\[ \mathbf{C}_1 \quad 128 \text{ filters} \quad \mathbf{C}_2 \quad 256 \text{ filters} \quad \mathbf{C}_3 \quad 768 \text{ filters} \quad \mathbf{C}_4 \quad 1536 \text{ filters} \quad \mathbf{C}_5 \quad 1536 \text{ filters} \]
Aesthetic ranking

**Ranking model: Linear SVM Rank**
- For each scene, dataset consists of thumbnail pairs:
  - \((d_i, d_j)\) where \(d_i\) is ranked higher than \(d_j\)
- Equivalent to a linear SVM on pairwise difference vectors

Mathematically:

\[
\begin{align*}
\text{minimize} & \quad \frac{1}{2} \| w \|^2 + C \sum_{i,j,k} \epsilon_{i,j,k} \\
\text{subject to} & \quad \forall (d_i, d_j) \in r_1^*: w\phi(d_i) \geq w\phi(d_j) + 1 - \epsilon_{i,j,1} \\
& \quad \cdots \\
& \quad \forall (d_i, d_j) \in r_n^*: w\phi(d_i) \geq w\phi(d_j) + 1 - \epsilon_{i,j,n} \\
& \quad \forall i, j, k : \epsilon_{i,j,k} \geq 0
\end{align*}
\]
Retrieval

Query

Match with most similar concept in semantic space

$R_{scene}(q) = \max_{s \in scene} \left( \alpha P(s, u) + (1 - \alpha) \max_{d \in s} w\phi(d) \right)$

Shots of the scene

Aesthetic ranking of shot $s$

Probability that concept $u$ appears in shot $s$

As thumbnail, select the one that maximizes $w\phi(d)$
Outline

• Introduction
• Scene-driven retrieval with thumbnail selection
  • Scene detection
  • Semantic concept detection
  • Aesthetic ranking
• Experimental results
Evaluation

BBC Planet Earth dataset

- 11 videos, approximately 50 minutes each
- 4900 shots and 670 scenes
- 3802 unigrams

Annotation

Three (paid) users asked to mark whether each key-frame was aesthetically relevant for its scene:

- Training pairs build according to the number of times a key-frame was selected
Comparison with baseline: **Ranking CNN**
- Pre-trained VGG-16
- End-to-end learning with MSE

3.73% error reduction
Thumbnail selection

Ranking of a sample scene

 Thumbnails with good quality and clearly visible object in the middle are preferred
Thumbnail selection

Ranking of a sample scene

Thumbnails with good quality and clearly visible object in the middle are preferred
Thumbnail selection

Ranking of a sample scene

Failure case:
Retrieval results (query “penguin”)

Shallow Seas
Probability: 0.78%

Being flightless, the returning penguins must cross the open beach on foot. Fur seals, that have come to the beach to breed are waiting for them... Fur seals normally live on krill, but these have now acquired an unexpected taste for blubber-rich penguins....

Shallow Seas
Probability: 0.75%

Penguins may be the featherweights by comparison, but they have razor sharp bills and a feisty character. The seal could easily lose an eye. The only safe way to grab a penguin is from behind... and the birds are well aware of that. Both animals are clumsy on this terrain. But the penguin has...

From Pole to Pole
Probability: 0.72%

The penguins stay when all other creatures have fled because each guards a treasure a single egg resting on the top of its feet and kept warm beneath the downy bulge of its stomach. There is no food and no water for them and they will not see the sun again for four months. Surely, no greater ord...
Retrieval results (query “conifer”)

Seasonal Forests
Probability: 0.78%
The Pacific coast of North America. The land of hemlock, Douglas fir and giant redwood. Here, water is never locked-up in ice. And even if rains fail, the needles can extract moisture from the fogs that roll in from the sea. The sun’s energy powers these forests not for one month, as it does...

Seasonal Forests
Probability: 0.70%
The American conifer forests may not be the richest in animal life, but their trees are extraordinary. This giant sequoia, a relative of the redwood, is the largest living thing on earth. Known as General Sherman, it's the weight of ten blue whales. Higher up in the nearby mountains, bristle-co...

Seasonal Forests
Probability: 0.64%
Trees. Surely among the most magnificent of all living things. Some are the largest organisms on earth, dwarfing all others and these are the tallest of them all. The deciduous and coniferous woodlands that grow in the seasonal parts of our planet are the most extensive forests on earth. Their shee...
Retrieval results (query “fall”)

Fresh Water
Probability: 0.62%
In their upper reaches mountain streams are full of energy. Streams join to form rivers, building in power, creating rapids...

Fresh Water
Probability: 0.58%

Fresh Water
Probability: 0.53%

Moisture, rising as water vapour from the surface of the sea, is blown inland by wind. On reaching mountains, the moist air is forced upwards and, as it cools, it condenses into cloud and finally rain, the source of all Fresh Water. There is a tropical downpour here almost every day of the year...
User study

- 12 undergraduate students
- 20 queries each
- Comparison with respect to the results of a full-text search inside text:
  - With random thumbnails
  - With thumbnail selection
Conclusions

• We proposed a video retrieval pipeline
  • Specifically designed for broadcast videos
  • Relies on temporal video segmentation (scenes)
  • Retrieval is carried out with semantic concept detection and aesthetic rankings.
  • Quantitative and qualitative evaluation
Thank you
Any questions?

lorenzo.baraldi@unimore.it, costantino.grana@unimore.it, rita.cucchiara@unimore.it

http://imagelab.ing.unimore.it

Thanks to "Educating City" Project funded by Italian MIUR National Technology Cluster program with UNIMORE, CNR, UNITN, RAI, ATI_Clttàedu, RCI, Almaviva