Panagiotis Sidiropoulos

Shot Descriptors for Video Temporal Decomposition

CERTH-ITI
Video Segments

- **Shot**
  - Video segment taken without interruption by a single camera

- **Scene**
  - Logical Story Unit (LSU): a series of temporally contiguous shots characterized by overlapping links that connect shots with similar content
  - A division of an act presenting continuous action in one place

- **Story**
  - Only in news broadcasts
Video Temporal Decomposition (1)

- Partition of video sequence $V$ into convex sub-sets

\[ \bigcup V_i = V \]
\[ V_i \cap V_j = \emptyset, \ \forall i \neq j \]

\[ \forall V_i \text{ if } x_1, x_2 \in V_i \text{ then all } x, x_1 \leq x \leq x_2 \text{ belong also to } V_i \]
Video Temporal Decomposition (2)

- **Shot segmentation**
  - State-of-the-art F-score level of 95% [1]
  - Eliminated from TRECVID in 2008

- **Scene (story) segmentation**
  - Still open issue

Basic Assumption

- Each shot belongs to exactly one scene
  - Scene boundaries are a subset of shot boundaries
  - Not valid in story segmentation
    - 9% of story boundaries not shot boundaries [1]
  - Shot grouping = Scene segmentation

Scene Segmentation Overview

- Feature Extraction
- Shot Segmentation
- Shot Linking
- Video Segmentation
- Ground truth Estimation
- Evaluation
Scene Segmentation Points

- Shot Descriptor Extraction
- Descriptor use/fusion
- Scene Disambiguation
- Development of evaluation measures
Scene Segmentation Points

- Shot Descriptor Extraction
- Descriptor use/fusion
- Scene Disambiguation
- Development of evaluation measures
Shot index or frame index of a representative frame

Temporal similarity is a function of their temporal distance

- Binary
  - Prune the set of candidate shot links
- Continuous
  - Filter shot content similarity (exponential)
Representative key-frame extraction

Low-level descriptors
  • Hint in relevant literature that descriptor selection does not play critical role
  • HSV or L*u*v histogram
High-level visual descriptors

Visual concept detectors representing key-frame semantic visual content
  - Confidence value (estimated probability the visual concept present)

Confidence-value feature vector

Motion Descriptors

- Based on video spatio-temporal nature
- Pair-wise comparison of frames and extraction of global motion properties
- Spatio – temporal slices (one axis in time, one axis in space)
  - Tensor Histograms for shot motion descriptor
- Require computations in frame level
  - Computational expensive
Low-level Audio Descriptors

- Volume
- Energy
- Zero-crossing Rate
- Mel-frequency cepstral coefficients
- Etc…

- Comparisons between adjacent shots
  - Discontinuity recognition
The distribution of speakers across two shots can measure audio similarity.

Speaker diarization.
- Identifying in an audio stream segments homogeneous according to the speaker identity.
- Assign a speaker ID in each speaker segment.

The histogram of the speakers present in each shot is estimated.

Audio Events

- High-level audio descriptors
- Audio-corresponding to visual concepts
- Confidence-value feature vectors
- Audio events and speaker histogram experimentally tested during Vidi-Video project.
  - Enhance low-level visual results

Other

- ASR Results
- Light source estimation
- Low-level visual descriptors from key-frame areas corresponding to background
- Face recognition or at least face detection
Scene Segmentation Points

- Shot Descriptor Extraction
- **Descriptor use/fusion**
- Scene Disambiguation
- Development of evaluation measures
Uni-descriptor Approaches

- Common approach: Graphs, estimate cuts.
- Scene Transition Graph (STG) [1]
  - Use visual similarity and temporal proximity
  - Two thresholds: one visual, one temporal.
  - Scene Convexity
  - Generalization to all kind of descriptors and modalities

Multi-descriptor Approaches/ Descriptor Fusion

- **Early Fusion**
  - Append descriptors to a “bigger” one
  - Employ a uni-descriptor approach

- **Late Fusion**
  - For each descriptor extract results exclusively based on it
  - Combine results
 INCLUDED TYPES OF STGS

• Low-level visual (HSV)
• High-level visual (visual concepts)
• Speaker Histogram
• High-level audio (audio events)
For each type of STG

- Generalization of multiple STGs for different, randomly selected values of content similarity and temporal proximity parameters
- Approximation of the probability value for each shot boundaries to be also a scene boundary, based on the descriptor of the STG type
- Random walk to parameter space

Probability values linear combination

- Thresholding

Four Parameters to tune

- 3 Weights, 1 Global Threshold
Documentary Base
- 15 Documentaries
- 513 minutes
- 3459 shots
- 525 scenes

Film Base
- 6 movies
- 643 minutes
- 6665 shots
- 357 scenes
Linear Combination: Limited Scalability

- Curse of dimensionality
- Space dimension = Number of descriptors
- As the number of descriptors increase tuning with dense uniform sampling leads to a prohibitively high number of sample points

In General

- Probability late fusion is a function (linear or not) in the descriptor space.
- Metric space for fully exploiting dimensionality reduction field.
- Measure: estimates distance of experimental segmentation from the ground truth segmentation.
Scene Segmentation Points

- Shot Descriptor Extraction
- Descriptor use/fusion
- Scene Disambiguation
- Development of evaluation measures
Temporal Decomposition Measures

- Not common ground for comparison
- Evaluation left to the reader
- Recall-Precision
  - Counting false negatives and false positives.
  - Feasible for shot segmentation since start and end are well defined.
  - Not adequate for scene segmentation (or story segmentation)
    - Do not communicate error magnitude
Method A: Recall 0% Precision 0%
Method B: Recall 33% Precision 25%
Two assumptions

1. The content of a scene is dissimilar from the content of a succeeding scene.
2. Within a scene shots with similar content are repeated.

Overflow measures to what extent assumption 1 is met (optimal value 0%)

Coverage measures to what extent assumption 2 is met (optimal value 100%)

Good modelling of segmentation flaws

- Over-segmentation (Overflow)
- Under-segmentation (Coverage)

Coverage-Overflow Inadequacy

- No obvious way to combine Coverage-Overflow
  - Two algorithms, one performing better in terms of coverage and the other in terms of overflow, which is overall better?

- Coverage-Overflow or their geometrical mean (F-Score) do not define a metric space.

- DED:
  - Single Measure
  - Metric Space
Differential Edit Distance (DED)

- **Idea:**
  - Best system is the one that minimizes the work that is left for human

- **Formally:**
  - The minimum quantity of set elements that need to change sub-set to transform the initial partition into the final.

- **Scene Segmentation:**
  - The minimum number of shots that need to change scene to transform the experimentally estimated partition into the ground truth partition

- **Analogous to Earth Mover’s Distance**

- **Resembles Edit Distance**
Scene Segmentation as shot labeling

- Labels are arbitrate
- Same scene ⇔ Same label
- Different scene ⇔ Different label
Differential Edit Distance (DED)

- Differentially equivalent label strings
  - If two corresponding elements have the same label in the first sequence they will also have the same in the second sequence
  - If two corresponding elements have different labels in the first sequence they will also have different in the second sequence
- Differential edit distance (DED) of label strings
  - the minimum number of label modifications that are required to transform the first string into a string that is differentially equivalent to the second.
Set of labels of first and second string

Occurrence matrix
- 2-d histogram of shots: position \((x,y)\) is the number of shots with label \(x\) in the first string and \(y\) in the second

In the minimum distance\((=\text{DED})\) solution
- If \(m\) and \(n\) the number of elements in label strings then \(\min(m,n)\) labels of the first set are assigned to a label to the second
- The total number of shots related to the assignment labels is maximized

Job Assignment Problem: Hungarian Algorithm [1]

DED Efficiency Optimization

- **Job Assignment Computational Complexity**
  - $O(N^3)$, $N$ is the number of scenes.

- **Computational Optimization Property**:  
  - Two adjacent shots with different labels in both label strings identify a “splitting” point.  
  - Video can be divided into two sub-videos, one ending to the splitting point and one starting from it.
DED Estimation Algorithm

- Find common “label” boundaries
- Split video into sub-videos
- For each sub-video
  - Estimate Ocurrence Matrix
  - Find sub-videos DED
- Sum all DEDs
d(x,y) = 0 iff x = y

• d(x,y) = d(y,x)

• d(x,z) \leq d(x,y) + d(y,z) (transitivity)
  • Suppose d(x,z) > d(x,y) + d(y,z)
    • There are more shot labels that need change between x and z than between x and y and y and z.
    • There are shot labels that change between x and z but do not change between both x and y and y and z.
    • This cannot stand since we can transform string x to z either by first “passing” from y or not, and the results need to be identical.
DED Advantages

- Metric
- Uni-dimensional
- Polynomial Complexity
- Easily implemented
Thanks

Questions?