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

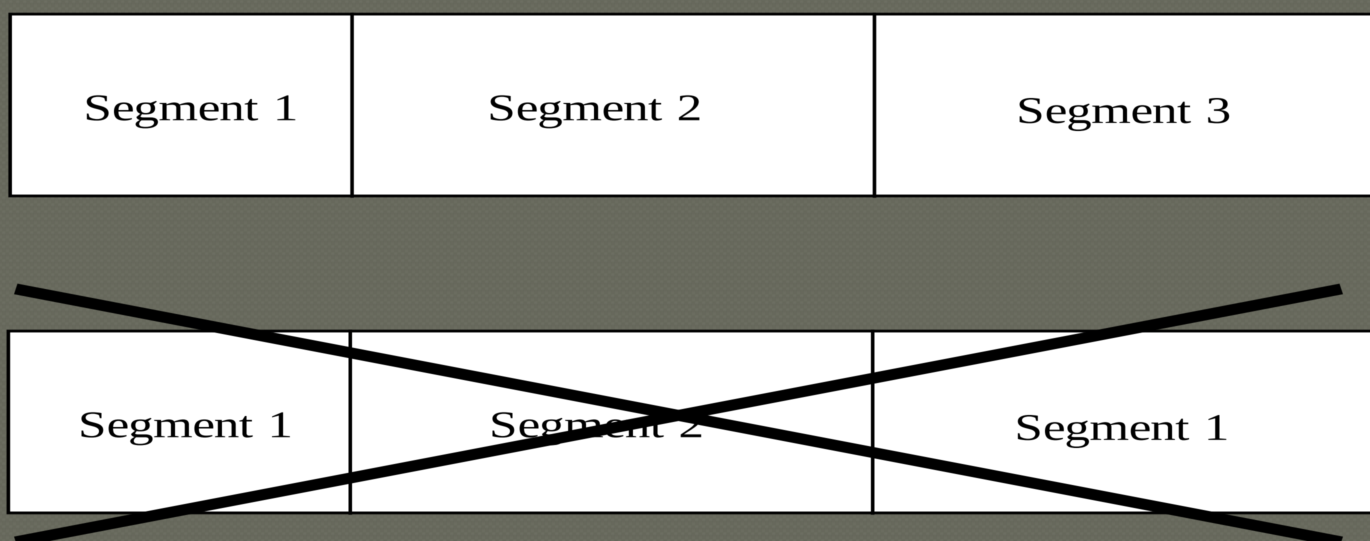
Video Temporal Decomposition (1)

- Partition of video sequence V into convex sub-sets

$$\bigcup V_i = V$$

$$V_i \bigcap V_j = \emptyset, \forall i \neq j$$

$$\forall V_i \text{ if } x_1, x_2 \in V_i \text{ then all } x,\ x_1 \leq x \leq x_2 \text{ belong also to } V_i$$



Video Temporal Decomposition (2)

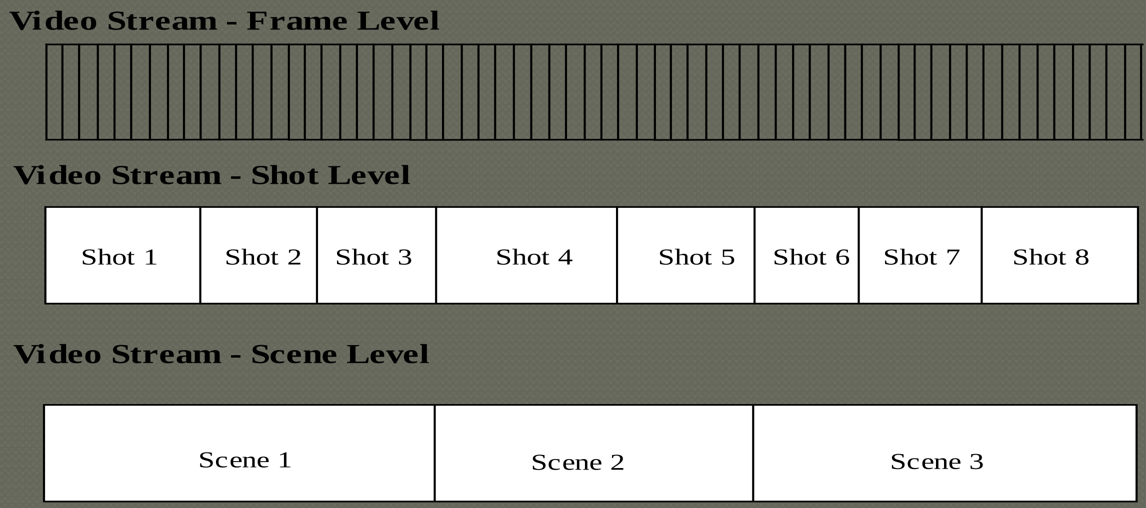
- Shot segmentation
  - State-of-the-art F-score level of 95% [1]
  - Eliminated from TRECVID in 2008
- Scene (story) segmentation
  - Still open issue

[1] Z. Liu, E. Zavesky, D. Gibbon, B. Shahraray, and P. Haffner, "AT&T research at TRECVID 2007," 2007.

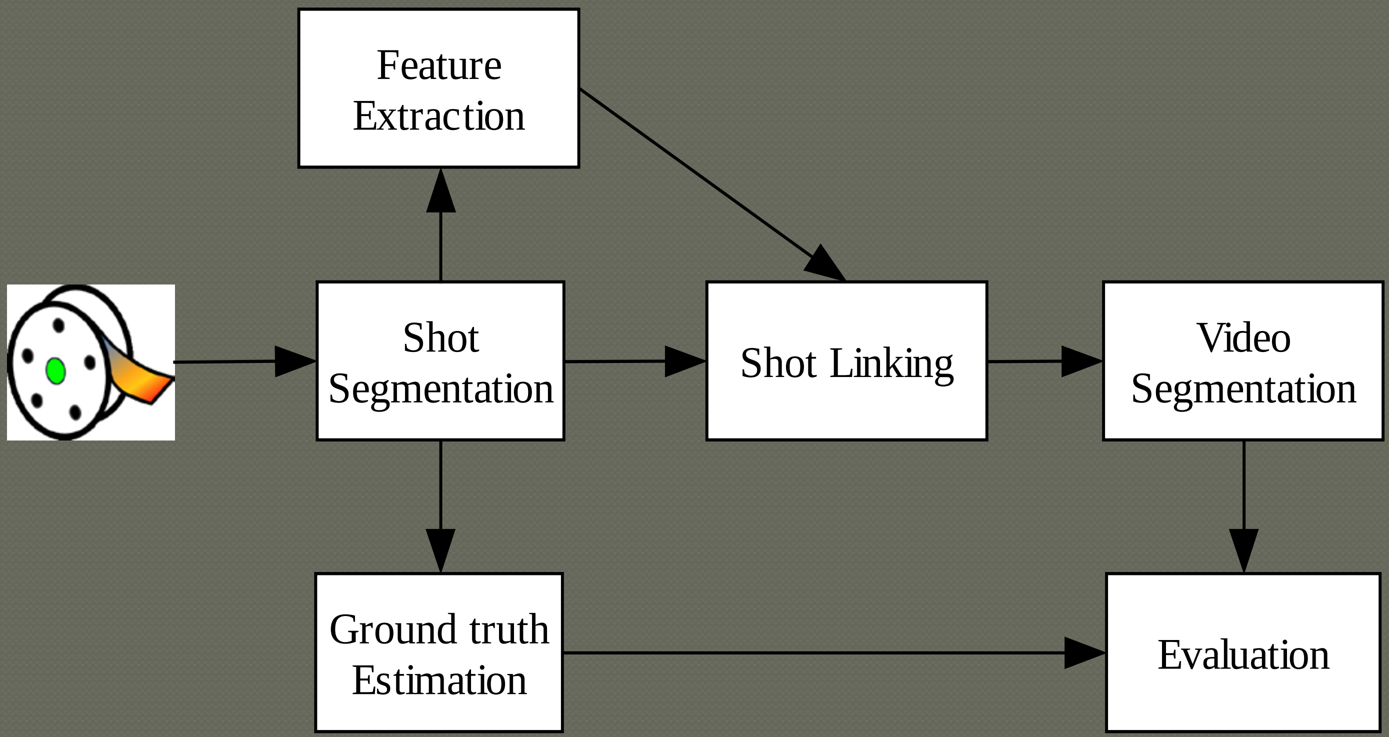
Basic Assumption

- **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

[1] Winston Hsu et al. "Discovery and Fusion of Salient Multi-modal Features towards News Story Segmentation", Proc. of Storage and retrieval methods and applications for multimedia, vol. 5307, 2004, pp. 244–258.



Scene Segmentation Overview



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

Temporal Position

- 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).

Low-level Visual Descriptors

- 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

Visual Concepts

- 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

- [1] V. Mezaris, P. Sidiropoulos, A. Dimou, I. Kompatsiaris, "On the use of visual soft semantics for video temporal decomposition into scenes," IEEE Fourth International Conference on Semantic Computing (ICSC), 2010, pp. 141-148

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

Speaker Histogram

- 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

- [1] P. Sidiropoulos, V. Mezaris, I. Kompatsiaris, H. Meinedo, I. Trancoso, "Multi-modal scene segmentation using scene transition graphs", ACM International Conference on Multimedia (ACM MM), 2009, pp. 665-668.

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

[1] P. Sidiropoulos, V. Mezaris, I. Kompatsiaris, H. Meinedo, I. Trancoso, "On the use of audio events for improving scene segmentation", 11th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS), 2010.

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

- [1] Minerva Yeung, Boon-Lock Yeo, and Bede Liu, Segmentation of video by clustering and graph analysis, Computer Vision and Image Understanding 71 (1998), no. 1, 94–109.

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

STG-based Late Fusion Probabilistic Framework (1)

- Included types of STGs
  - Low-level visual (HSV)
  - High-level visual (visual concepts)
  - Speaker Histogram
  - High-level audio (audio events)

STG-based Late Fusion Probabilistic Framework (2)

- 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 thedescriptor of the STG type
  - Random walk to parameter space
- Probability values linear combination
  - Thresholding
- Four Parameters to tune
  - 3 Weights, 1 Global Threshold

Experimental Results

- Documentary Base
  - 15 Documentaries
  - 513 minutes
  - 3459 shots
  - 525 scenes
- Film Base
  - 6 movies
  - 643 minutes
  - 6665 shots
  - 357 scenes

Method	Coverage(%)	Overflow(%)	F-Score(%)	Coverage(%)	Overflow(%)	F-Score(%)
GSTG ($y \in \{V, VC, A, AE\}$)	**86.30**	**10.91**	**87.67 (87.40)**	**87.91**	17.89	**84.91 (84.64)**
Method of [12]	70.90	24.13	73.30	76.43	**16.15**	79.97
Method of [21]	77.59	17.31	80.06	75.12	24.29	75.41
Method of [24]	78.22	16.73	80.67	79.50	21.17	79.16

Framework Limitations

- 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.

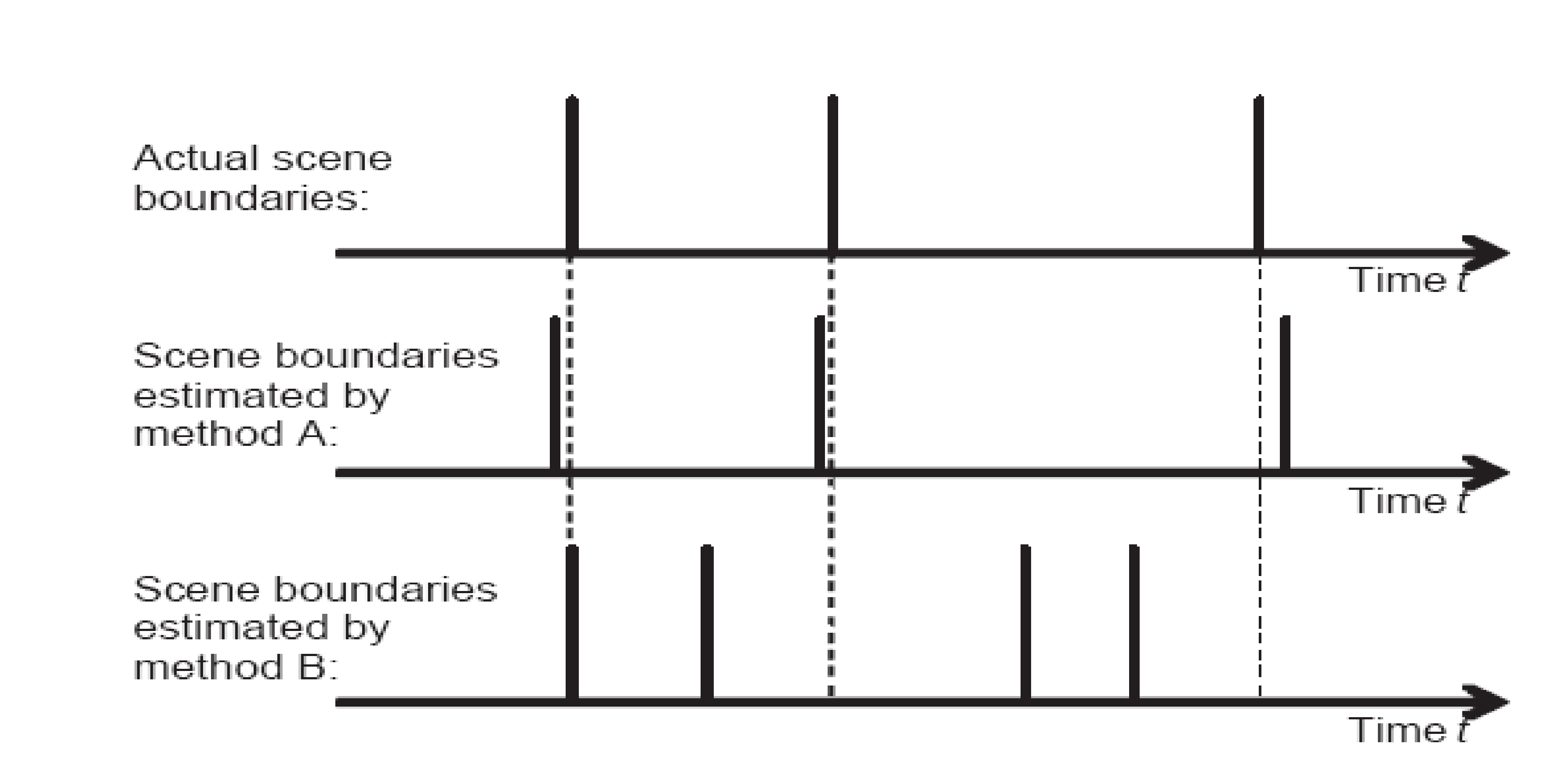
Scene Segmentation Points

- 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

Recall-Precision Inadequacy



- Method A: Recall 0% Precision 0%
- Method B: Recall 33% Precision 25%

Coverage - Overflow

- 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)

[1] Jeroen Vendrig and Marcel Worring, Systematic evaluation of logical story unit segmentation,IEEE Transactions on Multimedia 4 (2002), no. 4, 492–499.

Coverage-Overflow Inadequacy

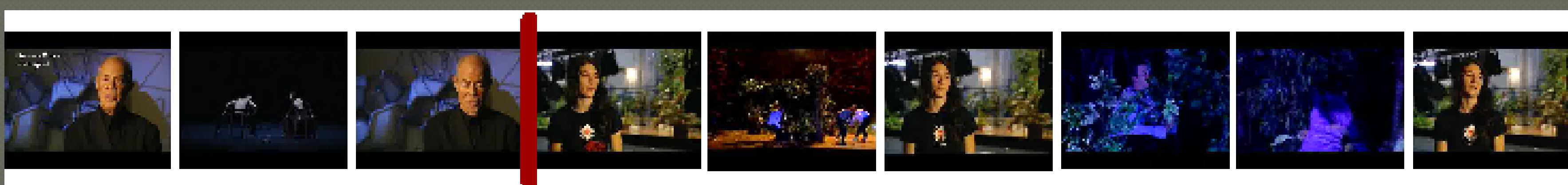
- No obvious way to combine Coverage-Overflow
  - Two algorithms, one performing better in terms of coverage and the other in terms of overflow, which is overall better?
- Coverage-Overflow or their geometrical mean (F-Score) do not define a metric space.
- DED:
  - Single Measure
  - Metric Space

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



1	1	1	2	2	2	2	2	2
2	2	2	1	1	1	1	1	1
*	*	*	+	+	+	+	+	+

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
  - Strings 'AABBCC', '112233', '221133', 'BB11AA', '++–**' are differentially equivalent
- Differential edit distance (DED) of label strings
  - <u>the minimum number of label modifications that are required to transform the first string into a string that is differentially equivalent to the second.</u>

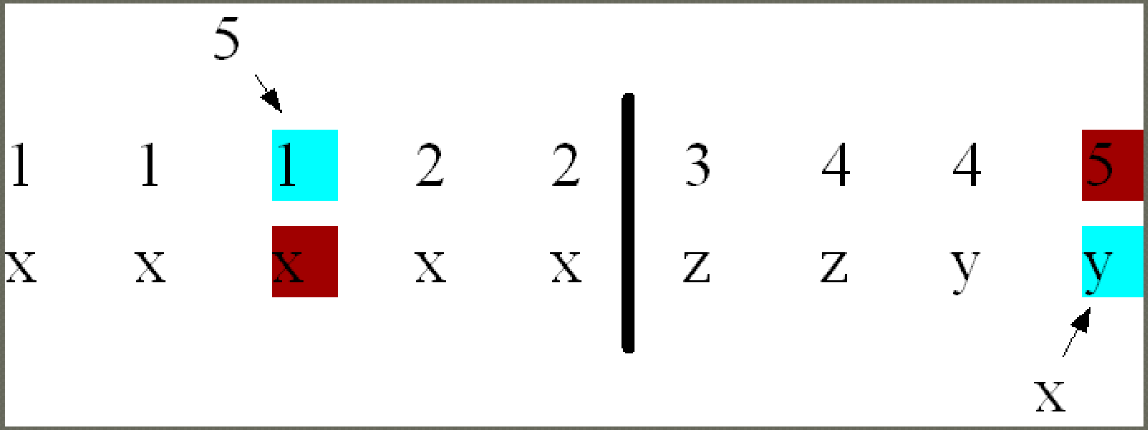
DED Estimation

- Set of labels of first and second string
- Ocurrence 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(=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]

[1] H. W. Kuhn, "The Hungarian Method for the assignment problem", Naval Research Logistics Quarterly, vol. 2, 1955, pp. 83-97.

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

DED Metric Properties

- $d(x,y) = 0$ 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 can not 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?