

## Motivation

summarize rushes  
(raw material, unedited, redundant video)

- scenes are shot from different camera positions
- several alternative takes (mistakes by actors, technical failures, trying different artistic options)
- action performed in takes is similar, but not identical (e.g. omissions, insertions)

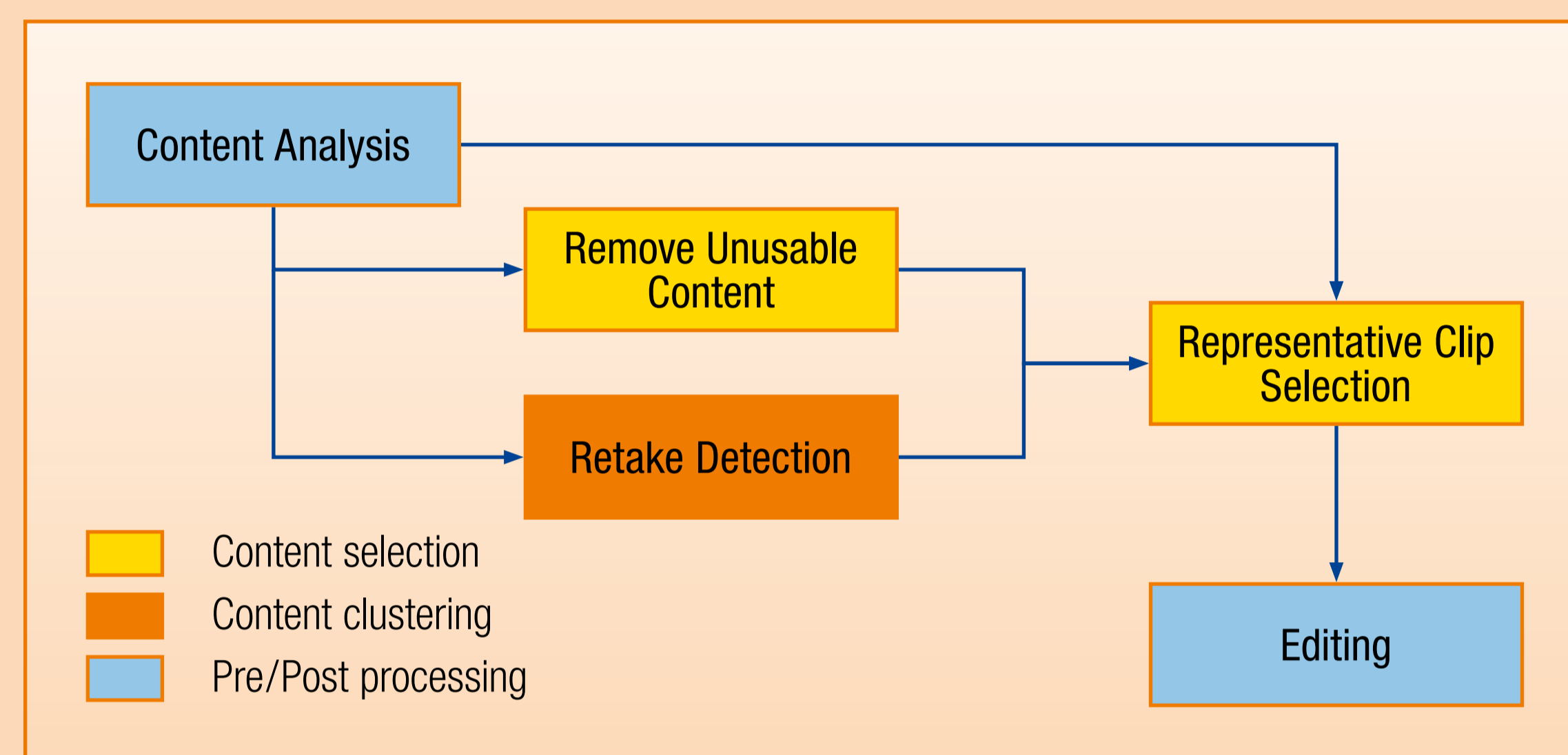


Figure 1: Process of summarizing rushes video.

Figure 2:  
Example of LCSS match between 1d feature  
sequences ( $\epsilon_\gamma = 0.5$ ,  $\theta_{len} = 3$ ,  $\gamma = 1$ ).

	2.3	4.7	1.2	3.3	2.2	1.4
1.3	0	0	1	1	1	1
2.5	1	1	1	1	2	2
3.4	0	1	1	2	2	2
2.4	1	1	1	2	3	3
4.6	1	2	2	2	3	3
1.5	1	2	3	3	3	4
2.6	1	2	3	3	4	4

## Approach

- retake = take of same scene, from same camera
- split videos into parts (shot boundary detection, then split at short-term local maxima of visual activity, e.g. clapboard moving in/out, production staff walking around)
- pair-wise matching of parts
  - MPEG-7 ColorLayout and EdgeHistogram descriptors, visual activity
  - extracted from sequence temporally subsampled by factor 10
- apply modified Longest Common Subsequence (LCSS) algorithm
- clean up matches (remove contained and largely overlapping matches), yields set of (partial) "take candidates"
- result is a similarity matrix of the take candidates
- cluster take candidates
  - hierarchical single-linkage clustering
  - distance between clusters: 1 – minimum of normalised LCSS
  - constraint: assign single takes to cluster before merging clusters (avoid merging similar scenes before all takes of one scene are clustered)
  - clustering stops when distance reaches the minimum length of a match between takes

## Modified LCSS Algorithm

- transform problem of matching parts to problem of matching sequences of feature vectors
- requirements on matching algorithm
  - match similar, but mostly not identical feature sequences
  - enforce minimum length of matches
- accept gaps and insertions, but enforce maximum length of gap/insertion
- LCSS variant proposed by [Vlachos, Kollios and Gunopoulos, 2002] for 2d trajectories

$$\begin{cases} 0 & \text{if } A \text{ or } B \text{ is empty,} \\ 1 + LCSS_{\delta, \epsilon}(Head(A), Head(B)), & \text{if } |a_{x,n} - b_{x,m}| < \epsilon \text{ and } |a_{y,n} - b_{y,m}| < \epsilon \text{ and } |n - m| \leq \delta \\ \max(LCSS_{\delta, \epsilon}(Head(A), B), LCSS_{\delta, \epsilon}(A, Head(B))), & \text{otherwise} \end{cases}$$

- our modifications:
  - replace  $\epsilon$  by a vector of thresholds  $\{\epsilon_1, \dots, \epsilon_m\}$  for  $m$  features, which are weighted by weights  $\{w_1, \dots, w_m\}$
  - discard  $\delta$  (absolute temporal distance of feature vectors is irrelevant)
  - introduce maximum gap  $\gamma$  between two consecutive matching feature vectors
  - accept all matches longer than minimum length of a take  $\theta_{len}$

## Evaluation

- 6 randomly selected videos from TRECVID 2007 BBC rushes test data set (~ 3 hours)
- manually created ground truth (takes from same camera position, partial takes treated like complete takes, color bars and monochrome frames excluded)
- evaluation method
  - associate take in result with one in ground truth (max. temporal overlap)
  - count number of takes assigned to a scene
  - calculate precision and recall
- results
  - segmentation of shots into parts fails: some takes are merged with others
  - little action and similar location (e.g. dialog scenes) causes wrong assignment of takes

Video	Nr. of scenes		Nr. of takes		Precision	Recall
	ground truth	detected	ground truth	detected		
MRS07063	6	4	26	17	0.7647	0.5000
MRS025913	8	9	28	28	0.8571	0.8571
MRS044731	7	9	34	31	0.5484	0.5000
MRS144760	6	5	24	26	0.8077	0.8750
MRS157475	8	8	36	32	0.6875	0.5903
MS2162107	1	0	26	21	0.6190	0.5000
Mean	7.00	7.50	29.00	25.83	0.7141	0.6405
Median	7.00	8.50	27.00	27.00	0.7261	0.5556

Table 1: Number of scenes and takes in the test set and precision and recall results.