Content Analysis and Summarization for Video Documents

Oral Defense for the degree of Master of Philosophy

Presented by Lu Shi

Committee Members

Prof. T .T. Wong

Prof. M. C. Lee

Prof. Michael R. Lyu (Supervisor)

Prof. Irwin King (Supervisor)

10 December 2004 - 2pm to 4pm

RM 121, Ho Sin Hang Engineering Bldg

Outline

- Introduction
  - Background and motivation
  - Related work
  - Goals
  - Our contributions
- Solution 1: Video summarization by graph modeling and optimization
  - Video structure analysis
  - Video skim length distribution
  - Spatial-temporal graph modeling
  - Optimization based video shot selection
- Solution 2: Video summarization with semantic knowledge
  - Video content annotation
  - Mutual reinforcement principle
  - Video skim selection
- Conclusion

Background and Motivation

- Huge volume of video data are distributed over the Web
- Browsing and managing the huge video database are time consuming
- Video summarization helps the user to quickly grasp the content of a video
- Two kinds of applications:
  - Dynamic video skimming
  - Static video summary
- We mainly focus on generating *dynamic video skimming* for movies

Related work

- Video summarization systems
  - MOCA (dynamic)
  - InforMedia (dynamic)
  - CueVideo (dynamic)
  - Hitchcock (static)
- Limitations:
  - Based on detected feature distribution
  - Neglect that the a video is structured document
  - Lack specific goals that a video summary should achieve

Goals

- Goals for video summarization
  - Conciseness
    - Given the target length of the video skim
  - Content coverage
    - Visual diversity and temporal coverage
    - Balanced structural coverage
  - Visual coherence

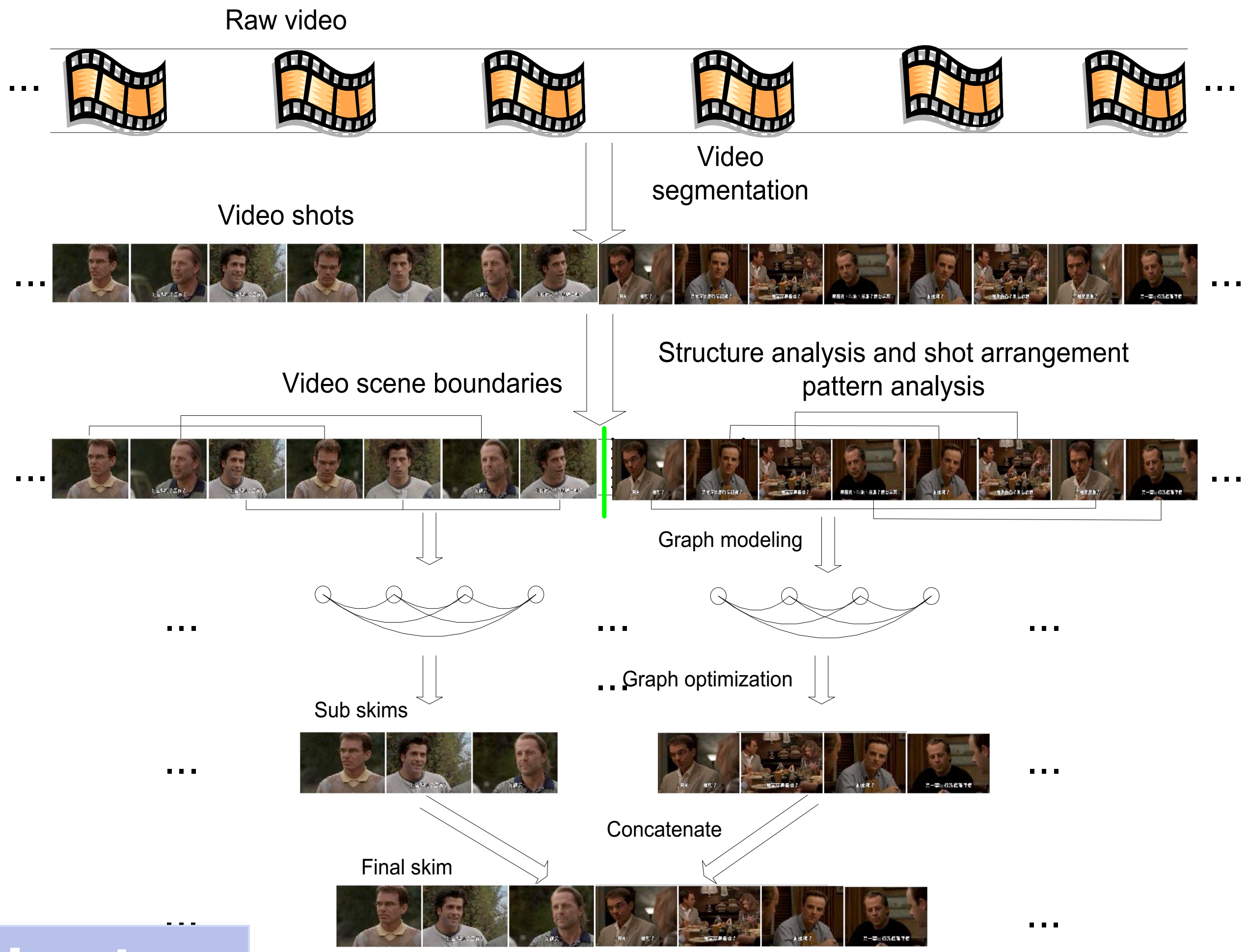
Contributions

- Our contributions:
  - Propose several goals for a good video skim
  - Analyze the video structure information and use it to guide the video skim generation
  - Utilize the video shot arrangement patterns to achieve better coherence
  - Propose the graph optimization based video shots selection to ensure both the visual diversity and the temporal content coverage
  - Employ the semantic knowledge to ensure the quality of the video skimming

Outline

- Introduction
  - Background and motivation
  - Related work
  - Goals
  - Our contributions
- Solution 1: Video summarization by graph modeling and optimization
  - Video structure analysis
  - Video skim length distribution
  - Spatial-temporal graph modeling
  - Optimization based video shot selection
- Solution 2: Video summarization by semantic knowledge
  - Video content annotation
  - Mutual reinforcement principle
  - Video skim selection
- Conclusion

Workflow



Solution 1

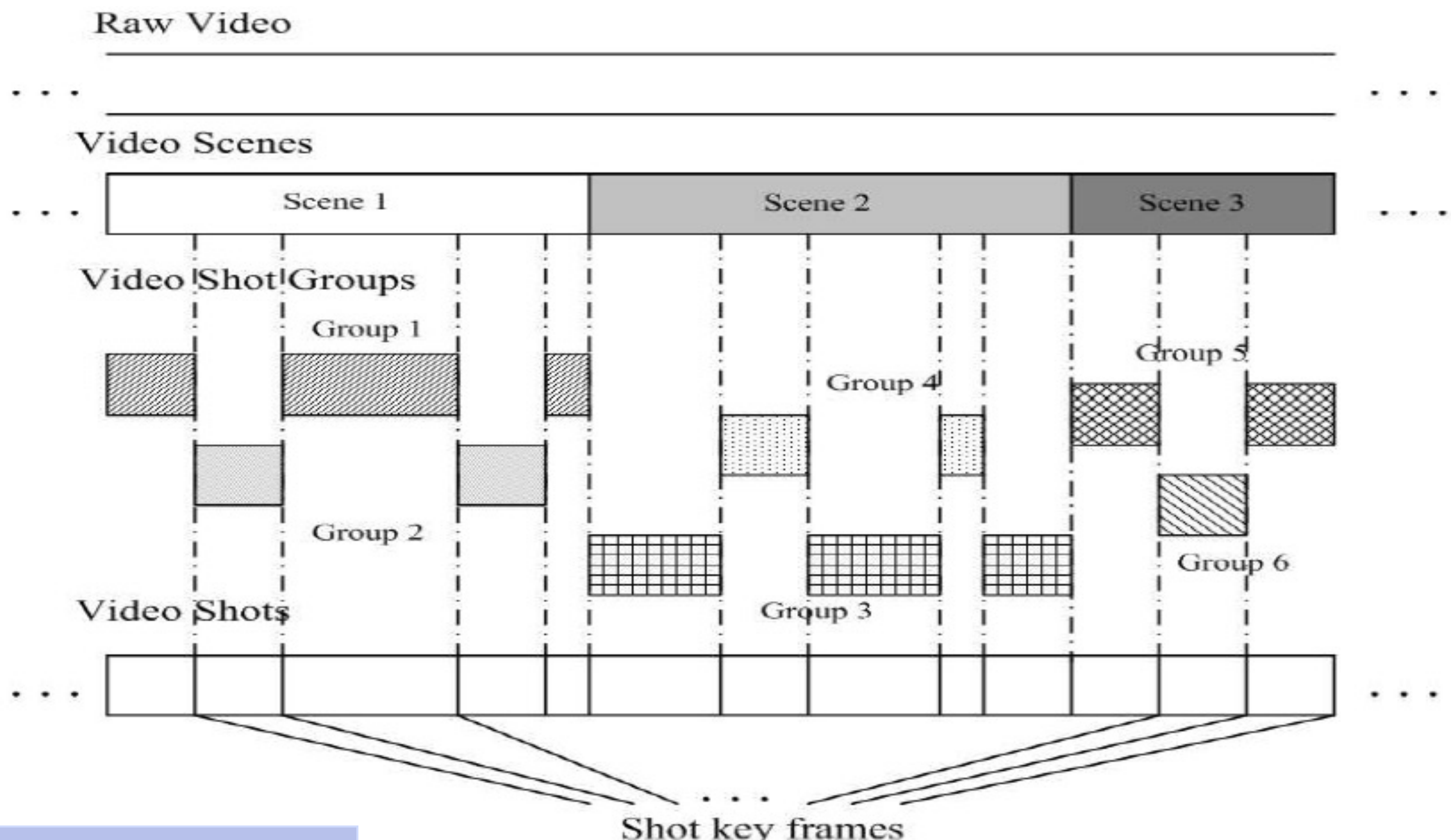
Video Structure

- Video ⇔article
  - Video (story)
  - Video scenes (paragraph)
  - Video shot groups (similar sentences)
  - Video shots (sentence)
  - Video frames

Solution 1

Video Structure

- Hierarchical video structure (Video Table Of Contents)



Solution 1

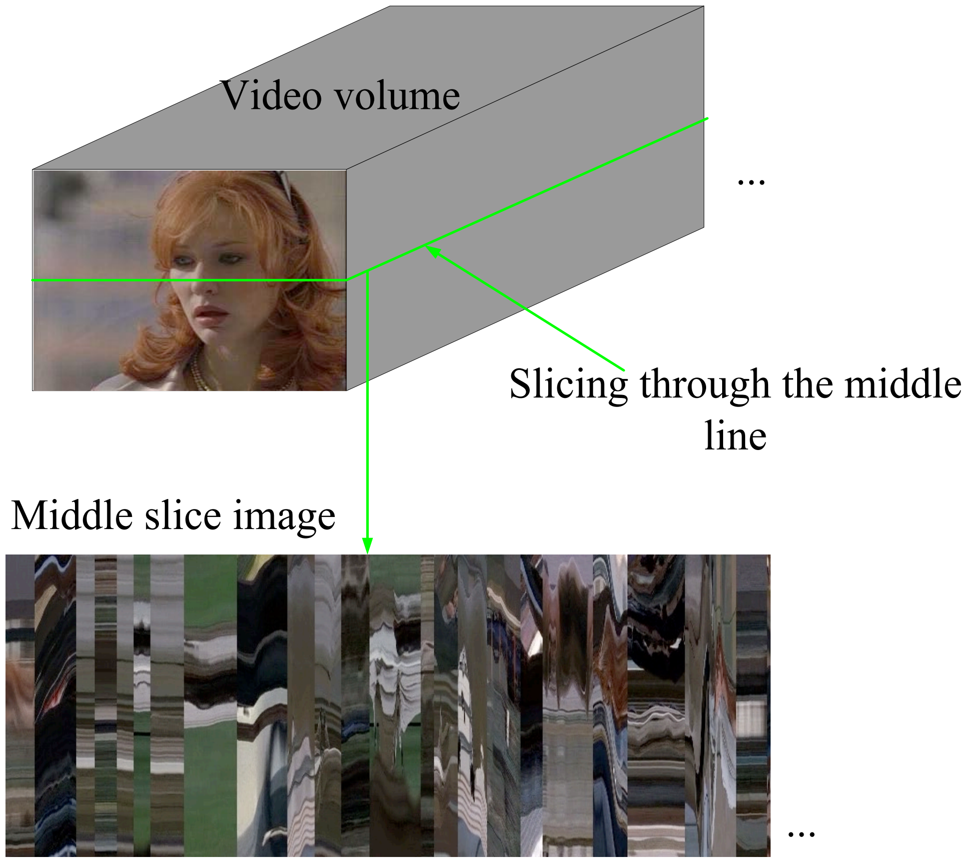
VToC Construction

- Can be built up in a bottom-up manner
  - Video shot detection
  - Video shot grouping
  - Video scene formation

Solution 1

Video Shot Detection

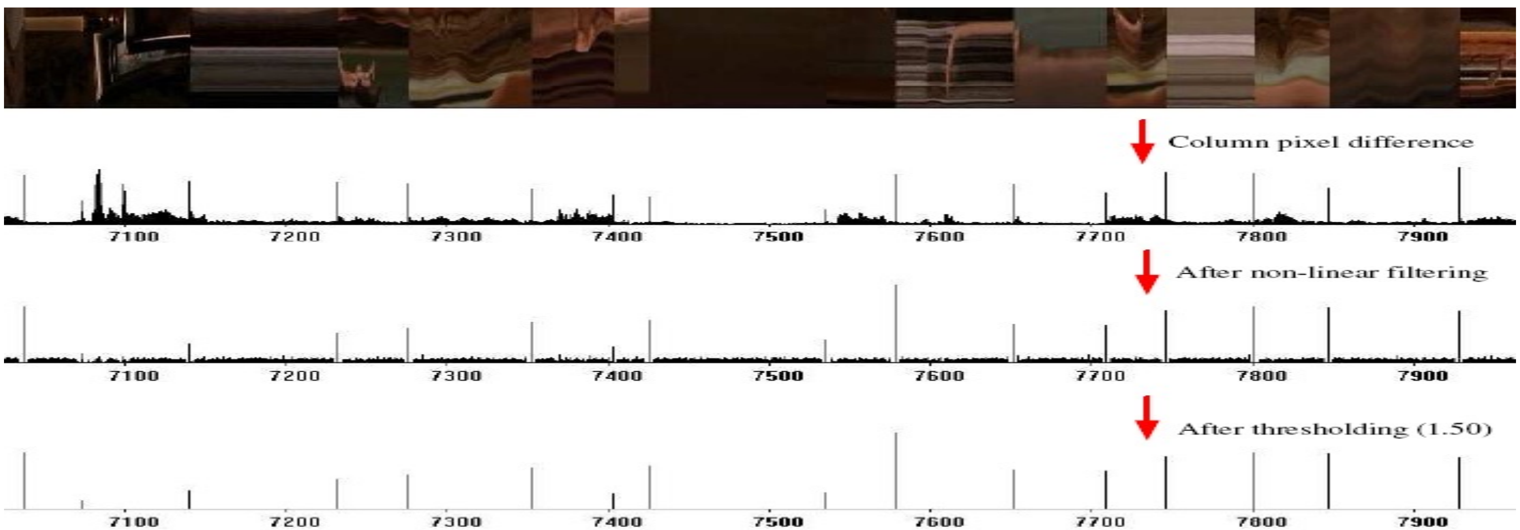
- Video shot detection
  - Video slice image (cut the video from middle line)



Solution 1

Video Shot Detection

- Video shot detection from the middle slice
  - Column - pairwise distance
  - Neighborhood window filtering and thresholding



- …

…

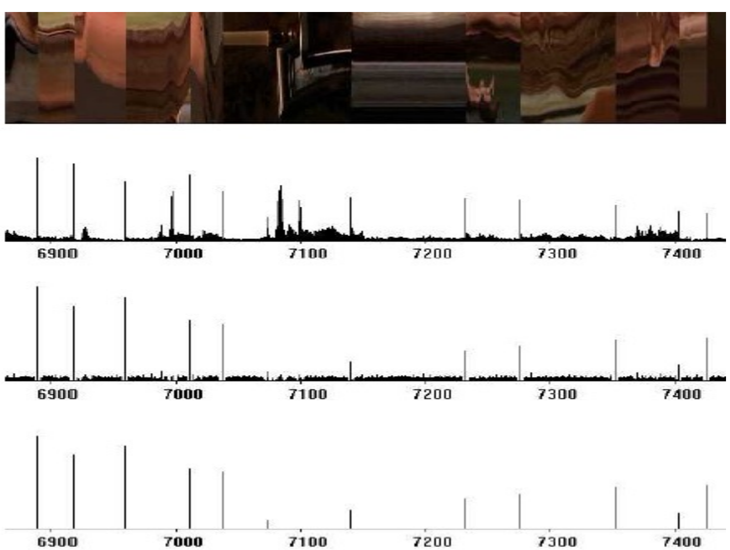
Solution 1

Video Shot Detection

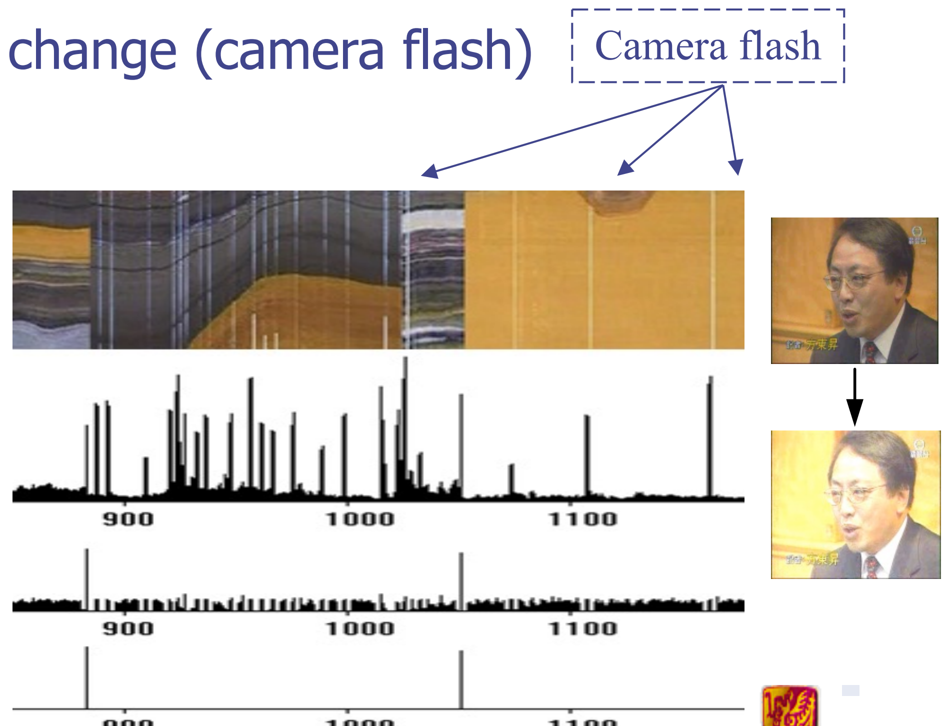
- Neighborhood window filtering
  - Shot cut cues:
    - Local maxima
    - Jump width is 1

$$D'_i = \frac{D_i}{\max_{j=-w, j\neq 0}^{w}(D_{i+j})}$$

  - Robust to sudden lightness change (camera flash)
  - Low computation cost



Normal situation



Flash effect elimination

Solution 1

Evaluation

- Video shot detection result

Table 1: Shot cut detection result for several video clips

Video type	Ground truth	Detected	F. D.	M. D.	Right Per.
Movie	166	157	0	9	94.6
News	40	39	1	1	95
Movie	138	137	2	3	97.8

Solution 1

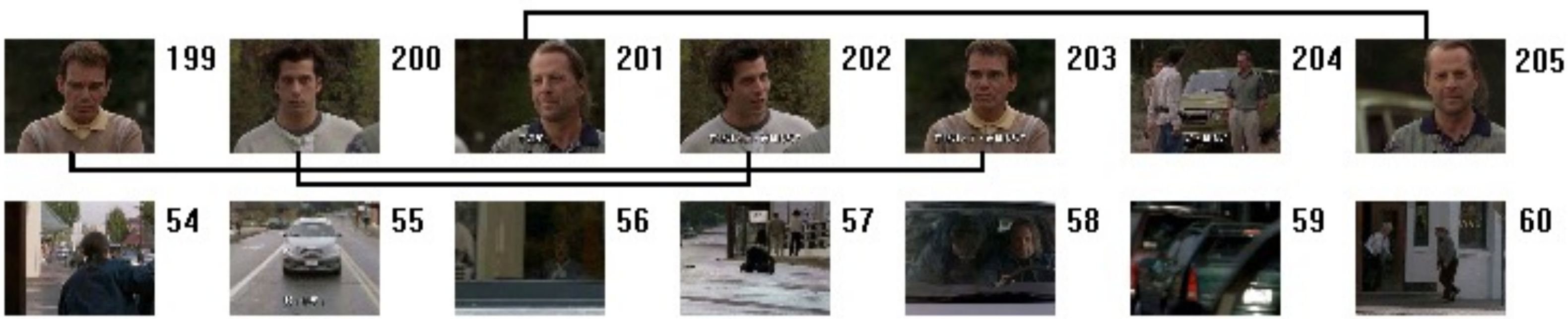
Video shot grouping

- Two methods in the literature:
  - ToC method by Y. Rui, et al
  - Spectral graph partitioning by J. B. Shi, et al

Solution 1

Video Scene Formation

- Loop scenes and progressive scenes
  - Group the visually similar video shots into groups
  - Intersected groups forms loop scenes



  - *Loop scenes* depict an event happened at a place
  - *Progressive scenes*: "transition" between events or dynamic events
- Summarize each video scene respectively

Solution 1

Shot Arrangement Patterns

- The way the director arrange the video shots conveys his intention
- For each scene, video shot group labels form a string (e.g 1232432452……)
- K-Non-Repetitive String (*k-nrs*)
- Minimal content redundancy and visually coherent—good video skim candidates
- String coverage
  - {3124} covers {312,124,31,12,24,3,1,2,4}
- For loop scenes only

Solution 1

Shot Arrangement Patterns

- Several detected *nrs* strings

Shot Arrangement Patterns

- Visual similarity between video shot strings
  - Shot to shot similarity
  - Shot to string similarity
  - String to string similarity
- Break a scene into a set of video shot strings
  - Given the upper bound of the string length l_{nrs}
  - Directly break from left to right
  - Example: {1234343152} is broken into a set of *nrs* strings {123, 43, 431, 52} under $l_{nrs} = 3$

Solution 1

Video Scene Analysis

- Scene importance: length and complexity
- Content entropy for loop scenes
- Measure the complexity for a loop scene

Length of a member video shot group

$$Entropy(Sc_i) = \sum_j -\frac{l_{Sg_j}}{l_{Sc_i}} \log\left(\frac{l_{Sg_j}}{l_{Sc_i}}\right)$$

Total length of the video scene

- For progressive scenes, we only consider its length

Solution 1

Skim Length Distribution

- Determine each video scene's target skim length, given L_{vs}
  - Determine each progressive scenes' skim length
    - If $l_{Sc_i} \times \frac{L_{vs}}{L_v} < t_1$, discard it, else $L^i{}_{vs} = l_{Sc_i} \times \frac{L_{vs}}{L_v}$
  - Determine each loop scenes' skim length
    - If $L^i{}_{vs} = L'_{vs} \times \frac{l_{Sc_i} \times Entropy(Sc_i)}{\sum_j l_{Sc_j} \times Entropy(Sc_j)} < t_2$,discard it
    - Redistribute L'_{vs} to remaining scenes

Graph Modeling of Video Scenes

- Visual-temporal dissimilarity function
  - Linear with visual dissimilarity
  - Exponential with temporal distance

$$Dis(str_i, str_j) = 1 - VisualSim(str_i, str_j) \times e^{-k(TemporalDis(str_i, str_j))}$$

Visual similarity (color, motion, texture…)

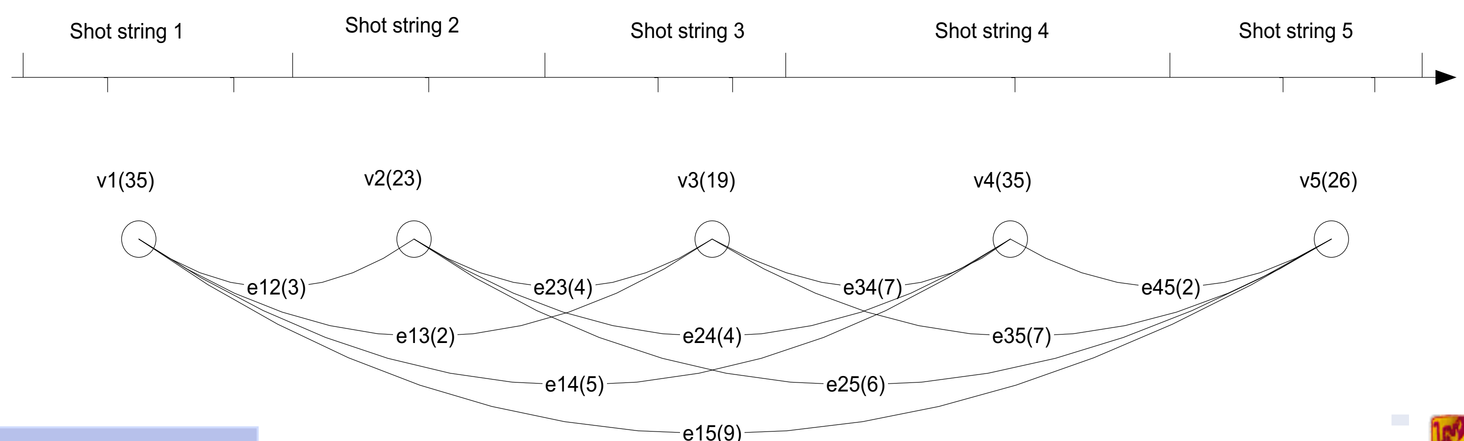
Slope control

Temporal distance between shot middle frames

Solution 1

Graph Modeling of Video Scenes

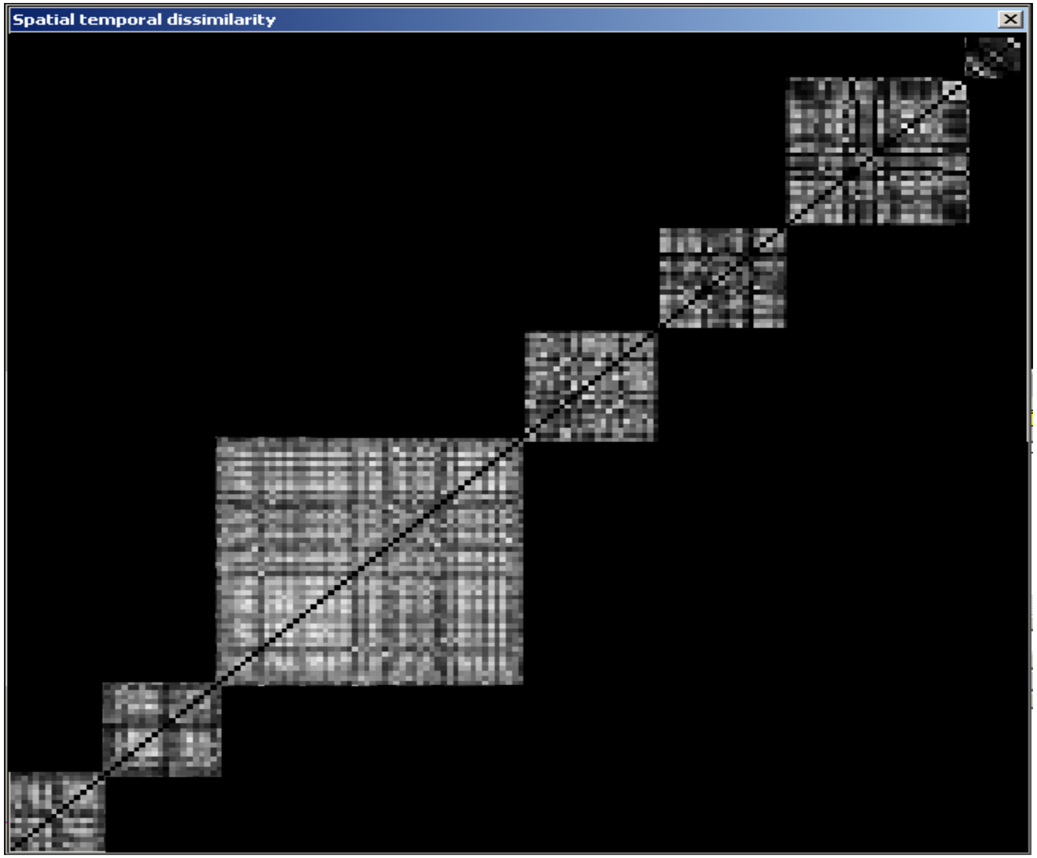
- The visual temporal relation graph
  - Each vertex corresponds to a video shot string
  - Each edge corresponds to the dissimilarity function between shot strings
  - Directional and complete



Solution 1

Graph Modeling of Video Scenes

- Dissimilarity function between video shots in a video with 7 scenes



Solution 1

Skim Generation

- The goal of video skimming
  - Conciseness: for each scene, given the target skim length L^{i}_{vs}
  - Content coverage
  - Coherence
- The visual temporal relation graph
  - A path corresponds to a series of video shot strings
  - Vertex weight summation
  - Path length is the summation of the dissimilarity between consecutive vertex pairs

Constrained Longest Path

- Objectives:
  - Search for a path p_s for each scene, such that:
    - Maximize the path length (dissimilarity summation)
    - Vertex weight summation should be close to L^i_{vs} but not exceed it
- The objective function

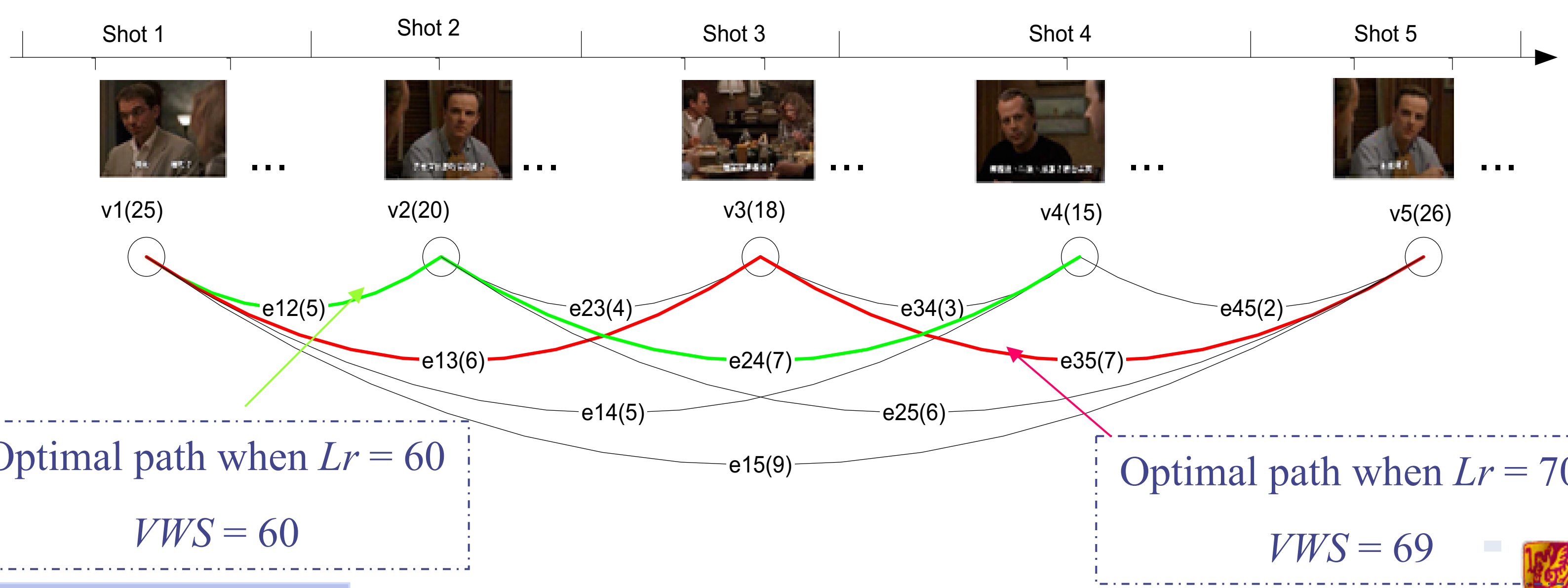
$$f_{obj}(p_s, L^i_{vs}) = L_{p_s} + w \times (VWS(p_s) - L^i_{vs}), VWS(p_s) <= L^i_{vs}$$

Path length

Summation of the shot length

Constrained Longest Path

- Global optimal solution
- Let $\{p^i_{v_x,L_r}\}$ denote the paths begin with v_x , whose vertex weight summation is upper bounded by L_r
- The optimal path is denoted by $f_{obj}(p^o_{v_0,L_r}) = \max_i f_{obj}(p^i_{v_0,L_r})$



Solution 1

Graph Optimization

- Optimal substructure

$$f_{obj}(p^{o}_{v_x,L_r}) = \max_{v_i=v_x+1}^{v_n}(f_{obj}(p^{o}_{v_i,L_r-l_{stri}}) + Dis(str_x, str_i) + w \times l_{sh_i}), x < n$$

$$f_{obj}(p^{o}_{v_n,L_r}) = w \times (l_{sh_n} - L^{i}{}_{v_s}), x = n$$

- Dynamic programming
  - Effective way to compute the global optimal solution
  - Trace back to find the optimal path
  - Time complexity $O(n^2 \times L^{i}{}_{vs})$, space complexity $O(n \times L^{i}{}_{vs})$

Evaluation

- Key frames of selected video shots

Solution 1

Evaluation

- Subjective experiment:10 people were invited to watch video skims generated from 4 videos with rate 0.15 and 0.30
- Questions about major events: Who has done What? (Meaningfulness)
- Which video skim looks better? (Favorite)
- Mean scores are scaled to 10.00
- Parameters: $t_1 = 3\sec, t_2 = 4\sec, w = 0.01, k = 250$

Video Clip	Duration	Major events	Skim Rate	Mfn.	Fav.
Movie 1	1403 sec.	7	0.15	82.9/**85.7**	4/**6**
			0.30	94.3/**92.9**	3/**7**
Movie 2	1230 sec.	8	0.15	83.8/**81.3**	4/**6**
			0.30	92.9/**96.3**	2/**8**
Movie 3	477 sec.	5	0.15	82.0/**86.0**	4/**6**
			0.30	94.0/**92.0**	5/**5**
Sitcom 1	1183 sec.	9	0.15	71.1/**76.7**	3/**7**
			0.30	84.4/**88.9**	3/**7**

TABLE I

USER TEST RESULTS. THE SCORES WITH l_{str} IS EQUAL TO 3 ARE IN **BOLD**

Solution 1

Summary

- Video structure analysis
  - Scene boundaries, sub-skim length determination
- Graph modeling for video scenes
- Model the sub skim generation problem as a constrained longest path problem
- Generate a video skim

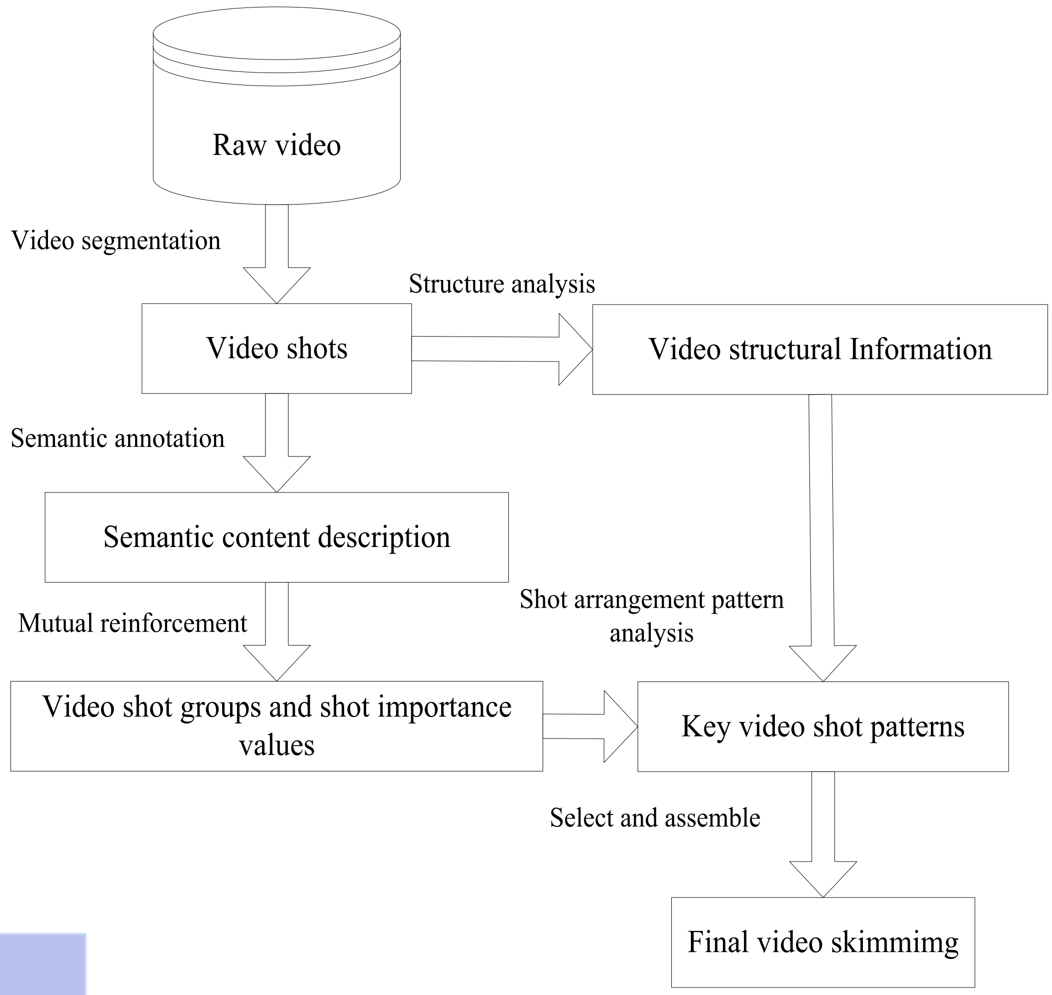
Outline

- Introduction
  - Background and motivation
  - Related work
  - Goals
  - Our contributions
- Solution 1: Video summarization by graph modeling and optimization
  - Video structure analysis
  - Video skim length distribution
  - Spatial-temporal graph modeling
  - Optimization based video shot selection
- Solution 2: Video summarization by semantic knowledge
  - Video content annotation
  - Mutual reinforcement principle
  - Video skim selection
- Conclusion

Video Semantics

- Low level features and high level concepts: semantic gap
- Summary based on low level features is not able to ensure the perceived quality
- Solution: obtain video semantic information by manual/semi-automatic annotation
- Usage:
  - Retrieval
  - Summary

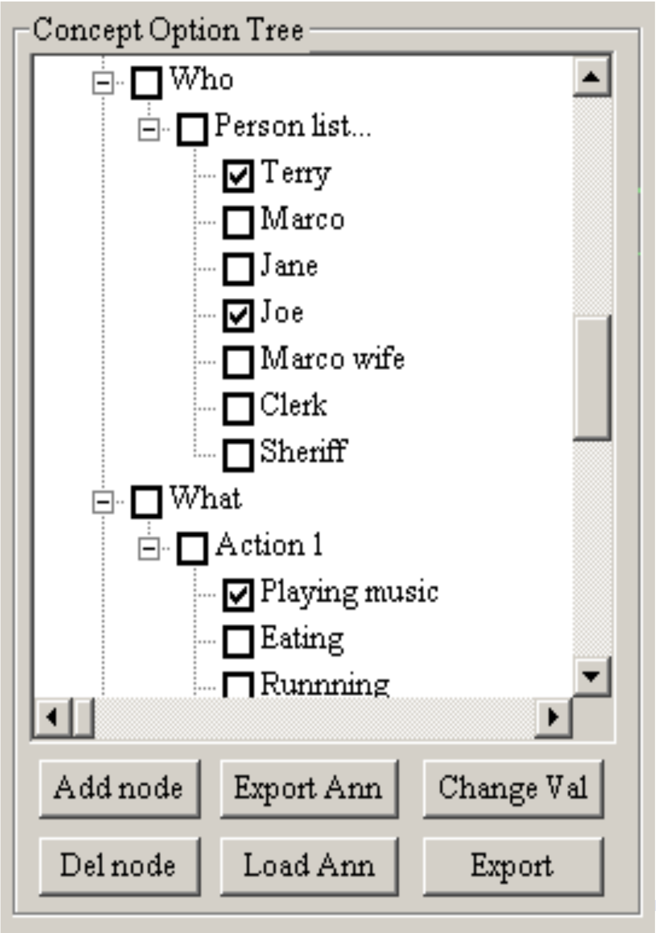
System Overview



Solution 2

Video Semantics

- Concept representation for a video shot
  - The most popular question: who has done what?
  - The two major contexts: who, what action
- Concept term and video shot description (user editable and reusable)



Solution 2

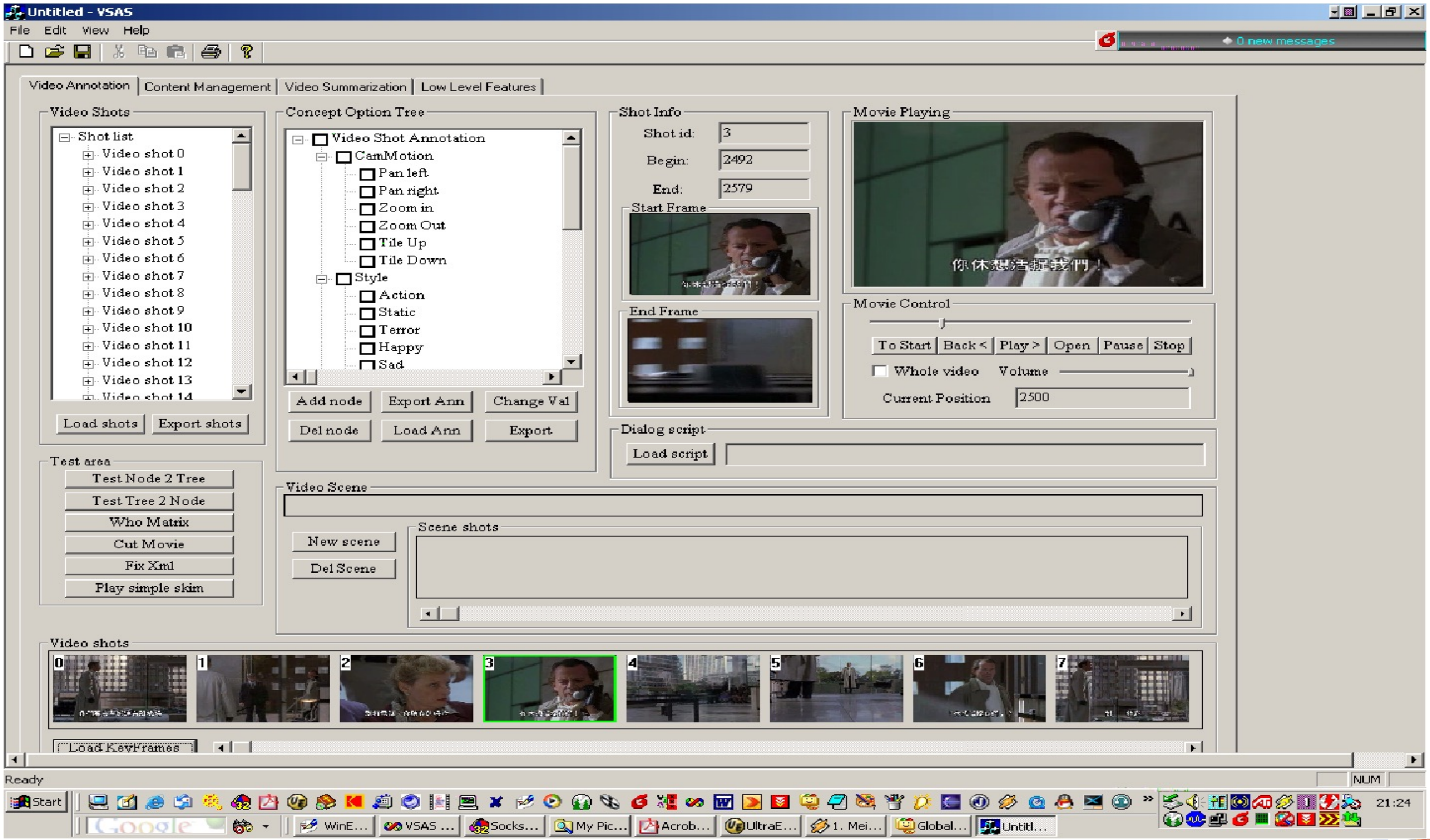
Video Semantics

- Concept term and video shot description
  - Term (key word): denote an entity, e.g. "Joe", "talking", "in the bank"
  - Context: "who", "what action"…
  - Shot description: the set comprising all the concept terms that is related to the shot $\{t_1, \ldots, t_n\}$
- Obtained by semi-automatic or video annotation

Solution 2

Video Content Annotation

- Annotation interface



Solution 2

Video Summarization

- Obtain the structure of the video
- Derive an importance measure for video shots
- Reselect some “important” shots then arrange them into a trailer
- An “inversion” of video editing

Mutual Reinforcement

- How to measure the priority for a set of concept terms and a set of descriptions?
  - A more important description should contain more important terms;
  - A more important term should be contained by more important descriptions
- Mutual reinforcement principle

Solution 2

Mutual Reinforcement

- Let *W* be the weight matrix describes the relationship between the term set and shot description set (elements in *W* can have various definitions, e.g. the number of occurrence of a term in a description)
- Let *U*, *V* be the vector of the importance value of the concept term set $\{d_i\}$ and video shot description set $\{t_i\}$
- We have

$$U = \frac{1}{k_1} WV, \qquad V = \frac{1}{k_2} W^T U$$

  Where k_1 and k_2 are constants.
- *U* and *V* can be calculated by SVD of *W*

Solution 2

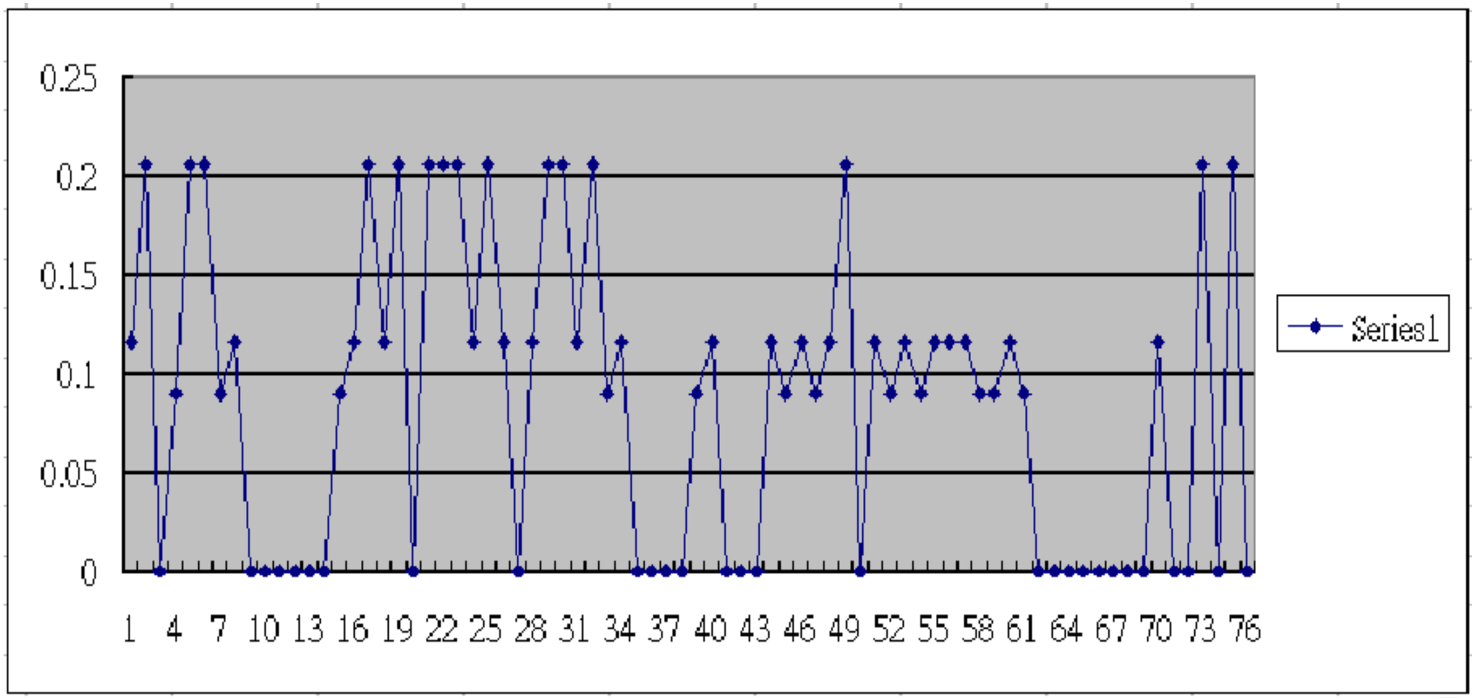
Mutual Reinforcement

- For each semantic context:
- We choose the singular vectors correspond to W's largest singular value as the importance vector for concept terms and sentences
- Since W is non-negative , the first singular vector V will be non-negative

Solution 2

Mutual Reinforcement

- Importance calculation on 76 video shots
- Based on context "who"



Solution 2

Mutual Reinforcement

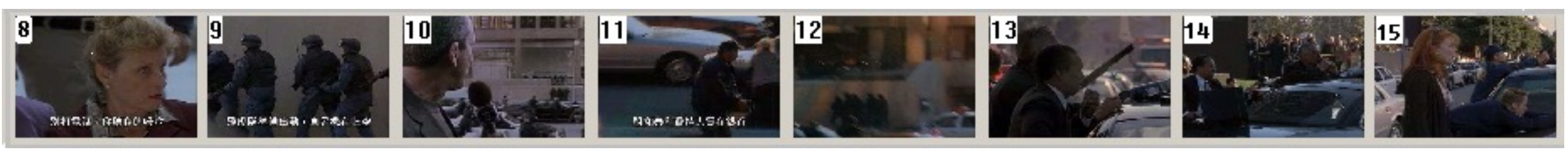
- Shots with different importance values "who"

Joe and Terry

Terry

Joe

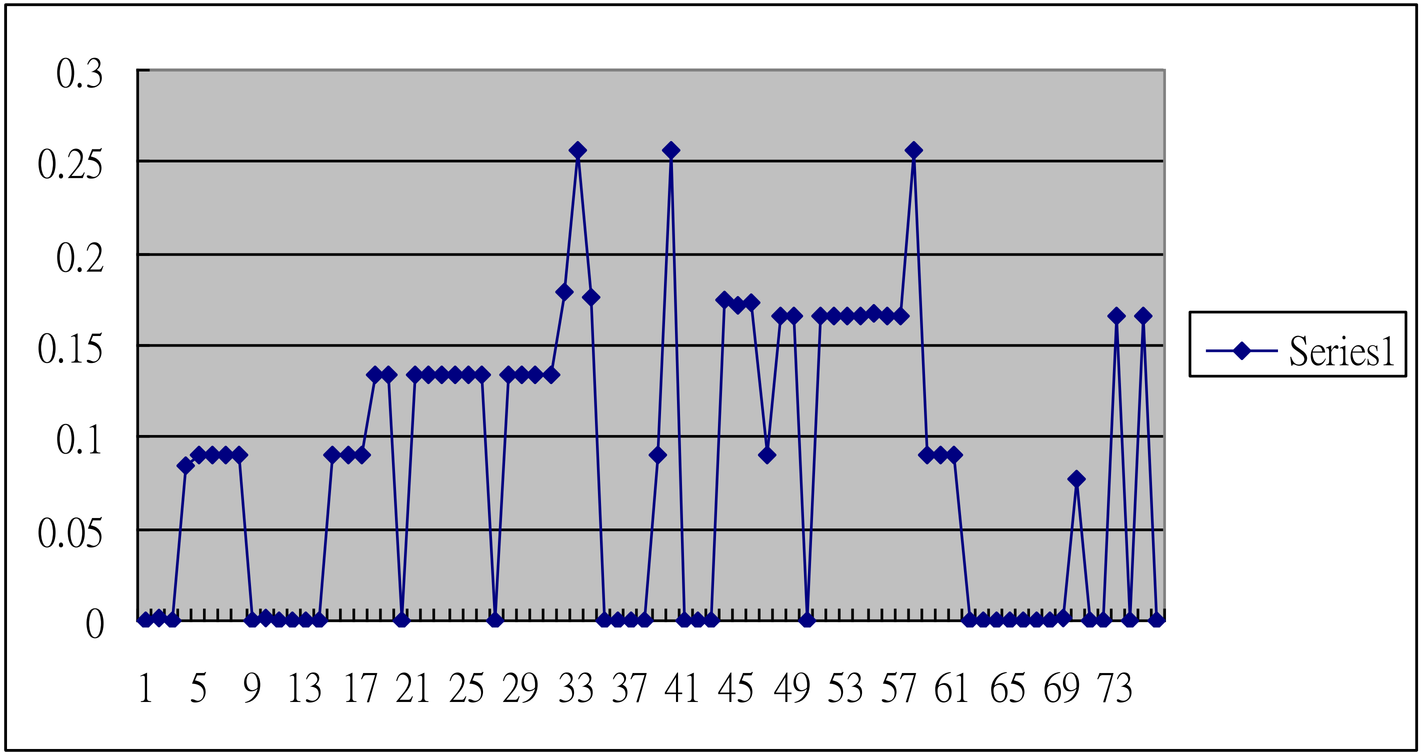
Background people



Solution 2

Mutual Reinforcement

- Priority calculation
- Based on context “what action”



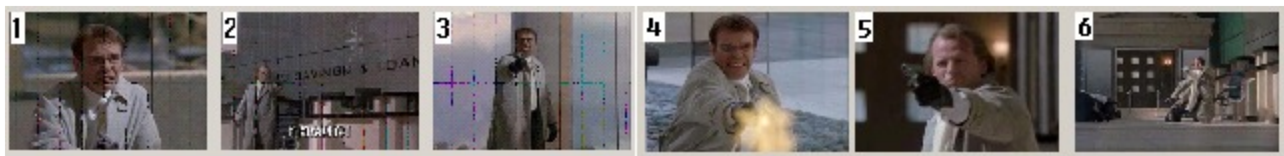
Solution 2

Mutual Reinforcement

- Shot groups

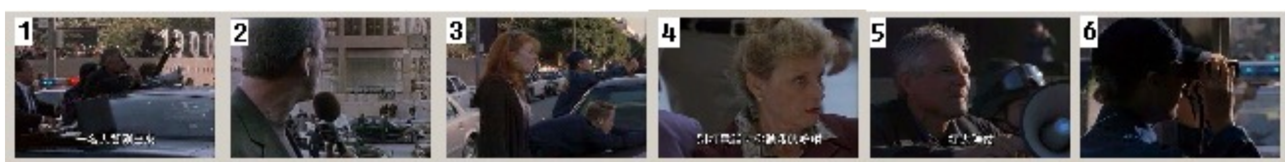
Gun shot and quarrel

Gun shot



Quarrel

Observing



No “action”

Solution 2

Video Summarization

- Based on the result of mutual reinforcement, we can determine the relational priority between video shots

$$V = V_{what} + V_{who}$$

- The generated skim can ensure the semantic contents coverage

Shot Arrangement Patterns

- The way the director arrange the video shots conveys his intention
- Minimal content redundancy and visual coherence
- Semantic video shot group label form a string
- K-Non-Repetitive Strings (*k-nrs*)
- String coverage
  - {3124} covers {312,124,31,12,24,3,1,2,4}
- The importance value of a *nrs* string: summation of the member shots

Video Skim Selection

- Input: the decomposed *nrs* string set from a scene
- do
  - Select the most important *k-nrs* string into the skim shot set
  - Remove those *nrs* strings from the original set covered by the selected string
- Until the target skim length is reached

Video Skim Selection

Input: The set of all nrs strings NRS; The target skimming length L_{vs};
Output: The selected nrs set $SKIM$ that form the video skimming
BEGIN $SKIM = \emptyset$
STEP 1: Sort the nrs strings in NRS according to their importance value;
while $L_{vs} > 0$ **do**
Select the best nrs string nrs_{opt}, such that:

1. $L_{nrs_{opt}} < L_{vs}$
2. $\forall nrs_i \in N$ and $L_{nrs_i} < L_{vs}$, $I_{nrs_{opt}} \geq I_{nrs_i}$

if Found **then**
1. $SKIM = S \cup \{nrs_{opt}\}$
2. $L_{vs} = L_{vs} - L_{nrs_{opt}}$
3. $NRS = NRS - \{nrs_t | nrs_{opt} \text{ covers } nrs_t\}$

else if Not found **then**
GOTO END
end if
end while
END

Solution 2

Evaluation

- We conduct the subjective test
- Compared with the previous graph based algorithm
- Achieve better coherency

Video Clip	Duration	Events	Skim Rate	Mfn.	Fav.
Movie1	1403 sec.	7	0.15	82.9/**78.6**	3/7
			0.30	94.3/**97.1**	2/**8**
Movie2	1230 sec.	8	0.15	83.8/**85.0**	2/**8**
			0.30	92.9/**96.3**	2/**8**
Movie3	477 sec.	5	0.15	82.0/**88.0**	4/**6**
			0.30	94.0/**94.0**	2/**8**
Sitcom1	1183 sec.	9	0.15	71.1/**73.3**	3/7
			0.30	84.4/**88.8**	3/7

Table 1: User test results. Scores for the new approach are **bold**

Outline

- Introduction
  - Background and motivation
  - Related work
  - Our contributions
- Video summarization by graph modeling and optimization
  - Video structure analysis
  - Video skim length distribution
  - Spatial-temporal graph modeling
  - Optimization based video shot selection
- Video summarization by semantic knowledge
  - Video content annotation
  - Mutual reinforcement principle
  - Video skim selection
- Conclusion

Conclusion

- In this presentation, we have:
  - Discussed the video summarization problem
  - Proposed three goals that a good video skim should achieve
  - Described two solutions to generate useful video skims
    - Graph modeling and optimization
    - Mutual reinforcement principle
- Future work:
  - More efficient way to annotate video shots
  - Augment the semantic template
  - Comply to MPEG-7 standard
  - Personalized video summary
  - New evaluation method

Publication list

- “Video summarization by greedy method in a constraint satisfaction framework”, S. Lu, I. King and M. R. Lyu, in proceedings of DMS 2003
- “Video summarization by spatial-temporal graph optimization”, S. Lu, M. R. Lyu and I. King, in proceedings of ISCAS 2004
- “Video summarization by video structure analysis and graph optimization”, S. Lu, I. King and M. R. Lyu, in proceedings of ICME 2004
- “Semantic video summarization by mutual reinforcement principle and shot arrangement patterns”, S. Lu, M. R. Lyu and I. King, accepted by MMM2005, to appear
- “A novel video summarization framework for document preparation and archival applications”, S. Lu, I. King and M. R. Lyu, accepted by IEEE Aerospace05, to appear

Q & A

Thank you!