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 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
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
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
Video Shot Detection

- Video shot detection
  - Video slice image (cut the video from middle line)
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
- Robust to sudden lightness change (camera flash)
- Low computation cost

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

Solution 1

Normal situation

Flash effect elimination
Evaluation

Video shot detection result

Table 1: Shot cut detection result for several video clips

<table>
<thead>
<tr>
<th>Video type</th>
<th>Ground truth</th>
<th>Detected</th>
<th>F. D.</th>
<th>M. D.</th>
<th>Right Per.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Movie</td>
<td>166</td>
<td>157</td>
<td>0</td>
<td>9</td>
<td>94.6</td>
</tr>
<tr>
<td>News</td>
<td>40</td>
<td>39</td>
<td>1</td>
<td>1</td>
<td>95</td>
</tr>
<tr>
<td>Movie</td>
<td>138</td>
<td>137</td>
<td>2</td>
<td>3</td>
<td>97.8</td>
</tr>
</tbody>
</table>
Video shot grouping

Two methods in the literature:

- ToC method by Y. Rui, et al
- Spectral graph partitioning by J. B. Shi, et al
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
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

Solution 1
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 $l_{nrs} = 3$
Video Scene Analysis

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

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

- Length of a member video shot group
- 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 $L_{vs}^i = l_{Sc_i} \times \frac{L_{vs}}{L_v}$
  - Determine each loop scenes’ skim length
    - If $L_{vs}^i = L_{vs}' \times \frac{l_{Sc_i} \times Entropy(Sc_i)}{\sum_j l_{Sc_j} \times Entropy(Sc_j)} < t_2$, discard it
    - Redistribute $L_{vs}'$ to remaining scenes

Solution 1
Graph Modeling of Video Scenes

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

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

- 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
Graph Modeling of Video Scenes

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

\[ f_{obj}(p_s, L^i_{vs}) = L_{p_s} + w \times (VWS(p_s) - L^i_{vs}), VWS(p_s) \leq L^i_{vs} \]
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**

Optimal path when \( L_r = 60 \)

\( VWS = 60 \)

Optimal path when \( L_r = 70 \)

\( VWS = 69 \)
Graph Optimization

Optimal substructure

\[ f_{\text{obj}} (p_{v_x,L_r}^o) = \max_{v_i=v_x+1}^{v_n} (f_{\text{obj}} (p_{v_i,L_r-l_{\text{str}i}}^o) + \text{Dis}(\text{str}_x, \text{str}_i) + w \times l_{sh_i}), x < n \]

\[ f_{\text{obj}} (p_{v_n,L_r}^o) = w \times (l_{sh_n} - L_{vs}^i), x = n \]

Dynamic programming

- Effective way to compute the global optimal solution
- Trace back to find the optimal path
- Time complexity \( O(n^2 \times L_{vs}^i) \), space complexity \( O(n \times L_{vs}^i) \)
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: $t_1 = 3 \text{ sec}, t_2 = 4 \text{ sec}, w = 0.01, k = 250$

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

**TABLE I**

User test results. The scores with $l_{str}$ is equal to 3 are in **bold**
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

Solution 1
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
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
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

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

Where $k_1$ and $k_2$ are constants.

$U$ and $V$ can be calculated by SVD of $W$
Mutual Reinforcement

- For each semantic context:

  - We choose the singular vectors correspond to $\mathbf{W}$'s largest singular value as the importance vector for concept terms and sentences.

- Since $\mathbf{W}$ is non-negative, the first singular vector $\mathbf{V}$ will be non-negative.
Mutual Reinforcement

- Importance calculation on 76 video shots
- Based on context “who”
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**

1. Gun shot and quarrel
2. Gun shot
3. Quarrel
4. Observing
5. No “action”

**Solution 2**
Video Summarization

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

\[ V = V_{\text{what}} + V_{\text{who}} \]

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. $NRS = NRS - \{nrs_{i}|nrs_{opt} \text{ covers } nrs_{i}\}$  
    else if Not found then  
    GOTO END  
    end if  
end while  
END
Evaluation

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

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

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

Summary
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!