Video Content Description: From Low-Level Features to Semantics

A. Murat Tekalp
University of Rochester
www.ece.rochester.edu/~tekalp

PhD Students
Ahmet Ekin, A. Mufit Ferman,
Yaowu Xu
Overview

- Learning from Examples
- Background: MPEG-7
- Low-Level and Semantic-Level Modeling of Object Motion
- An Integrated Semantic-Syntactic Video Model and Model-Based Query Processing
- Automatic Frame-Based Video Summarization: From Low-Level Features to Semantic-Level
Learning from Examples
1. “Learn” from examples to extract and annotate semantic objects automatically
2. Perform low-level feature, such as color and shape, between database images and example template

Application: Personal Image Library System
Formation of the Initial Content Hierarchy

- Color Edge Detection
- Color Segmentation
  - Region Formation by integration of multiple cues
    - Split regions containing edges into multiple regions
    - Merge regions using a highest confidence decision method
- Initialization of the Content Hierarchy
### Examples - Shape Similarity Matching

<table>
<thead>
<tr>
<th>Original</th>
<th>Segmentation</th>
<th>Initial Hierarchy</th>
<th>Best Match</th>
<th>Final Hierarchy</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="307x259.jpg" alt="Image" /></td>
<td><img src="307x259.jpg" alt="Image" /></td>
<td><img src="307x259.jpg" alt="Image" /></td>
<td><img src="307x259.jpg" alt="Image" /></td>
<td><img src="307x259.jpg" alt="Image" /></td>
</tr>
<tr>
<td><img src="307x153.jpg" alt="Image" /></td>
<td><img src="307x153.jpg" alt="Image" /></td>
<td><img src="307x153.jpg" alt="Image" /></td>
<td><img src="307x153.jpg" alt="Image" /></td>
<td><img src="307x153.jpg" alt="Image" /></td>
</tr>
<tr>
<td><img src="307x260.jpg" alt="Image" /></td>
<td><img src="307x260.jpg" alt="Image" /></td>
<td><img src="307x260.jpg" alt="Image" /></td>
<td><img src="307x260.jpg" alt="Image" /></td>
<td><img src="307x260.jpg" alt="Image" /></td>
</tr>
</tbody>
</table>
### Examples - Color Similarity Matching

<table>
<thead>
<tr>
<th>Query template</th>
<th>Original</th>
<th>Segmentation</th>
<th>Initial hierarchy</th>
<th>Best match</th>
<th>Final hierarchy</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Query Image" /></td>
<td><img src="image2" alt="Original Image" /></td>
<td><img src="image3" alt="Segmentation Image" /></td>
<td><img src="image4" alt="Initial hierarchy Image" /></td>
<td><img src="image5" alt="Best match Image" /></td>
<td><img src="image6" alt="Final hierarchy Image" /></td>
</tr>
</tbody>
</table>
Hierarchical Content Matching

- **Query Modes**
  - High-level queries at the object level
    Example: Find “cars”
  - Low-level queries at the region level
    Example: Find a blue color region

- **Matching Measure**
  - Color histogram intersection
  - Hausdorff distance

- **Hierarchical Content Matching**
  - Top-down fashion
  - Highest-level (composite) nodes first.
  - No match in higher level, go to lower level
Semantic Segmentation

Images

Segmentations

Low Level

Semantic Level
Overview of Some Elements of MPEG-7
MPEG-7 Segment DS

Still Regions

Segment Decomposition

Still Region

Audio-visual segment

<table>
<thead>
<tr>
<th>Feature</th>
<th>Video Segment</th>
<th>Still Region</th>
<th>Moving Region</th>
<th>Audio Segment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time</td>
<td>X</td>
<td>.</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Shape</td>
<td>.</td>
<td>X</td>
<td>X</td>
<td>.</td>
</tr>
<tr>
<td>Color</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>.</td>
</tr>
<tr>
<td>Texture</td>
<td>.</td>
<td>X</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>Motion</td>
<td>X</td>
<td>.</td>
<td>X</td>
<td>.</td>
</tr>
<tr>
<td>Camera motion</td>
<td>X</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>Mosaic</td>
<td>X</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>Audio features</td>
<td>.</td>
<td>.</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>
Segment Features
- Text Annotation
- Time
- Mosaic

Segment Features
- Text Annotation
- Time
- Mosaic

Segment Relation

<table>
<thead>
<tr>
<th>Relation</th>
<th>Inverse Relation</th>
</tr>
</thead>
<tbody>
<tr>
<td>before</td>
<td>after</td>
</tr>
<tr>
<td>meets</td>
<td>metBy</td>
</tr>
<tr>
<td>overlaps</td>
<td>overlappedBy</td>
</tr>
<tr>
<td>during</td>
<td>contains</td>
</tr>
<tr>
<td>strictDuring</td>
<td>strictContains</td>
</tr>
<tr>
<td>starts</td>
<td>startedBy</td>
</tr>
<tr>
<td>finishes</td>
<td>finishedBy</td>
</tr>
<tr>
<td>equal</td>
<td>equal</td>
</tr>
</tbody>
</table>
MPEG-7: Semantic DS

Semantic Relations:
- Object to object
- Object to event
- Event to event
- STime to event
- SPlace to event
- SBase to segment

Narrative World

state change

"Carnegie Hall" "7-8pm, Oct. 14, 1998"

SemanticState DS

SemanticTime DS

SemanticPlace DS

Object DS

Event DS

Concept DS

PersonObject DS

"Tom's tutor" "Tom Daniels" "piano" "play"

interpretation, non-perceivable abstraction, non-perceivable

musician

interpretation, perceivable
MPEG-7: Summarization DS

Key frames; Key video clips; Key audio clips; Key events; mixed.
Video Analysis / Feature Extraction

- Shot-Based Processing
- Object-Based Processing
- Low-level descriptions

CM I
CM II
CM N

Mapping

Semantic Description
- People/Objects
- Location
- Actions/Events
- Time
Low-Level and Semantic-Level Modeling of Object Motion
From Low-Level to Semantic Level

Image Processing

- Color
- Texture
- Shape
- Motion
- Shot Boundaries

Low-Level Visual Features

Semantic Content

- People/Objects
- Location
- Events/Actions
- Time

Mapping

Context
Goal

Object-based motion description at the low level
- Parametric motion (PM) - Dominant motion
- Motion trajectory (MT)
- Motion activity (MA)

Object-based motion description at the semantic level
- Actions
- Interactions
- Events
Problems

- The lifetime of a video object or even a shot is too coarse a temporal resolution to describe its motion both semantically and at the low level.
- We define segments which enable meaningful description of semantic and low-level motion of objects and interactions between them.
- We describe scene motion (events) by composing object actions and object-to-object interactions.
Object-Based Motion Description

Object Descriptions

- Temporal Descriptions
  - comprised of
    - Action Unit Descriptions
    - Interaction Unit Descriptions

- Spatial Descriptions
  - comprised of
    - Elementary Motion Unit Descriptions
    - Elementary Reaction Unit Descriptions

in terms of
Parametric Motion Descriptor

<table>
<thead>
<tr>
<th>ModelCode</th>
<th>Meaning</th>
<th>Number of parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Translational model</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>Rotation/scale model</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>Affine model</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>Perspective model</td>
<td>8</td>
</tr>
<tr>
<td>4</td>
<td>Quadratic model</td>
<td>12</td>
</tr>
</tbody>
</table>

StartTime specifies the beginning of the temporal interval.
EndTime specifies the end of the temporal interval.

MotionParameters[] is a floating point array that keeps the values of the model parameters. Its size depends on the motion model specified by ModelCode.

SpatialRegion a pointer to the spatial region the model is associated with.

Xorigin, Yorigin are the coordinates of the origin of the spatial reference with respect to the image coordinates.
Extraction of Motion Descriptions

Video Data

Scene segmentation

Object segmentation

Camera Compensation

EMU Segmentation

Low-Level Motion Feature Extraction

AU Generation

Semantic Motion Content Abstraction

Representation & Database Index
Procedure

Identify/Segment Objects of Interest in the video

Segment the lifespan of each object into elementary motion units

Extraction descriptions of the elementary motion units

Identify action units

Extract description of action units
Detection of EMUs

1. Estimate the parameters of the dominant motion of the object between at each pair of frames.
2. Detect the initial set of elementary motion units by identifying segments with coherent object motion.
3. Refine the initial set of elementary motion units to obtain the final set of elementary motion units.
Example: Children Sequence

EMU 1  EMU 2
AU “throw the ball”

EMU 3  EMU 4  EMU 5
AU “pick up the ball and throw it”

EMU 6  EMU 7  EMU 8
AU “pick up the ball”

EMU 9  EMU 10
AU “throw the ball”

EMU 11  EMU 12
AU “pick up the ball”

EMU 13  EMU 14
AU “throw the ball”

EMU 15
AU “stand”
An Integrated
Semantic-Syntactic
Video Description Model
Integrated Semantic-Syntactic Model
Contributions

- **Integrated semantic-syntactic model:** Enables efficient mixed-level query processing; “Find all penalty kicks shot to the left of the goalie”

- **Actor entity:** Enables incorporation of context in object-event relationships

- **Model-based query processing:** Graph/tree matching; e.g., “Find all clips similar to this one”
Abstract Event Model
Query Processing

Database

Description
(XML Document)

XML parser

Index Generator

Index

Matching

Results

Query Formation

Example/Abstract Description

User

Query Graph
User Interface
Video Processing: Trajectory Estimation
Frame-Based Video Summarization and Shot Classification
Framework

- Shot Boundary Detection
- Extraction of Low-Level Shot Features
  - GoF Color
  - Spatial Layout
  - Motion Activity
  - Shot Duration, ...
- Key Frame Extraction - for visual summarization
- Fuzzy Clustering to Generate Domain Models
- Analyze New Content in the Domain using the Generated Domain Model (similar to VQ)
GoF Color Representation

- Common approach: Describe visual and color content of a shot with key frames and key frame histograms, respectively.
  - The provided color description may vary significantly with key frame selection criterion
- Alternative: Consider color content of all frames for representative histogram computation
  - Robust color histogram descriptors that are unaffected by outlier frames due to camera movement, occlusion, text/graphics overlays, brightness variations, etc.
Alpha-Trimmed Average Histograms
Key Frame Selection

For each GoF \( l \), compute

\[
E_i = \| H_i - H_l \| \quad i = 1, \ldots, N.
\]

The frame \( r \) that minimizes \( E \) is the key frame.
News Unit Generation
Location-Based Browsing Using Establishing Shots