The IBM Shot Boundary Detection System at TRECVID 2003

Arnon Amir

Shot Boundary Detection

- Temporal segmentation of the video into shots
- Detect location and duration of all shot transitions of all kinds:
  - cuts, dissolves, fades, wipes, checkerboard, etc.
- Desired: accurate detection, frame-accurate location
- Challenges: It would be so much easier if there were no
  - Fast object or camera motion
  - Fast illumination changes: pop concerts, driving, backlight, ...
  - Fire, flags in the wind, sea waves, ...
  - Specularities, shadows, reflections from glass, water, ...
  - Instantaneous illumination changes due to flash photography
  - Very short shots (up to single-frame “shots” in the Search test set)
  - Very long gradual transitions
  - Text overlay, graphics, animation
  - Screen split, video in video
  - Video artifacts: MPEG errors, compression noise, camera noise, ...
IBM SBD: System Overview

- Work on fully-decoded RGB video frames
- A single pass over the video
- Graph-based, multiple pair-wise frames comparisons
- Adaptive thresholds to determine significant changes
- Finite state machine processes the video frame by frame
  - Rule-based state to state transitions
- Aids:
  - Temporal rank filtering whenever applicable
  - Detectors for various simple “events” and video artifacts
- Dedicated monitoring and debugging tools

Graph-Based, Multiple Pair-Wise Frames Comparisons

Each frame is a node in a graph
Pairs of frames, up to 13 frames apart, are connected with arcs
A shot transition is like a cut in the graph

Motivation: more pairs - a better decision [Amir & Lindenbaum, 1995]
Frames Representation and Comparison

- 3D RGB Color histogram \( H(r, g, b) \)
  - 3 bits per color channel
  - Total 512 bins
  - Histogram comparison: \( L_i(H_i, H_j) \)

- 3D Localized Edges direction histogram \( H(x, y, |\nabla I|) \)
  - Frame is divided to 8x8 blocks (6 bits)
  - Gradient magnitude (3 bits)
  - Total 512 bins
  - Histogram comparison: \( L_i(H_i, H_j) \)

- Gray-level Thumbnails comparison

- Average frame luminance

- Black detector, monochrome detector

- Non linear, state-based fusion

Adaptive Thresholds

- Observe:
  - Frames similarity changes significantly across different shots and videos
  - Frames dissimilarity increases as the frames are wider apart of each other

- Adaptive thresholds are required

- Collect statistics of frames similarity in a 61-frames buffer around the processed frame
  - Assumption: at least 25% of the buffer content is within-shot
  - Use rank filtering to determine the threshold: \( C \cdot (\text{Buff @ 25%}) \)
  - Apply a hysteresis when change between shot and transition states. \( C_{\text{within shot}} > C_{\text{within transition}} \)
Example: Adaptive Thresholds

Threshold 5
\( \text{Diff}(F_i, F_{i-5}) \)

Threshold 3
\( \text{Diff}(F_i, F_{i-3}) \)

Threshold 1
\( \text{Diff}(F_i, F_{i-1}) \)

Ground truth

System output

SBD - Processing Schedule

Scene Cuts Detector Scheduling

<table>
<thead>
<tr>
<th>Input</th>
<th>Output</th>
<th>Buffer</th>
<th>Buffer</th>
<th>Buffer</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Video</td>
<td>Scenes</td>
<td>Frames</td>
<td>Memory</td>
<td>Buffer</td>
</tr>
<tr>
<td>Buffer</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Memory</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Buffer</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Development Methodology

- Apply to development data with ground truth
- Analyze causes for errors, classify them
- Address one class of errors at a time:
  - Use a visual monitoring tool to analyze system behavior at errors
  - Modify the finite state machine - modify/add rules and states
  - Add new detectors
- Test again
Example: Analyzer View of 19980203_CNN.MPG

matlab_example_edit_subframe2.avi
Example: Analyzer View of 19980203_CNN.MPG

matlab_example_edit_subframe2.avi

Example – Dissolves Misclassified as Cuts

Ground truth: Dissolve
System: Dissolve

In shot

10 frames

Start Dissolve

Ground truth: Fade-out-in
System: cut

Start Dissolve
Shot Boundary Detection - System Improvements 2003

- Baseline system is the IBM CueVideo SBD best run at TREC 2002
- Algorithmic improvements for TREC 2003
  - Flash photography detector. Eliminates false detections on rapid flashlights. 19980426_CNN.mpg, 19980422_CNN.MPG
  - Better handling of fades (better detection, linking fade out-in, abrupt fades)
  - Improved detection of graduals’ boundaries (improves gradual accuracy)
  - Detection & handling of partial-frame MPEG errors (reduces insertion errors)

Results on SBD Development Set

- Changes were tested on a development set with manually labeled frame-accurate ground truth: 10 video segments, 5 minutes each. Total of 363 cuts, 145 graduals.

<table>
<thead>
<tr>
<th>10 Runs on development set</th>
<th>All Transitions</th>
<th>Cuts</th>
<th>Gradual transitions</th>
<th>Graduals accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>alm1</td>
<td>0.951</td>
<td>0.881</td>
<td>0.99</td>
<td>0.94</td>
</tr>
<tr>
<td>N1047</td>
<td>0.947</td>
<td>0.887</td>
<td>0.98</td>
<td>0.94</td>
</tr>
<tr>
<td>N110</td>
<td>0.950</td>
<td>0.905</td>
<td>0.98</td>
<td>0.95</td>
</tr>
<tr>
<td>N119</td>
<td>0.941</td>
<td>0.938</td>
<td>0.98</td>
<td>0.98</td>
</tr>
<tr>
<td>N120</td>
<td>0.944</td>
<td>0.938</td>
<td>0.98</td>
<td>0.98</td>
</tr>
<tr>
<td>N122</td>
<td>0.944</td>
<td>0.938</td>
<td>0.98</td>
<td>0.98</td>
</tr>
<tr>
<td>N123</td>
<td>0.950</td>
<td>0.930</td>
<td>0.98</td>
<td>0.98</td>
</tr>
<tr>
<td>N126</td>
<td>0.941</td>
<td>0.940</td>
<td>0.98</td>
<td>0.97</td>
</tr>
<tr>
<td>N127</td>
<td>0.941</td>
<td>0.947</td>
<td>0.98</td>
<td>0.98</td>
</tr>
<tr>
<td>N128</td>
<td>0.944</td>
<td>0.941</td>
<td>0.98</td>
<td>0.98</td>
</tr>
</tbody>
</table>
NIST Results on the Test Set

- Trends are similar to the results on the Development set
- Absolute values are typically 0 to 0.06 lower
- Cuts – one frame shift

10 Runs on NIST Test Set

<table>
<thead>
<tr>
<th>System</th>
<th>All Trans.</th>
<th>Cuts</th>
<th>Graduals</th>
<th>Graduals Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>2001 (alm1)</td>
<td>0.915</td>
<td>0.916</td>
<td>0.793</td>
<td>0.793</td>
</tr>
<tr>
<td>2002 (N047)</td>
<td>0.895</td>
<td>0.944</td>
<td>0.778</td>
<td>0.784</td>
</tr>
<tr>
<td>2003 (N110)</td>
<td>0.893</td>
<td>0.936</td>
<td>0.784</td>
<td>0.784</td>
</tr>
<tr>
<td>2004 (N119)</td>
<td>0.897</td>
<td>0.936</td>
<td>0.792</td>
<td>0.792</td>
</tr>
<tr>
<td>2005 (N120)</td>
<td>0.897</td>
<td>0.938</td>
<td>0.802</td>
<td>0.802</td>
</tr>
<tr>
<td>2006 (N122)</td>
<td>0.897</td>
<td>0.938</td>
<td>0.802</td>
<td>0.802</td>
</tr>
<tr>
<td>2007 (N123)</td>
<td>0.898</td>
<td>0.936</td>
<td>0.809</td>
<td>0.809</td>
</tr>
<tr>
<td>2008 (N126)</td>
<td>0.896</td>
<td>0.932</td>
<td>0.776</td>
<td>0.776</td>
</tr>
<tr>
<td>2009 (N127)</td>
<td>0.892</td>
<td>0.937</td>
<td>0.784</td>
<td>0.784</td>
</tr>
<tr>
<td>2010 (N128)</td>
<td>0.894</td>
<td>0.935</td>
<td>0.798</td>
<td>0.798</td>
</tr>
</tbody>
</table>

Shot Boundary Detection - System Results 2001-2003

System Errors 2001-2003

<table>
<thead>
<tr>
<th>System</th>
<th>All Trans.</th>
<th>Cuts</th>
<th>Graduals</th>
<th>Gradual Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>2001 (alm1)</td>
<td>10.86%</td>
<td>6.88%</td>
<td>20.14%</td>
<td>27.34%</td>
</tr>
<tr>
<td>2002 (N047)</td>
<td>11.46%</td>
<td>7.85%</td>
<td>20.73%</td>
<td>22.50%</td>
</tr>
<tr>
<td>2003 (N127)</td>
<td>9.18%</td>
<td>5.80%</td>
<td>17.75%</td>
<td>16.40%</td>
</tr>
</tbody>
</table>

Rel. Improvement 19.9% 26.1% 14.4% 27.1%
All top seven runs in all measures are IBM.


Four of the ten IBM runs are highlighted:

- Systematic work with an SBD development set

Summary

- Multiple pairs of frames
- Finite State Machine
- Frame comparison: RGB, Edges, Thumbnails
- Adaptive thresholds
- System monitoring tools
- Systematic work with an SBD development set
Detection of All Transitions

Detection of Cuts
Detection - Gradual Changes

Gradual Changes – Frame Accuracy