Structuralizing Educational Videos Based on Presentation Content

Chitra Dorai (IBM T. J. Watson Research Center)
Vincent Oria (New Jersey Inst. of Technology)
Viswanath Neelavalli (New Jersey Inst. of Technology)
Motivation and Approach

**Motivation**
- A need to index and query educational videos on their teaching content
- A legacy of educational videos in higher education institutions

**Approach**
- Identify slides or parts of slides in the educational video (the purpose of this paper)
- Match the detected slides in the video with slides used during the presentation (future work)
- Generate semantics based some pre-knowledge (future work)
Structure Induced by Material Presented in Educational Videos

Educational Video

Non-Narrator (text) Frame Sequence

Slide Test Frames

Pure Slide Text Frames (text close-up)

Web (page) Text Frames

Slide Text Mixed (with instructor, whiteboard)

Whiteboard Frames

Other (slide text with video playing)
Educational Video Frame Samples

Slide Text

Web Page

Narrator

Whiteboard
Structuralizing Educational Videos

- First level segmentation (distinguish between narrator and non-narrator sections) based on color
  - Inter-frame color-histogram
  - Color moments
- Second level segmentation (finer classification of non-narrator sections) based on texture
  - Horizontal Texture Energy
  - Vertical Texture Energy
First Level Segmentation:

Color Histogram

- 4 bin Color Histogram for R, G and B, computed for each frame
- Inter-frame distance between consecutive frames
Color Histogram Analysis

- Points of transition between sections containing text from those depicting narrator
- Points of interest: change in type of content, transition from frames showing narrator to frames showing gathering
- Extremely sensitive to camera jitter, contrast adjustment and brightness variations in the video
- Need for advanced analysis
First Level Segmentation

Color Moments

- Color Moments based classification using a machine learning algorithm, Decision Tree (c4.5)
- First 10 central moments computed are from the three normalized color histograms for R, G and B
- 30 valued feature vector for each frame

\[
m_1 = \frac{\sum (i \cdot H(i))}{\sum (H(i))}
\]
\[
m_p = \frac{\left\{ H(i) \cdot (i - m_1)^p \right\}}{\sum (H(i))} \quad 10 \geq p \geq 2
\]
# First Level Classification: Confusion Matrices

<table>
<thead>
<tr>
<th>Video</th>
<th>Training and Test Samples</th>
<th>Misclassification Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>eseminar.mpg</td>
<td>15554</td>
<td>Nearly 0%</td>
</tr>
<tr>
<td>universaldl.mpg</td>
<td>14065</td>
<td>6.2%</td>
</tr>
<tr>
<td>eseminar.mpg</td>
<td>Narrator: 1603 Non-Narrator(text): 0</td>
<td>1</td>
</tr>
<tr>
<td>Narrator</td>
<td></td>
<td>13950</td>
</tr>
<tr>
<td>Non-narrator (text)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>universaldl.mpg</td>
<td>Narrator: 5456 Non-Narrator(text): 6</td>
<td>861</td>
</tr>
<tr>
<td>Narrator</td>
<td></td>
<td>7742</td>
</tr>
<tr>
<td>Non-Narrator</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
First Level Classification Results

- First level classification results are good.
- Non-Narrator (text) Frames in *universaldl.mpg* contained predominantly text along with occasional presence of narrator in a small window.
- Errors in Narrator class of *universaldl.mpg* video are observed due to the presence of frames depicting a gathering of people, including narrator.
Second Level Classification: Non-narrative Sections

- Extraction of texture features for the various classes of text appearing in the video.
- Texture feature vectors (2640 values)
  - using features derived from the DCT coefficients of each frame and input to the Decision Tree algorithm

\[
E_{\text{hor}}(i, j) = \sum_{1 \leq i \leq 7} |C_{o}(i, j)|
\]

Horizontal Energy

\[
E_{\text{ver}}(i, j) = \sum_{1 \leq j \leq 7} |C_{o}(i, j)|
\]

Vertical Energy
## Second Level Classification Results

<table>
<thead>
<tr>
<th>Video</th>
<th>Samples</th>
<th>Misclassification Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>comclass.mpg</td>
<td>3000</td>
<td>4.9%</td>
</tr>
<tr>
<td>eseminar.mpg</td>
<td>140</td>
<td>1.4%</td>
</tr>
</tbody>
</table>

### Confusion Matrix

<table>
<thead>
<tr>
<th></th>
<th>Slide Text</th>
<th>Whiteboard Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>communicationclass.mpg</td>
<td>1956</td>
<td>301</td>
</tr>
<tr>
<td>Slide Text</td>
<td>118</td>
<td>896</td>
</tr>
<tr>
<td>Whiteboard Text</td>
<td>68</td>
<td>2</td>
</tr>
<tr>
<td>eseminar.mpg</td>
<td>0</td>
<td>70</td>
</tr>
<tr>
<td>Slide Text</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Web Text</td>
<td>70</td>
<td></td>
</tr>
</tbody>
</table>
Conclusion and Future Work

- Fine-grained classification results are *satisfactory*.
- Experiments are underway to identify the learning content of the slides and generate semantic annotation.
- Further Work:
  - Audio track processing
  - Flexible content browsing and management of educational video