Video scene description: an audio based approach

Hadi Harb
Outline

- Expectations from audio
- Technical difficulties
- Our approach for audio classification
- The audio browser
- conclusion
Introduction

Video needs:
- Searching
- Retrieving
- Describing
Audio expectations

- What can we expect from audio?
  - Action
  - Dialog calm
  - Dialog action
  - happiness
  - Sadness
Audio expectations

- Noise $\rightarrow$ action
- Speech $\rightarrow$ dialog
- Speech+Noise+Music $\rightarrow$ active dialog
- Music $\rightarrow$ calm/sad
- ...

…
Technical difficulties

- Audio signal:

- This representation contains no semantics
Variability
Classes similarity
Audio classification

- Pattern recognition
  - Feature extraction
  - classifier
Current state-of-the-art

- MFCC
- GMM or \( k \)-NN
  
  + easy implementation
  
  -- need for large amount of learning data
State of the art systems (France)

- SECK (IRISA) Speech/Music:
  - Test: 35 min, documentary
  - Learning: 55 min
  - Results: ~88%

- This is no good for general audio!!
Features

- Long Term (2s) features:
  - SCR
  - FT
  - DBAI
  - FCV
SCR-feature

Speech

Music

Noise
DBAI-feature

Speech

Music

Noise
Features histograms

- SCR histogram for 400s of speech
- FCV histogram for 400s of noise
- DBI histogram for 400s of music
Classifier-Neural Network
Experimental setup

- Audio Classification in “film” context
  - Classes are more diversified
  - Training != Testing

- Training: CNN, Songs, ERAZOR (30min)
- Testing: GLADIATOR (30min)
Past-Present-Future

Convergence time of NN

Context Blocks in Feature Vector
### Classification results

<table>
<thead>
<tr>
<th>Time in min</th>
<th>Time classified</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Music</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>83%</td>
</tr>
<tr>
<td><strong>Speech</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>10.4</td>
<td>87%</td>
</tr>
<tr>
<td><strong>Noise</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>11.2</td>
<td>93%</td>
</tr>
</tbody>
</table>
Description « Films »

Movies (7.5 h):

- Central do Brasil
- Matrix
- Forsaken
- Gladiator
- Bridget Jones

- Speech
- Music
- Noise
Description “TV materials”

- TV material (2.5 h):

  - News
  - Program + Music theme
  - Talk + commercials
  - Talk
  - Video Clips
  - Jazz
  - Speech
  - Music
  - Noise
Description “Scenes”

- Scenes GLADIATOR:
  - Action Scene
  - Music Scene
  - Music+Action Scene
Similarity/search

- Scene fingerprints:
Navigation experiments

Scene1          Scene2          Scene3     …   Scene9         Scene10
The Audio browser

- Something with colors that indicates how to navigate
The audio browser 2

- Demo
Conclusion

- More audio classes are required
- Our unpublished technique for audio classification works better
- Better if combined with visual information