Generic Architecture for Event Detection in Broadcast Sports Video

Chris Poppe, Sarah De Bruyne,
Rik Van de Walle
Multimedia Lab – Ghent University - IBBT
Context

• Broadcast sports video

• Automatic annotation
  • summarization
  • retrieval
  • adaptation (e.g., for mobile devices)
  • statistical analysis (coaching)
Context

• Generic!
  • soccer
  • tennis
  • football
  • cricket
  • hockey
  • …
• common aspects: broadcast video, edited, game rules
• Generic = easily configurable 😊
System

Low-level
- RGB-Color histograms
- Cumulative HS-Color Histogram
- MV
- MFCC
- ZCR

Mid-level
- Shot Detection
- Shot Info
- Color
- Motion

High-level
- Replay Detection
- Video Classification
- Audio Classification

Event detection
- State transitions
Low-level

- Video
  - Motion Vectors
  - RGB histogram
  - Cumulative HS(V)-histogram
    - determines dominant color in video
- Audio
  - MARSYAS (Music Analysis, Retrieval and Synthesis for Audio Signals) an open source software framework for audio processing
    - Mel-Frequency Cepstral Coefficients (MFCC)
    - Zero Crossings (ZCRS)
Mid-level

- Rely on low-level or mid-level modules
- Create features
  - information compression
Shot Detection

- Evaluate differences of histograms
- Local search window to find maxima
- Morphological processing for gradual transitions

Gradual transition

Abrupt transition
Shot detection results

Precision and recall values for the detection of shot transitions

<table>
<thead>
<tr>
<th></th>
<th>Precision (%)</th>
<th>Recall (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Roland Garros</td>
<td>99.44</td>
<td>98.90</td>
</tr>
<tr>
<td>Manchester United</td>
<td>93.86</td>
<td>100</td>
</tr>
</tbody>
</table>
Color Module

- Cummulative histogram for detection of playing field
- Back projection

- Gaussian smoothing
Color Module

- Calculate ratio of field pixels for each vertical line
  - ColorMax
  - ColorStDev
Motion Module

- MV Space
  - represent MVs as HSV values according to size and angle
  - Cluster (image processing)

- MotionEntropy, HorizontalMotion, VerticalMotion, DiagonalMotion
High-level

- Add semantic meaning based on Mid-level features
  - famous semantic gap
- Replay detection
  - assume gradual transitions correspond to replay
- Classification

![Diagram of multimodal processing pipeline](image-url)
**Video Classification**

- Labeled training set

<table>
<thead>
<tr>
<th>Video labels for soccer sequences</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Audience</td>
<td>Reaction of audience</td>
</tr>
<tr>
<td>Field</td>
<td>Overview of actions</td>
</tr>
<tr>
<td>Goal</td>
<td>Field view with goal</td>
</tr>
<tr>
<td>Close-up of player</td>
<td>No action</td>
</tr>
<tr>
<td>Tracking of player</td>
<td>Medium view of action</td>
</tr>
<tr>
<td>Medium image of player</td>
<td>Before or after an action</td>
</tr>
</tbody>
</table>
SVM Classification

1. Train SVM
2. Input feature vector
3. Output label

“Field”

“Tracking of Player”

ColorMax, MotionEntropy, …

SVM

“Field”

ColorMax, MotionEntropy, …
## Classification results

### Classification results for soccer (in %).

<table>
<thead>
<tr>
<th>SVM</th>
<th>Precision (%)</th>
<th>Recall (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Audience</td>
<td>100</td>
<td>30.43</td>
</tr>
<tr>
<td>Field</td>
<td>75.32</td>
<td>99.58</td>
</tr>
<tr>
<td>Goal</td>
<td>89.47</td>
<td>32.08</td>
</tr>
<tr>
<td>Close-up of player</td>
<td>90.03</td>
<td>92.18</td>
</tr>
<tr>
<td>Tracking of player</td>
<td>89.15</td>
<td>78.23</td>
</tr>
<tr>
<td>Medium image of player</td>
<td>61.76</td>
<td>42.00</td>
</tr>
<tr>
<td>Replay</td>
<td>88.99</td>
<td>97.00</td>
</tr>
</tbody>
</table>
### Audio Classification

- Labeled training set
- Feature vectors \((MFCC, ZCRS)\)
- Audio samples

<table>
<thead>
<tr>
<th>SVM</th>
<th>Precision (%)</th>
<th>Recall (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Whistle</td>
<td>95.00</td>
<td>97.44</td>
</tr>
<tr>
<td>Loud comments</td>
<td>96.00</td>
<td>72.73</td>
</tr>
<tr>
<td>Comments</td>
<td>77.78</td>
<td>87.50</td>
</tr>
<tr>
<td>Sheering</td>
<td>83.33</td>
<td>95.24</td>
</tr>
<tr>
<td>Audience</td>
<td>90.45</td>
<td>95.30</td>
</tr>
</tbody>
</table>
Event detection

- Detect events using cinematographic principles
  - Broadcast sports video is composed by director
  - Game-rules can also be used
- Search for patterns of video shots and combine with audio
  - Finite State Machine (FSM)
    - models an event (e.g., goal in soccer or ace in tennis)
Event description

- **Goal in Soccer**

![Diagram of Event Description]

- S0
  - <Loud, Field>
  - <Loud, Close-up>
  - <Sheer, Close-up>

- S1
  - <noWhistle, Field>
  - <noWhistle, Medium>
  - <Comment, Close-up>
  - <Audience, Close-up>
  - <noWhistle, Audience>

- S2
  - <noWhistle, Audience>
  - <Sheer, *>
  - <noWhistle, Close-up>
  - <noWhistle, Medium>
  - <noWhistle, Audience>

- S3
  - <Whistle, *>, <*, Replay>
  - <Audience, Medium>
  - <Loud, Medium>
  - <Comment, Medium>
  - <Comment, Close-up>
  - <Loud, Close-up>

- S4
Results

Detection results for tennis points and soccer goals (in %)

<table>
<thead>
<tr>
<th>Tennis Point</th>
<th>Precision (%)</th>
<th>Recall (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wimbledon</td>
<td>82.14</td>
<td>95.83</td>
</tr>
<tr>
<td>Roland Garros</td>
<td>84.44</td>
<td>95.00</td>
</tr>
<tr>
<td>US Open</td>
<td>97.67</td>
<td>85.71</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Goal</th>
<th>Precision (%)</th>
<th>Recall (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manchester United</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Marseille</td>
<td>100</td>
<td>66.67</td>
</tr>
</tbody>
</table>
Conclusions

• System to detect events in broadcast sports video
• Generic (configurable 😊)
  • Training data (SVM)
  • Event description (FSM)
• Future work
  • semi-automatic selection of lower-level modules
  • new events/ new modules/ new sports
Questions?

Low-level
- ViolaJones Features
- Audio sampling

Mid-level
- Face detection
- Lip detection
- Speech Transcription

High-level
- Speaker Detection
- Question Detection

Event detection
- Answer Prediction