CS365A - ARTIFICIAL INTELLIGENCE

Project Proposal

Automatic Highlights Extraction in Cricket

Anjani Kumar(11101)
Sumedh Masulkar(11736)

Guided By:
Dr. Amitabha Mukerjee

Aim

- Extracting highlights automatically from a sports video using audio and video features.

Related Works

- Highlights extraction using Hidden Markov Models(HMM) in [1][2][3].
    - ❑ The states and transitions in the game were represented using HMM.
- [3] fused in audio information along with motion information for the first time.

Related Works (2)

- In [4], the author proposed an unsupervised event discovery and detection framework which used color histograms(CH) or histograms of oriented gradients(HOG).
- [5] extracted event sequences from videos and classifies them into a concept using sequential association mining.

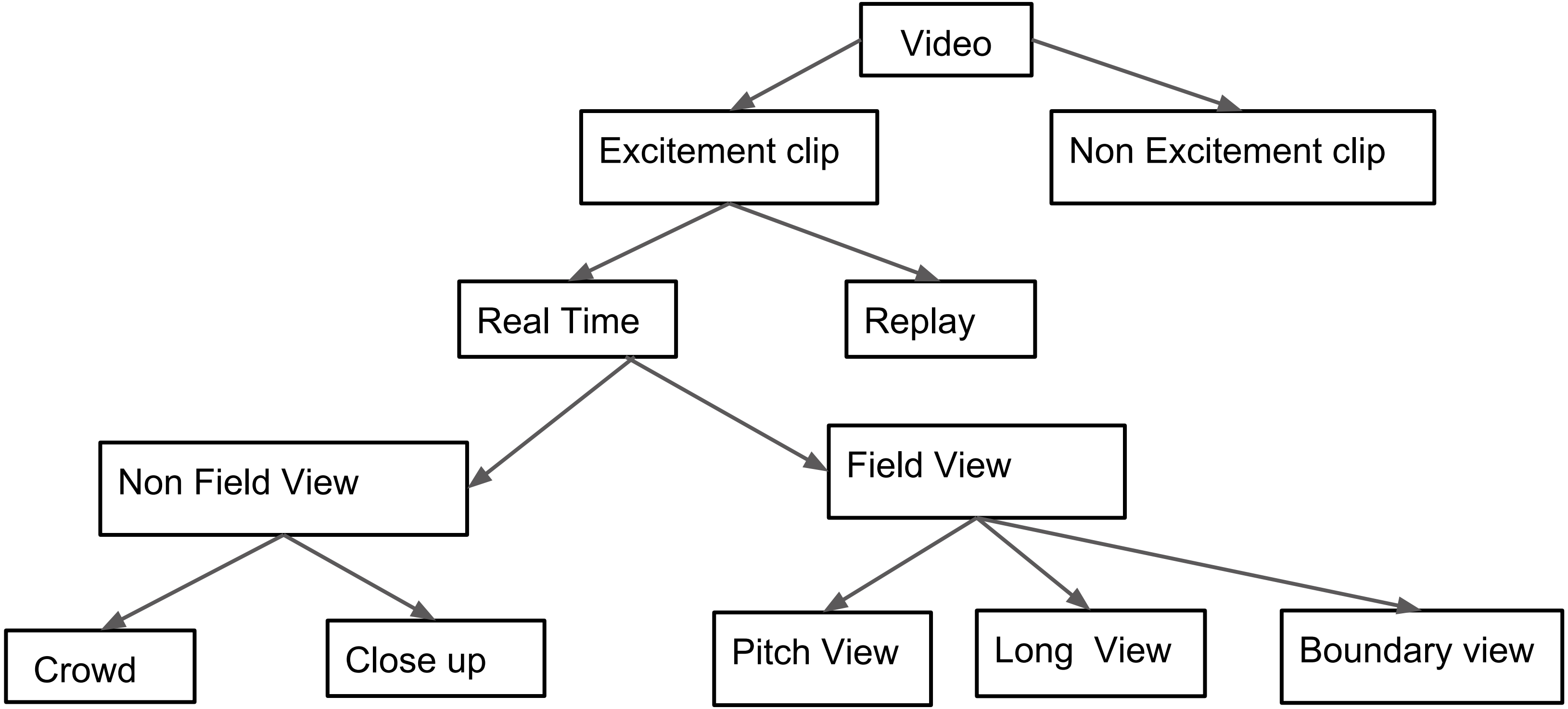
Related Works (3)

- [6] introduced a hierarchical framework for events detection and classification without shot detection and clustering.
    - ❏ We will be primarily following approach of [6] in our project.
    - ❏ [6] was an improved version of [5].
- [7] used text commentary processing and shot detection techniques.

Approach

- Divide the extraction process into multiple levels.
- Remove the uninteresting event sequences from the main video at each level.
- 5 levels of extraction for shot classification (pitch view, crowd view, field view etc.)

Hierarchical Framework

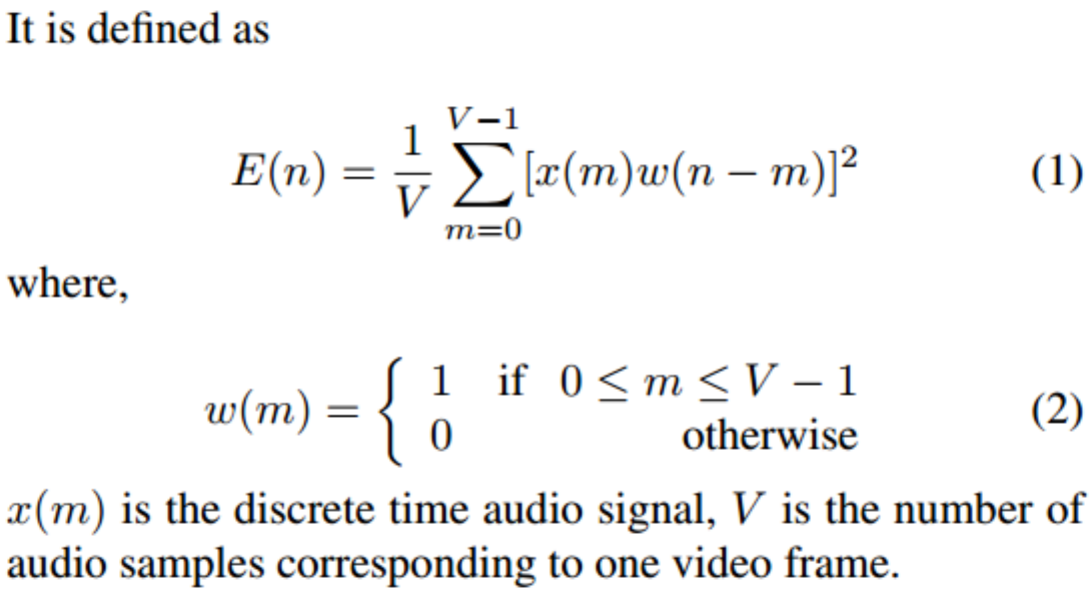


Level - I

- Excitement Detection
  - ❑ Spectator’s cheer and commentator’s speech analysis.
  - ❑ Two popular content analysis techniques - Short-time audio energy(*E*) and Short-time Zero Crossing Rate(*Z*).
  - ❑ If *E* * *Z* is greater than a given threshold, the particular frame is an excitation frame.

Level - I (2)

- Short-time audio energy



It is defined as

$$E(n) = \frac{1}{V} \sum_{m=0}^{V-1} [x(m)w(n-m)]^2 \qquad (1)$$

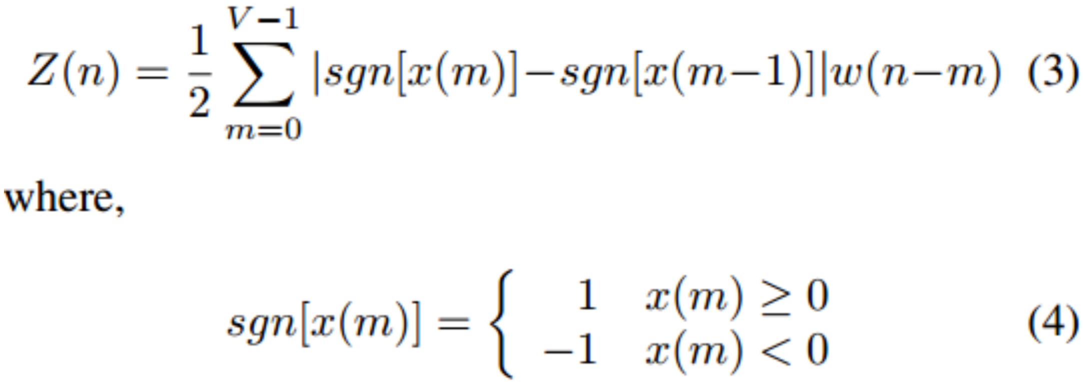
where,

$$w(m) = \begin{cases} 1 & \text{if} \quad 0 \leq m \leq V-1 \\ 0 & \text{otherwise} \end{cases} \qquad (2)$$

$x(m)$ is the discrete time audio signal, V is the number of audio samples corresponding to one video frame.

Level - I (3)

- Short-time zero-crossing rate



$$Z(n) = \frac{1}{2}\sum_{m=0}^{V-1} |sgn[x(m)] - sgn[x(m-1)]| w(n-m) \quad (3)$$

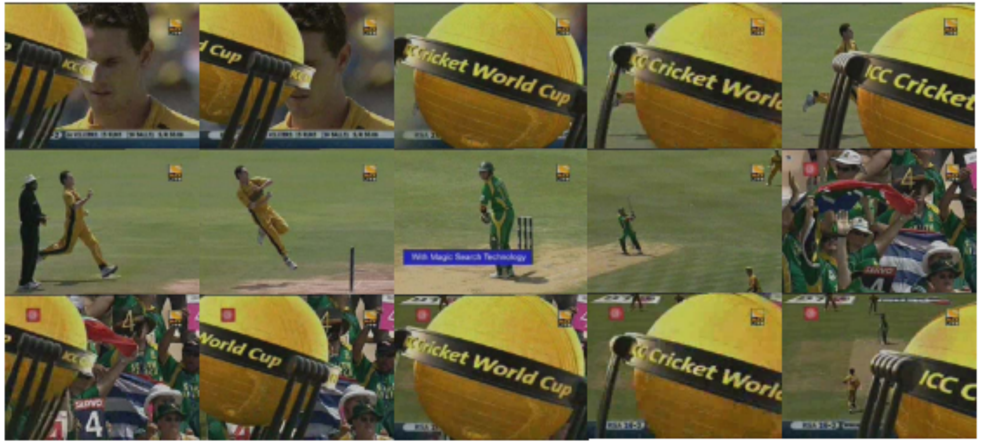
where,

$$sgn[x(m)] = \begin{cases} 1 & x(m) \geq 0 \\ -1 & x(m) < 0 \end{cases} \quad (4)$$

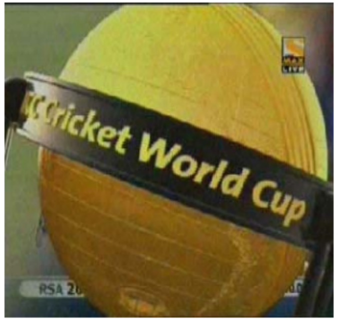
where w(m) is a rectangular window.

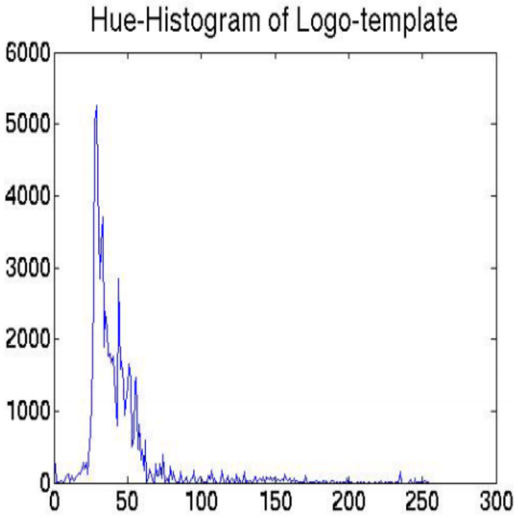
Level - II

- Replay Detection
  - ❑ A replay is sandwiched between two logo transitions and the score bar is removed.



Level - II (2)





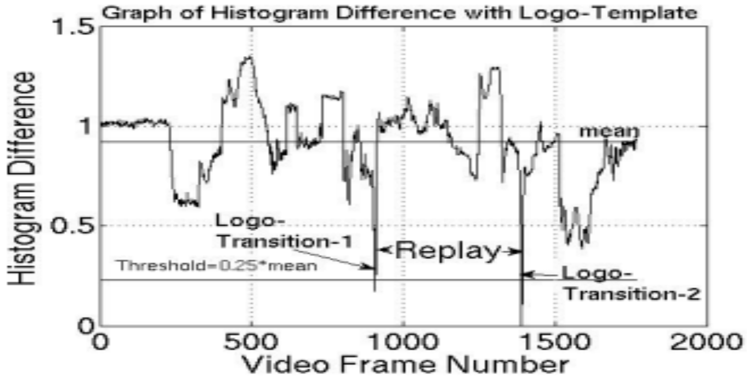


Figure 4. Graph of Hue-histogram difference vs video frame number for the video containing replay segment shown in figure 2

Level - III

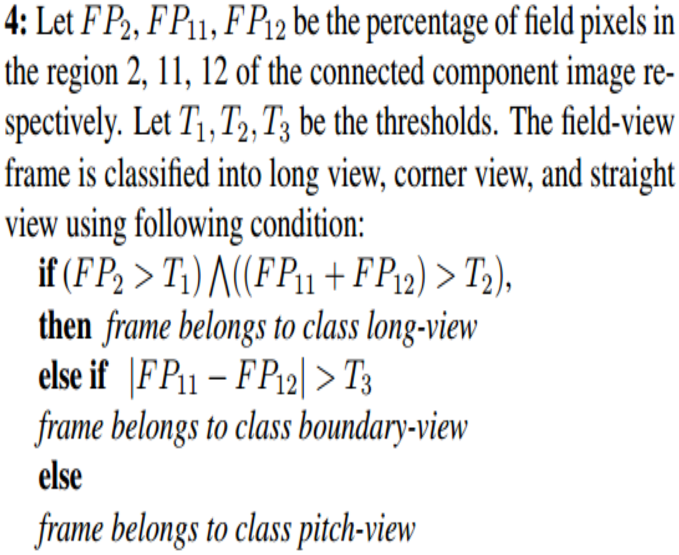
- Field view detection
  - ❑ Dominant Grass Pixel Ratio(DGPR) is used to classify frames.
  - ❑ DGPR = (x_g/x) where x_g is number of pixels of grass, and x is total number of pixels.
  - ❑ For field view, DGPR values is greater than 0.07 whereas DGPR is smaller for non-field views.

Level - IV

- 4a - Field view classification
  - ❑ Classified as pitch view, long view or boundary view.
  - ❑ Introduces the concept of *flux tensor* - temporal variations of the optical flow field within the local 3D spatiotemporal volume.
  - ❑ Percentage of field pixels used to differentiate between views.

Level - IV (2)

- 4a



4: Let FP_2, FP_{11}, FP_{12} be the percentage of field pixels in the region 2, 11, 12 of the connected component image respectively. Let T_1, T_2, T_3 be the thresholds. The field-view frame is classified into long view, corner view, and straight view using following condition:

```
if $(FP_2 > T_1) \bigwedge ((FP_{11} + FP_{12}) > T_2)$,
then *frame belongs to class long-view*
else if $|FP_{11} - FP_{12}| > T_3$
*frame belongs to class boundary-view*
else
*frame belongs to class pitch-view*
```

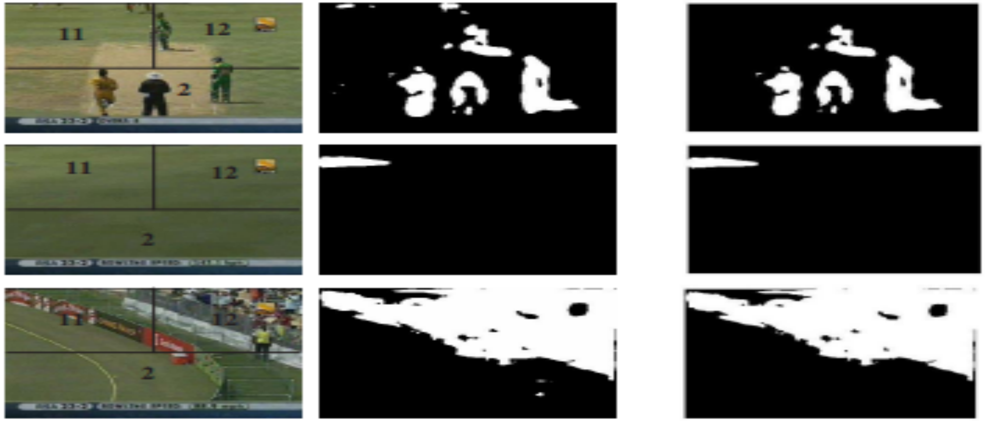
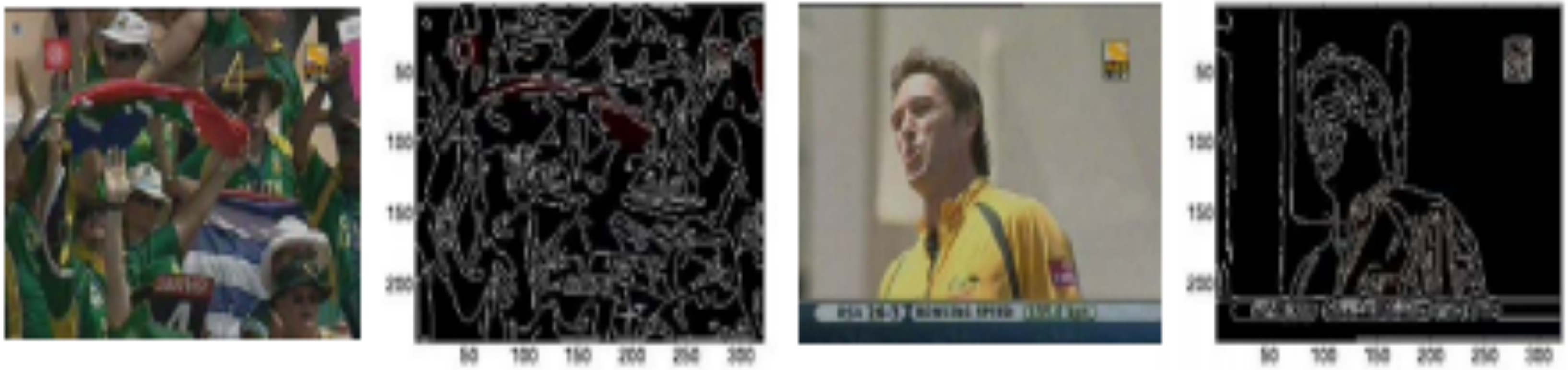


Figure 7. Row-1 shows pitch view: (a) Image (b) motion-mask (c) connected component image, Row-2 shows long view: (d) Image (e) motion-mask (f) connected component image, Row-3 shows boundary view: (g) Image (h) motion-mask (i) connected component image

Level - IV

- 4b - Close Up view
  - ❑ RGB image is converted to YC_bC_r.
  - ❑ Percentage of edge pixels(EP) are calculated using *Canny* operator.
  - ❑ A threshold for EP classifies frames as close up view or crowd view.

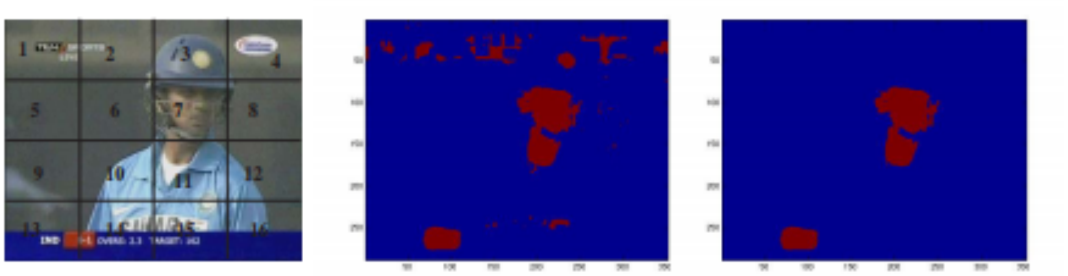
Level - IV (2)



- Percentage of Edge pixels greater for crowd view.

Level - V

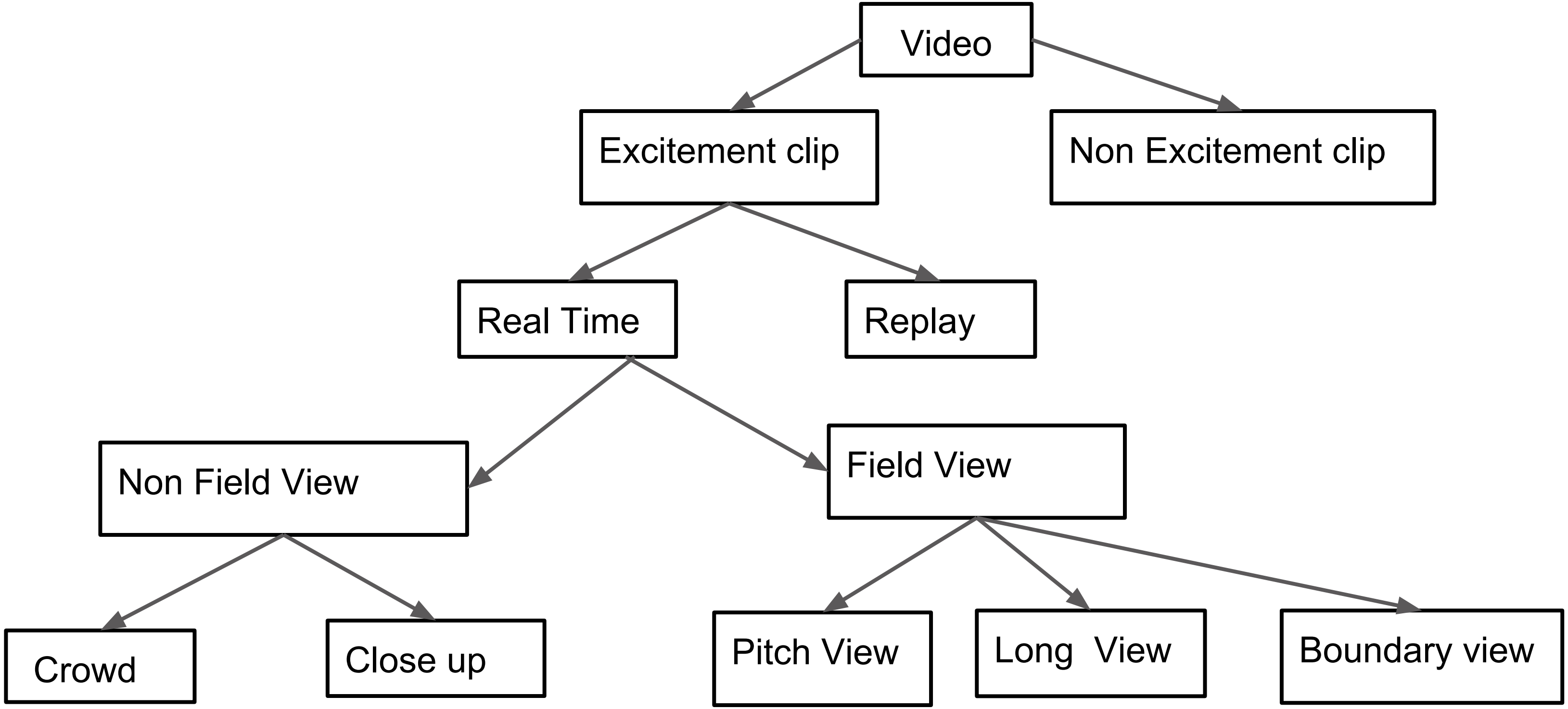
- 5a - Close up classification
- Detection of skin color by converting RGB image to YC_bC_r.



Level - V

- 5b - Crowd classification into spectators or fielders gathering.
- Fielders usually gather after an interesting event and have field as background, which should be kept in highlights.

Hierarchical Framework



References

[1] Kamesh Namuduri. “Automatic extraction of highlights from a cricket video using MPEG-7 descriptors”.

[2] Jinjun Wang, Changsheng Xu, Engsiong Chng, Qi Tian. “Sports Highlight Detection from Keyword Sequences Using HMM”, in Proceedings of the International Conference on Multimedia and Expo, 2004.

[3] Chih-Cheih Cheng, Chiou-Ting Hsu. “Fusion of Audio and Motion Infromation on HMM-Based Highlight Extraction for Baseball Games”, in Proceedings of the IEEE Transactions on Multimedia, vol. 8, no. 3, June 2006.

[4] Hao Tang, Vivek Kwatra, Mehmet Emre Sargin, Ullas Gargi. “Detecting Highlights in Sports Videos: Cricket as a test case”, 2011.

[5] Maheshkumar H. Kolekar, Somnath Sengupta. “Semantic concept mining in cricket videos for automated highlight generation”, 2009.

References

[6] M. H. Kolekar, K. Palaniappan, S. Sengupta. “Semantic Event Detection and Classification in Cricket Video Sequence”, in Proceedings of the Indian Conference on Computer Vision, Graphics & Image Processing, 2008.

[7] Dipen Rughwani. “Shot Classification and Semantic Query Processing on Broadcast Cricket Videos”. http://cse.iitk.ac.in/~vision/dipen/.

THANK YOU!!

QUESTIONS?