# Highlights Extraction from Unscripted Video

T-61.6030, Multimedia Retrieval
Seminar presentation 04.04.2008

Harrison Mfula
Helsinki University of Technology
Department of Computer Science,
Espoo, Finland

# Contents

1. Introduction
2. Audio marker recognition
3. Visual marker detection
4. Finer - resolution highlights extraction
5. Conclusion

# Motivation

Few real applications of multimedia retrieval have been accepted by the general public so far. Is sports highlights extraction, medical database retrieval, or web multimedia search engine going to be the next killer application? It remains to be seen. With no clear answer to this question, it is still a challenge do research and implement applications that are appropriate to real life in multimedia retrieval.

# Objective

Highlights extraction from unscripted content such as sports video.

Example applications –

Soccer, baseball, golf

# Introduction

An approach for highlights extraction for unscripted content such as sports video is described. The effectiveness of this algorithm is shown in three different sports: soccer, baseball and golf. The framework consists of four main parts summarized in the diagram below.

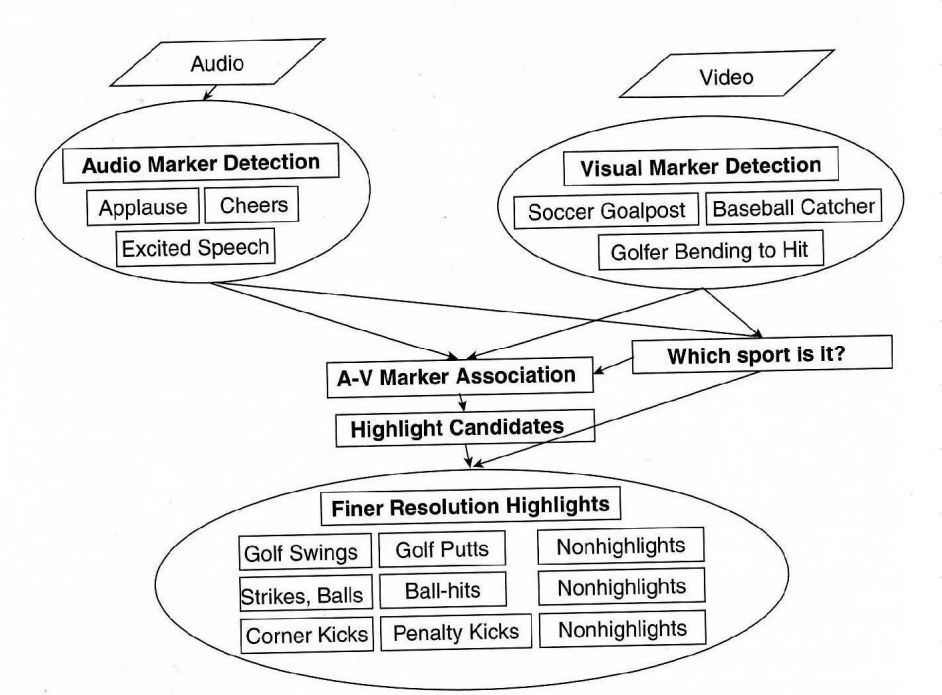



Fig. 1. Framework overview

# Audio marker recognition

Audience reaction to the interesting moments of the game can be used as audio markers.

Example audience reaction classes:

- applause, cheering, music, commentators exited speech, speech and music.

# Audio marker recognition cont…

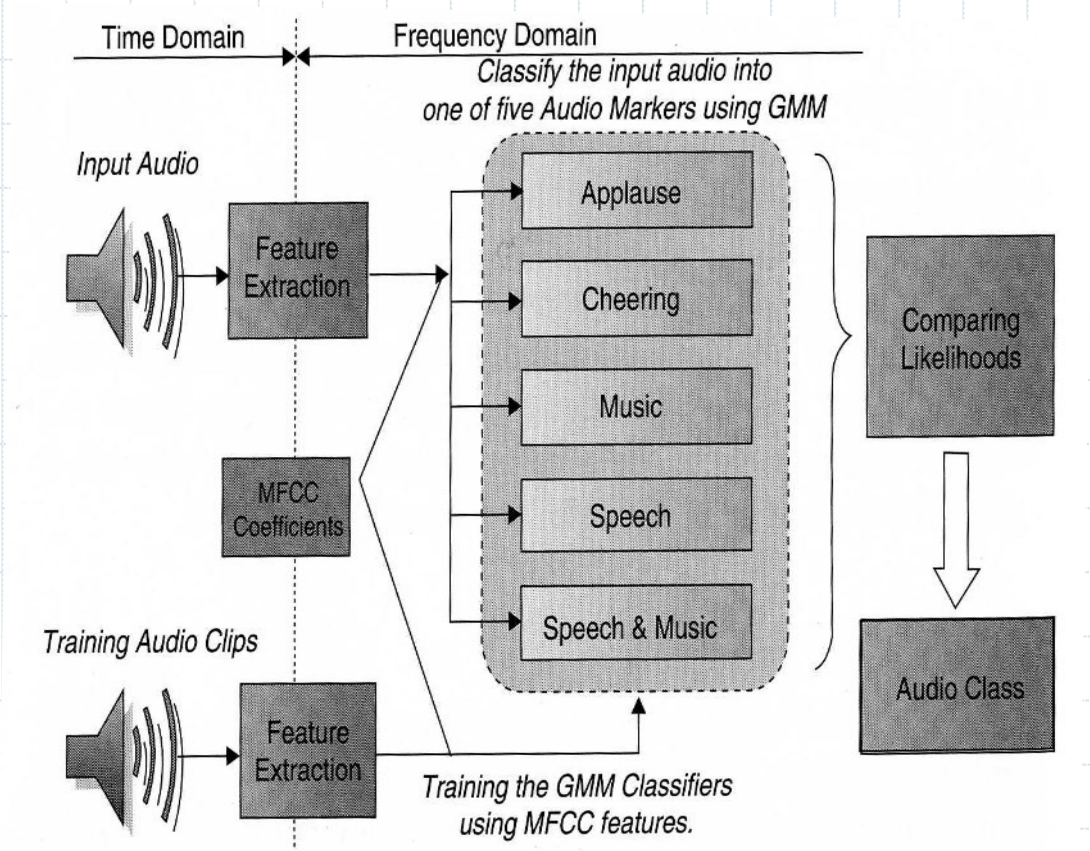



Fig. 2. Audio markers for sports highlights extraction

# Audio marker recognition algorithms

Based on Gaussian mixture models (GMMs)

- Number of Gaussian components (priori) not known, assumed constant for GMMs
- Chosen through cross validation

Problem

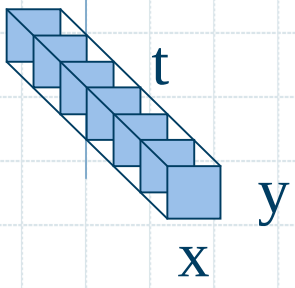
- May lead to overfitting of training data –if it is much less than the actual data and vice versa.

Solution

- Minimum Description Length (MDL) criterion in selecting the number of mixtures.

MDL-GMMs fit the training data to the generative process as closely as possible, avoiding the problem of overfitting or underfitting

# Learning a Probabilistic Model in Space-Time



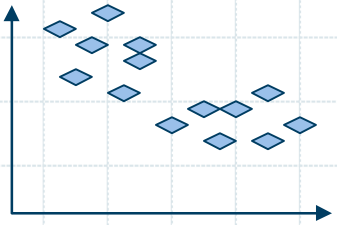

Expectation Maximization (EM)

Feature Vectors

Gaussian Mixture Model

**[L,a,b,x,y,t]**

(6 - dimensional space - Components)



# Audio/Video Representation via Gaussian Mixture Modeling

- Each Component of the GMM Represents a Cluster in the Feature Space (=Blob) and a Spatio-temporal region in the audio/video

- PdF for the GMM :

$$f(x|\theta) = \sum_{j=1}^{k} \alpha_j \frac{1}{\sqrt{(2\pi)^d |\Sigma_j|}} \exp\left\{-\frac{1}{2}(x-\mu_j)^T \Sigma_j^{-1}(x-\mu_j)\right\}$$

With the Parameter set $\theta = \{\alpha_j, \mu_j, \Sigma_j\}_{j=1}^{k}$

- Mixture coefficients, mean, variance

# Example MDL-GMM for Sound Classes

- The algorithm was applied on 679 TV audio clips of golf, baseball and soccer.
- Each clip was hand labeled into the 5 classes as ground truth: applause, cheering, music, speech, and speech with music
- The results were 105, 82, 185, 168 and 139 respectively,
- Total audio duration about 1 h and 12 min
- 100 12 dimension mel-frequency cepstrum coefficients (MFCC) per second using a 25-ms window were extracted
- $1^{st}$ and $2^{nd}$ order time deriviatives were also used to enhance performance
- For each class, begin with a big K (Gaussian Components), calculate the MDL score MDL(K, $\theta$) using all the training sound files, then merge the two nearest Gaussian Components to get the next MDL score MDL(K-1, $\theta$), iterate til K = 1.
- The optimal number of K is chosen as the one that gives the minimum of the MDL scores. For the training database used the results are shown on the next slide.

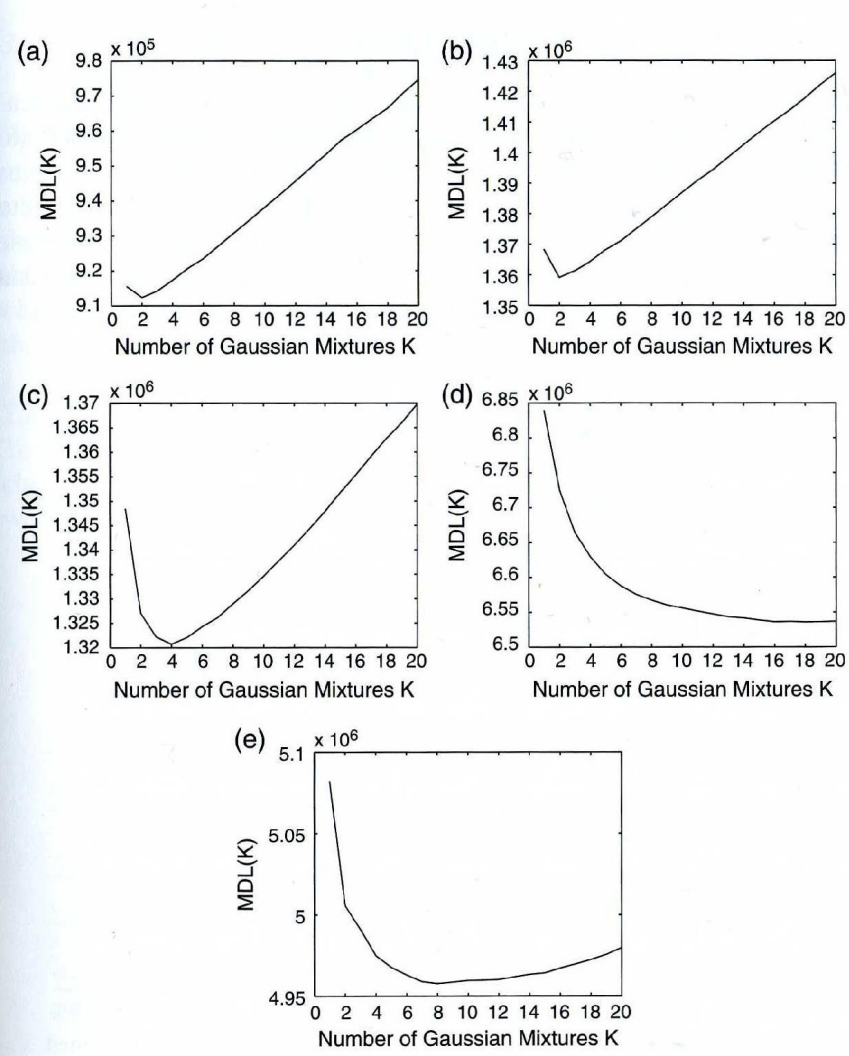



Fig. 3. MDL(K, θ ) (Y - axis) with respect to different numbers of GMM mixtures K(X -axis) to model the 5 classes of audio markers (K = 1…20). Optimal mixture numbers at lowest positions 2, 2, 4, 18 and 8 respectively

## Evaluation using the precision recall curve

- L length of contiguous applause or cheering
- N number of true highlights
- C candidates that have applause or cheering of length greater than or equal to L
- G Total number of highlights in the Ground Truth set.
- Precision (Pr = N/C)
- Recall (Re = N/G)

A curve that is closer to the upper right corner suggest better performance

# Performance comparison

Comparing (1) GMMs with (2 ) MDL-GMMs results, 5 classes, 90/10 traning/test sets. For (1) the number of MMs is taken as 10 For all classes and for (2), they are taken From Fig. 3.

Algorithm (2) improves overall classification accuracy by more than 8%

*Table 1* **Performance of traditional GMM, every class is modeled using 10 Gaussian mixtures: (1) applause, (2) cheering, (3) music, (4) speech, and (5) "speech with music." Classification accuracy on the 10% data by models trained on the 90% data.**

| | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| (1) | 88.8% | 5.0% | 3% | 2% | 1.2% |
| (2) | 5% | 90.1% | 2% | 0 | 2.9% |
| (3) | 5.6% | 0 | 88.9% | 5.6% | 0 |
| (4) | 0 | 0 | 0 | 94.1% | 5.9% |
| (5) | 0 | 0 | 6.9% | 5.1% | 88% |

*Average Recognition Rate: 90%*

*Table 2* **Performance of MDL-GMM. Classification accuracy on the 10% data by models trained on the 90% data. (1) to (5) are the same as described in Table 3.1.**

| | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| (1) | 97.1% | 0 | 0 | 0.9% | 2.0% |
| (2) | 0 | 99.0% | 1.0% | 0 | 0 |
| (3) | 0 | 1.0% | 99.0% | 0 | 0 |
| (4) | 0 | 0 | 0 | 99.0% | 1.0% |
| (5) | 0 | 0 | 1.0% | 0 | 99.0% |

*Average Recognition Rate: 98.6%*

# Experimental results on golf highlights generation using the Optimal (MDL- GMMs) models

- Training done on all the data
- Includes a few seconds of video before play action
- Results compared with those that are human labeled (Ground truth)

Precision –recall performance comparison

*Table 3.* **The confusion matrix on *all* the audio data. The results are based on MDL-GMMs with different "optimal" numbers of mixtures (see Fig. 3.4).**

| | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| (1) | 97.1% | 0 | 0 | 0.9% | 2.0% |
| (2) | 0 | 99.0% | 1.0% | 0 | 0 |
| (3) | 1.0% | 8.0% | 89.0% | 0 | 2.0% |
| (4) | 0 | 0 | 0 | 92.2% | 7.8% |
| (5) | 0 | 0 | 0.7% | 2.8% | 96.5% |

*Average Recognition Rate: 94.76%*

*Table 4.* **Recognition results of those in Table 3.3. The number of sound examples that are correctly or incorrectly classified for each class is shown.**

| | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| (1) | 102 | 0 | 0 | 1 | 2 |
| (2) | 0 | 81 | 1 | 0 | 0 |
| (3) | 2 | 15 | 164 | 0 | 4 |
| (4) | 0 | 0 | 0 | 155 | 13 |
| (5) | 0 | 0 | 1 | 4 | 134 |

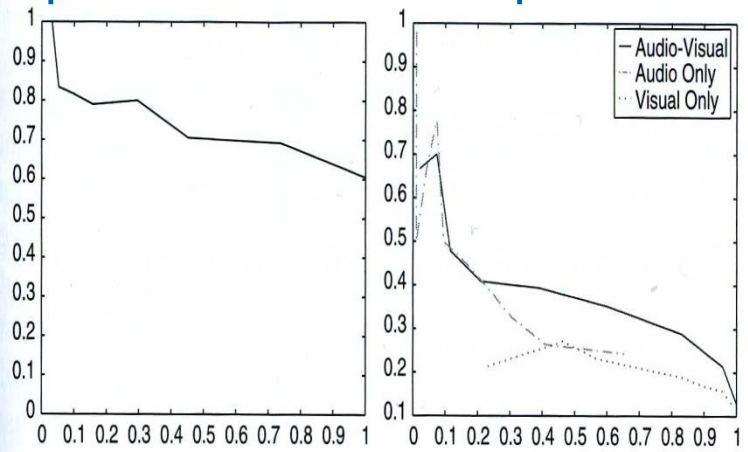



**Figure 4.** Precision-recall curves for the test golf game. Left: by the current approach right: by the previous approaches; Y-axis: precision; X-axis: recall.

# Performance comparison cont..

- Precison –percentage of highlights that are correct of all those extracted
- Recall –percentage of highlights that are in the ground truth set

## Analysis

- all plots are of events of duration L or more
- left Fig. shows plots from the current method (MDL-GMMs)
- Fig on the right also applies Hidden Markov Model (HMM) for every audio chunk (12s window) –to enhance performance
- Solid line curve –results for coupled audio and video using HMM
- Dotted line –video only curve
- Dash –dot –audio only curve

Though superior, Overall performance of Coupled HMM has poor performance at the rightmost part of the curve.

Solution

Take advantage of the fact that key audio classes such as applause and cheering (indicate more possible highlights).

# System Interface (SI)

For providing entry points to video content to viewers

Design aim

To provide a SI where users can adaptively choose other interesting content that is not necessarily modeled by training data

- Depends on the length of sequence highlights needed
- Content –adaptive threshold applied with lower likelihood limit at the bottom and the highest at the top

# System Interface (SI) cont…

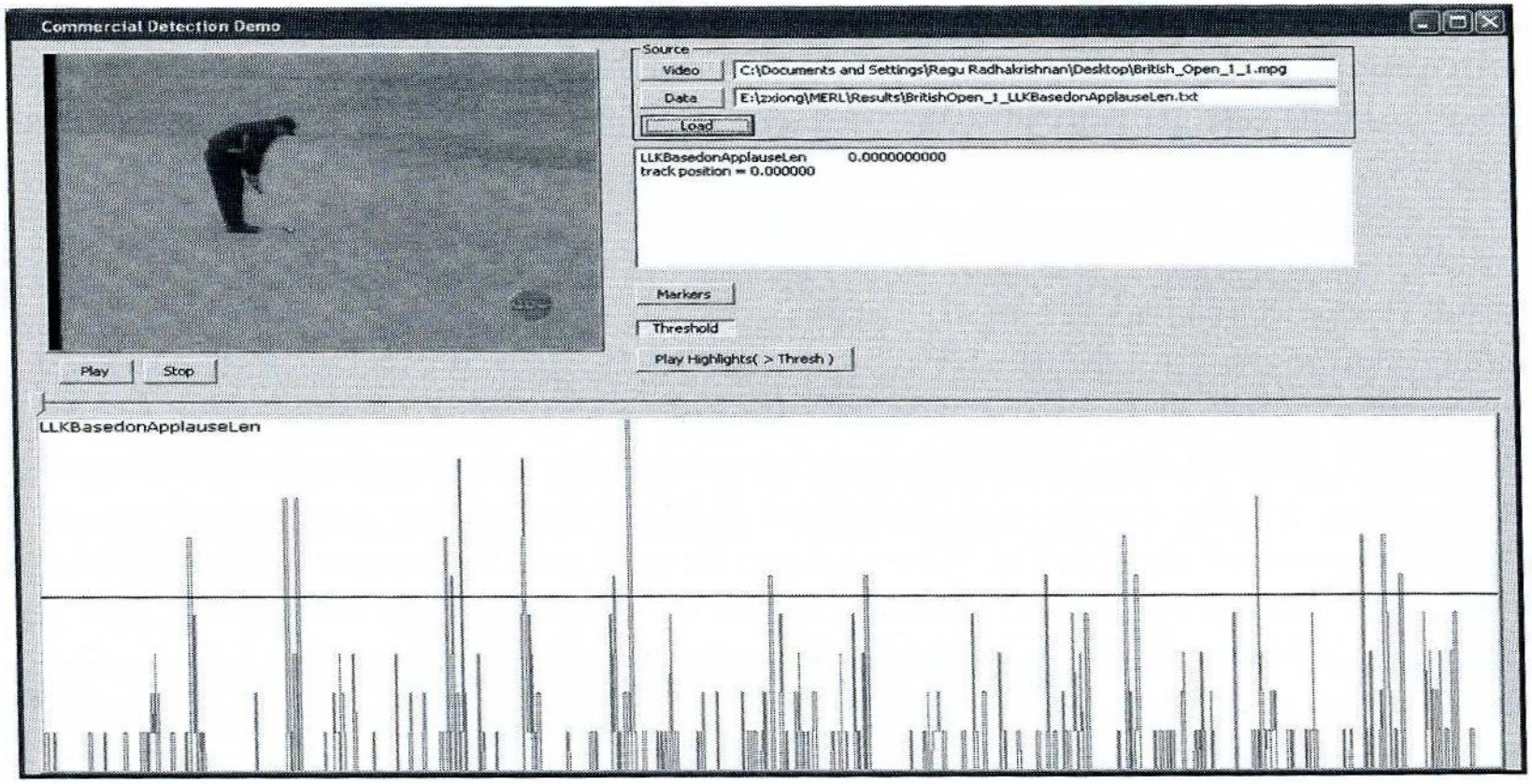



Fig. 5. Snaphot of the Interface. Horizotal line is the threshold value the user can choose to display those segments with a confidence level greater than the threshold [1]

# Visual Marker Detection

- Low level image features like color histogram, texture e.t.c. are suitable for shot detection for scripted video
- Not suitable for unscripted video, for example soccer visual features are so similar for a long period of time, almost all frames may be grouped as one shot
- Semantic level concepts like attacks on the goal, counter attacks in the mid-field provide a way to detect highlights-related visual objects.

# Choice of Visual Markers

Baseball – At he beginning, almost always faces the TV viewers squatting (frontal view)

- Position of the batter and the pitcher varies more than that of the catcher
- Robust identification of those video frames containing the catcher may bring us to the vicinity of the highlights

# Choice of Visual Markers cont…

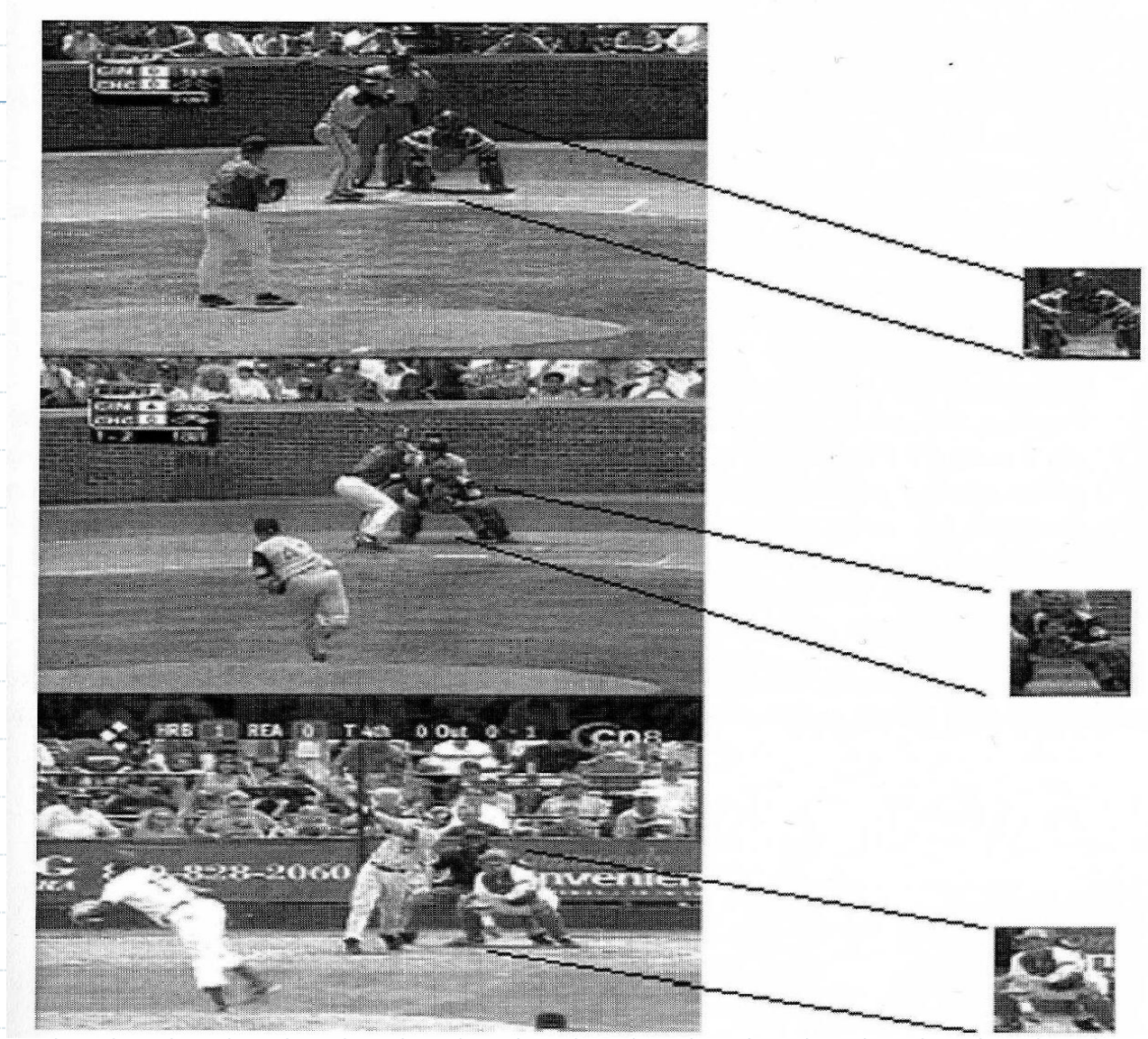


Fig. Some examples of the typical view of the squatting baseball catcher

# Choice of Visual Markers cont...

-Two main observations for soccer game
- Goalpost almost always in view during goals, corner, penalty kicks e.t.c.
- Cameras positioned on either side of the field (Fig.), two detectors used
- So, Robust identification of frames with either of the two goalposts can bring us to the vicinity of the soccer highlights

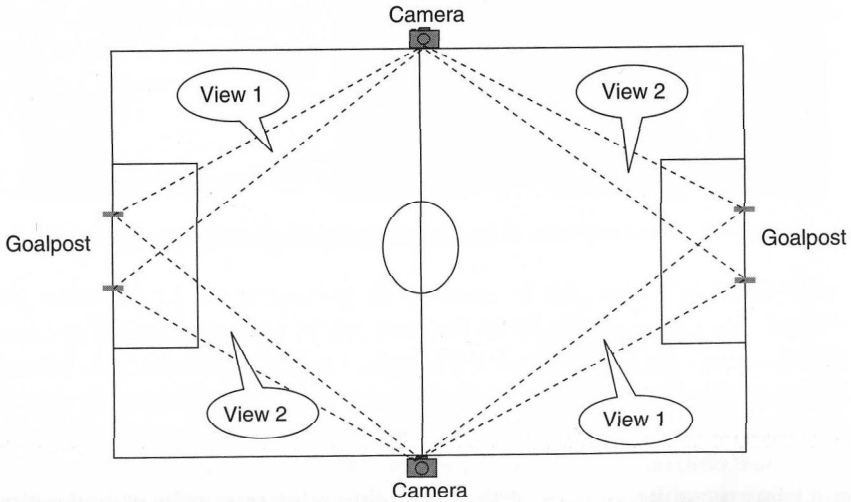



Fig. Camera setup for live soccer broadcast

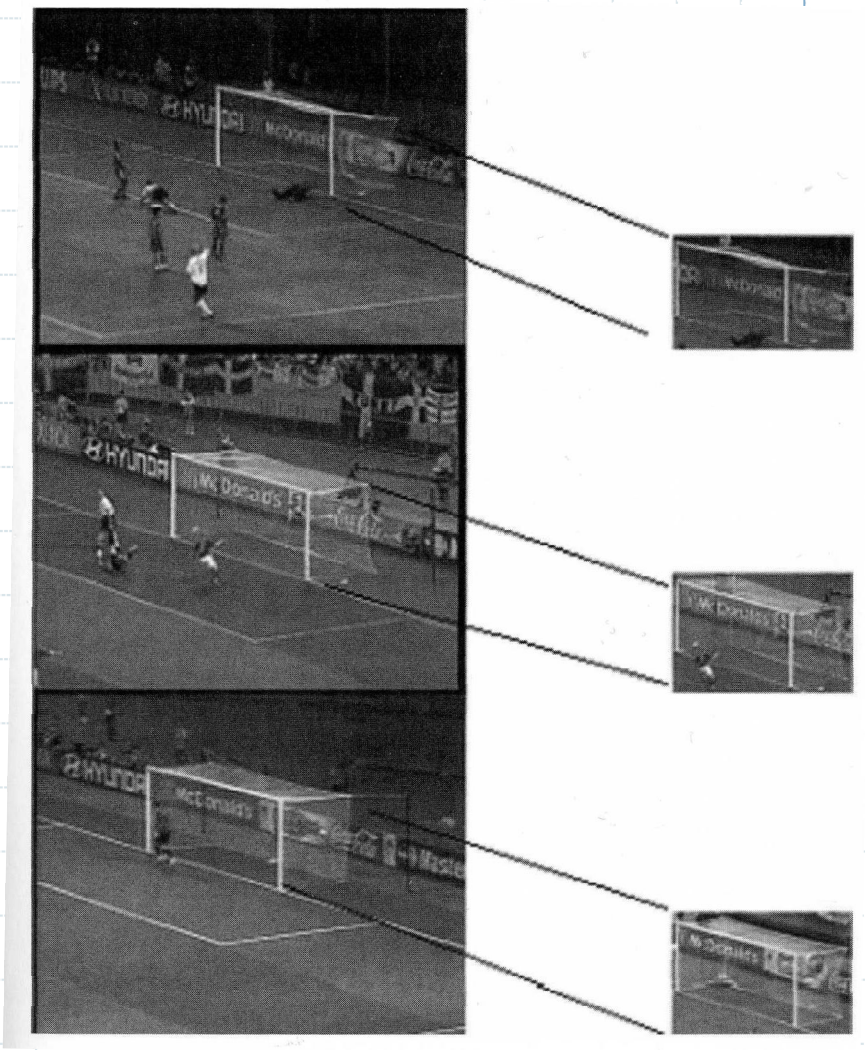


Fig. Examples of soccer goalpost views

# Choice of Visual Markers cont…

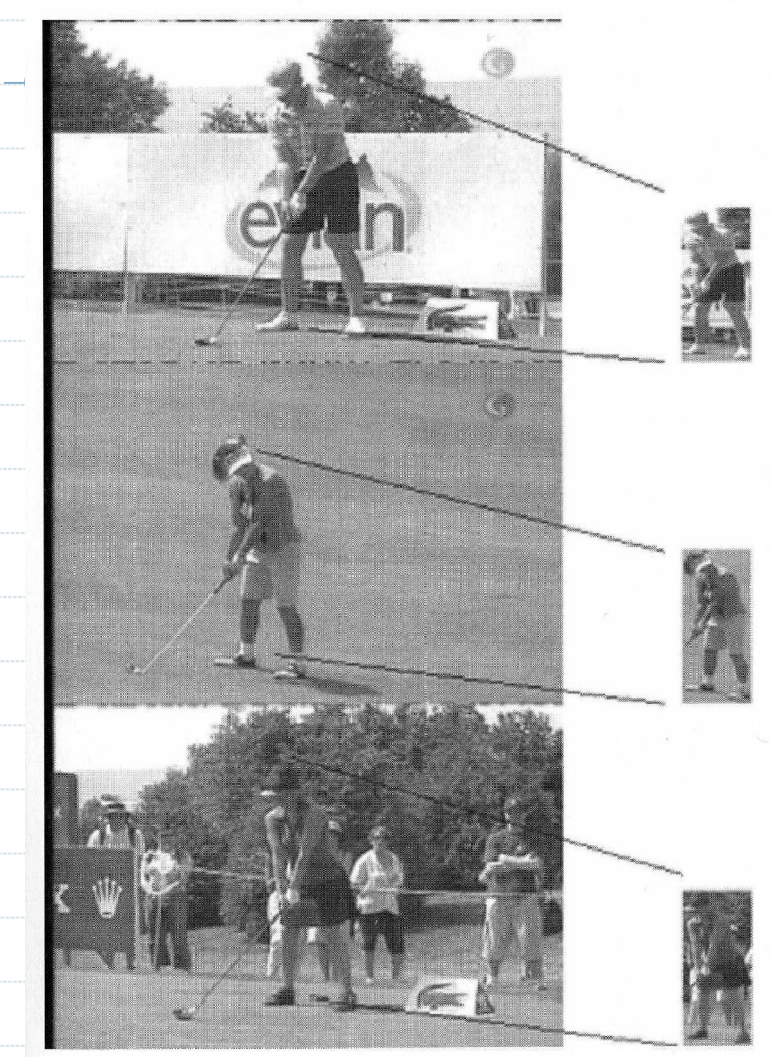


Fig. Examples of some views from golf

-Two main observations
- Golf club and golf ball
- Low accuracy because of: different orientation of the club, motion blur e.t.c.
- Small size of the golf ball

HIGHLIGHT MOVE ESTIMATION
- Based on three principal poses of the golfer
- Frontal (nearly frontal view)
- Side view with golfer bending to the left
- Opposite side view (See examples)
- All these choices are just a compromise

FOCUS
–Audio analysis –applause/cheering is key to effective detection

# Robust real time object detection algorithm

- Viola & Jones "Integral Image "algorithm
- The algorithm allows fast feture calculation
- Can computes various features using rectangles
- Uses AdaBoost learning algorithm

AdaBoost algorithm

- Uses a small number of critical rectangle features
- uses an adaptive algorithm for boosting weak classifiers and returns very good classifications by changing their weights based on previous errors learned

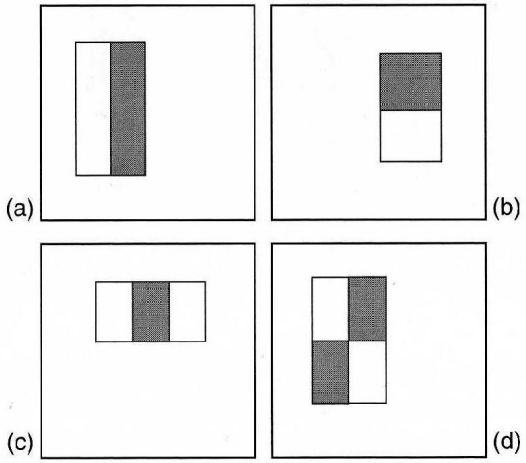



Fig. Example rectangle features shown relative to the enclosing detection window. The sum of the pixels that lie within the white rectangles are subtracted from the sum of pixels in the black rectangles, (a) & (b) show two features © three features and (d) four features

**AdaBoost**

**Inputs**

Training examples $(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)$ with $y_i \in \{0, 1\}$ and the number of iterations $T$.

**Initialize** weights $w_{1,i} = \frac{1}{2m}, \frac{1}{2l}$ for $y_i = 0, 1$ respectively, where $m$ and $l$ are the number of negative and positive examples respectively, with $l + m = n$.

**Train Weak Classifiers**

for $t = 1, \cdots, T$

(1) **Normalize** the weights, $w_{t,i} = \frac{w_{t,i}}{\sum_{j=1}^{n} w_{t,j}}$ so that $w_t$ is a probability distribution.

(2) For each feature $j$, train a weak classifier $h_j$. The error is evaluated with respect to $w_t$, $\epsilon_j = \sum_i w_{t,i} |h_j(x_i) - y_i|$.

(3) Choose the best weak classifier, $h_t$, with the lowest error $\epsilon_t$.

(4) **Update** the weights: $w_{t+1,i} = w_{t,i} \beta_t^{1-e_i}$ where $\beta_t = \frac{\epsilon_t}{1-\epsilon_t}$ and $e_i = 0$ if $x_i$ is classified correctly, $e_i = 1$ otherwise.

**Output**

The final strong classifier is: $h(x) = sign(\sum_{t=1}^{T} (\alpha_t (h_t(x) - \frac{1}{2})))$ where $\alpha_t = \log \frac{1}{\beta_t}$.

Fig. The AdaBoost algorithm

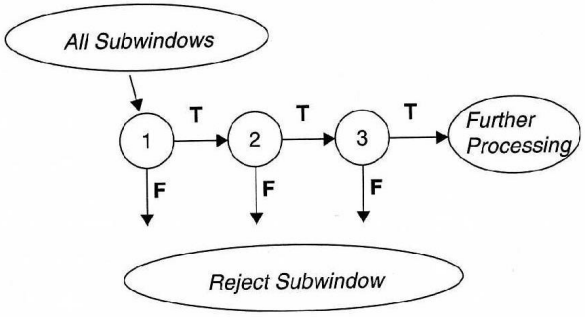



Fig.Detection cascade schematic

# Detection results

## Baseball Catcher

-1464 images, 240 cover different baseball and audience scenes.

-1224 Images not related to baseball

- Algorithm scales & remodels 9 more catcher images to mimic different catcher images.

## Interpretation of results

Three sections captured:

1. sum of hand region,
2. difference between chest region and & the sum of two arm regions.
3. image part that has ground & player s feet

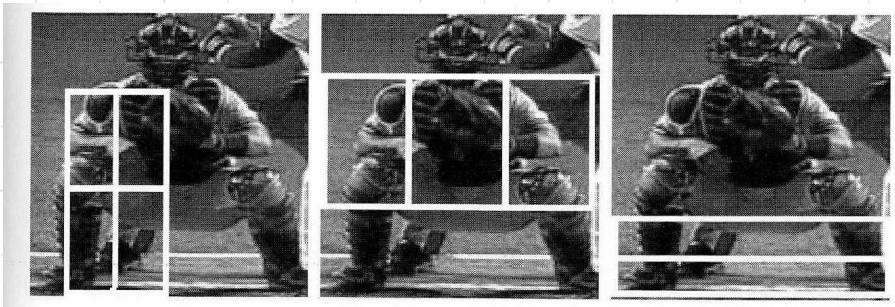


Fig. First few weak classifiers learned by AdaBoost algorithm of the Catcher model

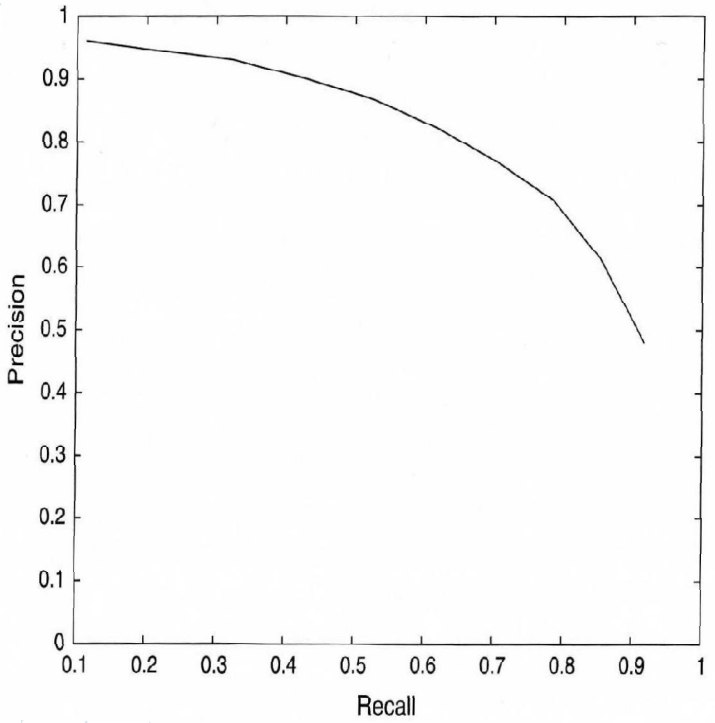



Fig. Precision –recall curve of the Baseball catcher detection.

The graphs show:

- About 80% precision for a recall of 70%
- Uses a threshold and scene frequency in a given number of frames before and after the selected region, the overall result is then regarded as final and is compared to the ground truth (GT) set

# Detection results cont…

## Soccer goalpost detection

- Image parameters and quantity same as in the baseball image collection
- Uses intensity normalization
- algorithm not as good as that used in the previous classifier & peforms badly for unlearned goalpost scenes.

Table. Precision-recall values of the goalpost detection

| *Threshold* | *Precision* | *Recall* | *Threshold* | *Precision* | *Recall* |
|---|---|---|---|---|---|
| 0.1 | 0.464 | 0.521 | 0.2 | 0.676 | 0.281 |
| 0.3 | 0.738 | 0.150 | 0.4 | 0.784 | 0.09 |
| 0.5 | 0.844 | 0.06 | 0.6 | 0.834 | 0.03 |
| 0.7 | 0.961 | 0.02 | 0.8 | 1.000 | 0.01 |

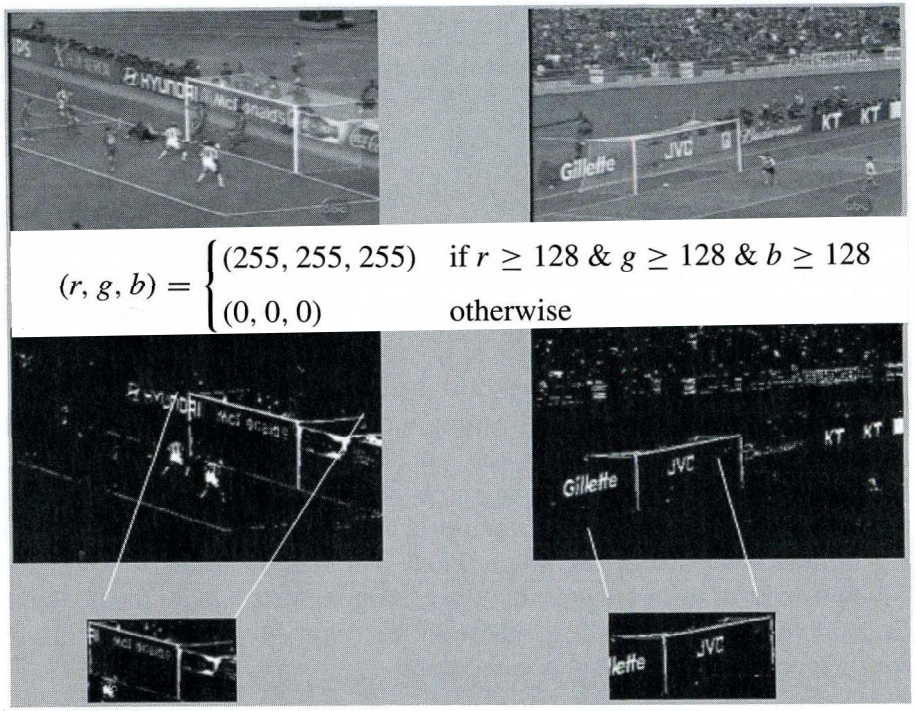



Fig. First few weak classifiers learned by AdaBoost algorithm of the goalpost views.
Second row shows the preprocessing (thresholding) step of the views.

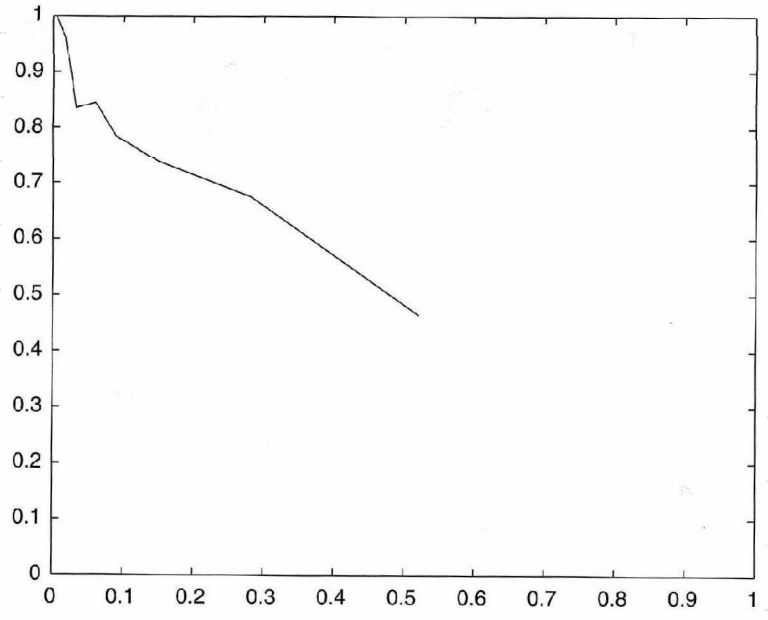



Fig. Precision- recall curve of the goalpost detection

# Detection results cont…

## Golfer model detection

- Image parameters and quantity same as in the baseball image collection
- Included here were golfer images of different sizes under different lighting conditions.
- 9 models of each image used
- algorithm not as good as that used in either of the of the previous classifiers & peforms badly for unlearned scenes.

FOCUS
- Golf audio –so easy to classify

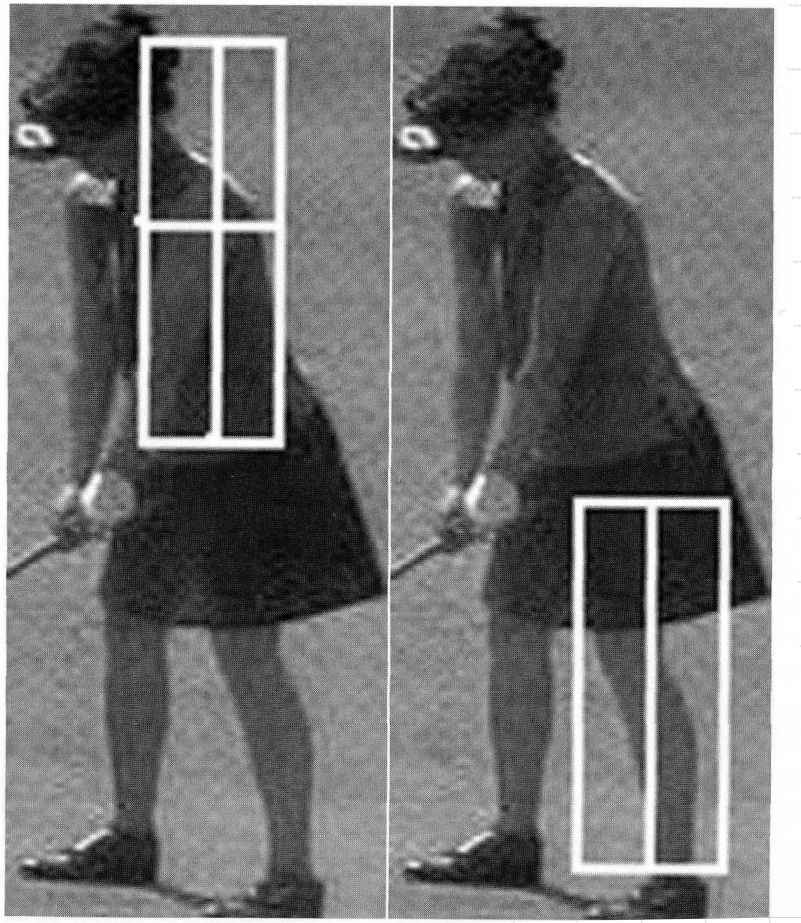


Fig. First few weak classifiers learned by AdaBoost algorithm of the golfer model

# Algorithm summary

Baseball catcher and Soccer goalpost detection algorithms used:

- AdaBoost – based visual object detection algorithm used

Golf

- MDL-GMMs – based audio classification algorithm for detecting long and contigous applause in golf

# Finer resolution Highlights classification

- Ideally, only one visual marker must be associated with one audio marker (pair)
- In reality this is not usually the case
- Preprocessing is done to minimize the grouping errors
- Association process - Match the video marker and the audio marker based on the overlapping threshold

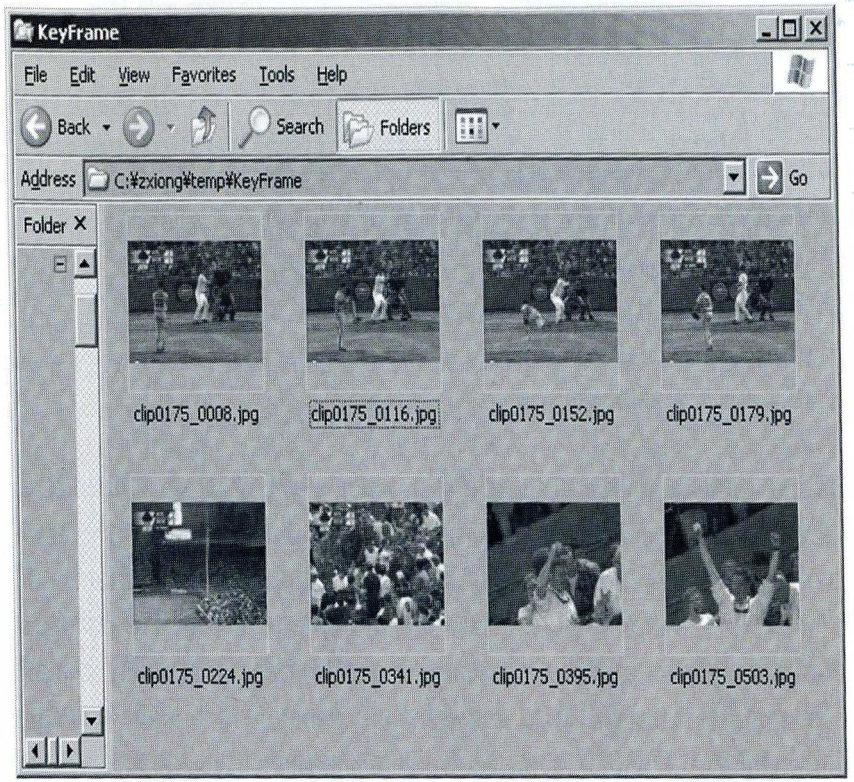



Fig. Example of color change characteristics in a baseball hit

# Methods

Method 1. Clustering

Baseball example

- Two major categories of candidate highlights

1. Balls or strikes –batter does not hit the ball
   - Camera fixed on the pitch scene –
   - Variance of scene color low
2. Ball hits –batter hits the ball to the field or audience

- Camera first shoots pitch scene
- Then follows the ball or audience
- Variance of color is therefore high

From the clip mean, STD & other characteristics, color histograms are constructed and clustering is done using k-means algorithm.

# Methods cont…

## Method 2. Color/Motion modeling using HMMs

Golf example

- Not all long contigous audio represents highlights
- visual pattern usually different from highlight type
- Visual pattern include change in global color and motion intensity
- E.g. in a golf putt, a player stands in the middle of the green (dominant color)
- during player introduction for example, camera usually focused on the announcer ( not so much green)
- During a swing – golfball goes from the ground up in the sky then down to the ground – change of color from color of the ground to sky color and vice versa (two dominant colors)
- Motion – camera follows the ball (up and down) – gives possibility to capture motion intensity features
- gather samples of putts, swings
- Train the model and test

# Method2 cont…

Method 2.1. Modeling highlights by motion using HMMs

- Motion highlight *m*, computed as average magnitude of effective motion vectors in the frame

$$m = \frac{1}{|\Phi|} \sum_{\Phi} \sqrt{v_x^2 + v_y^2}$$

- $\Phi$ –intercoded macroblocks
- V= ($v_x$+ $v_y$)- motion vector (MV) for each macroblock
- MV - Measure of motion intensity –estimation of the gross motion in the frame, includes object and camera motion + color features
- Indicates semantics, –high motion = player action
- Static wide shot = game pause

# Experimental results, observations and comparisons

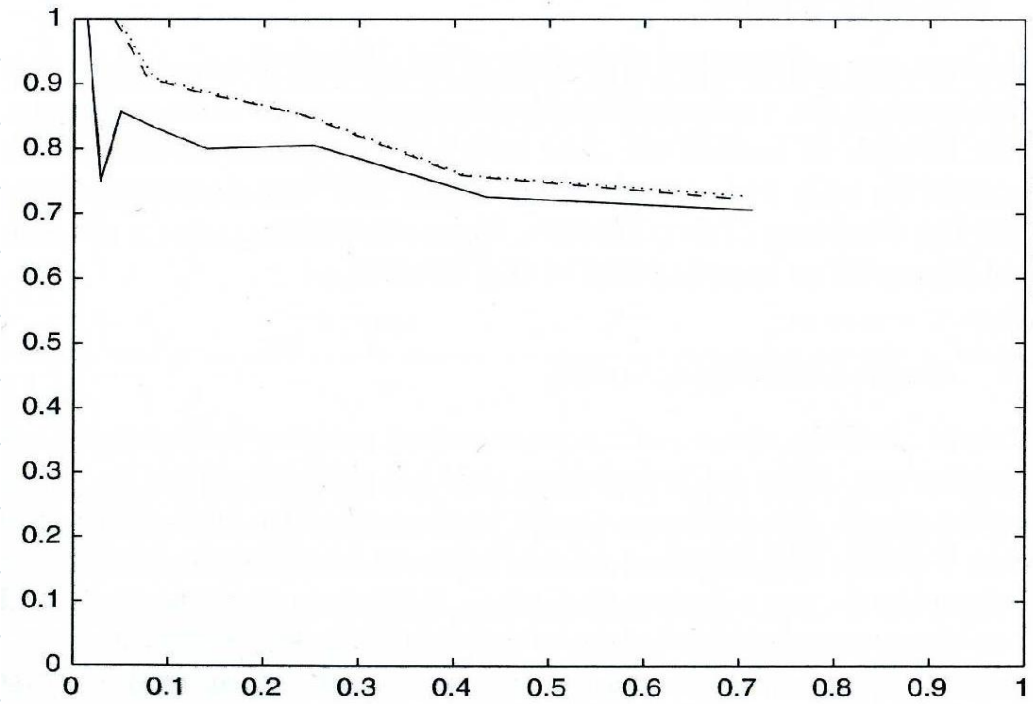
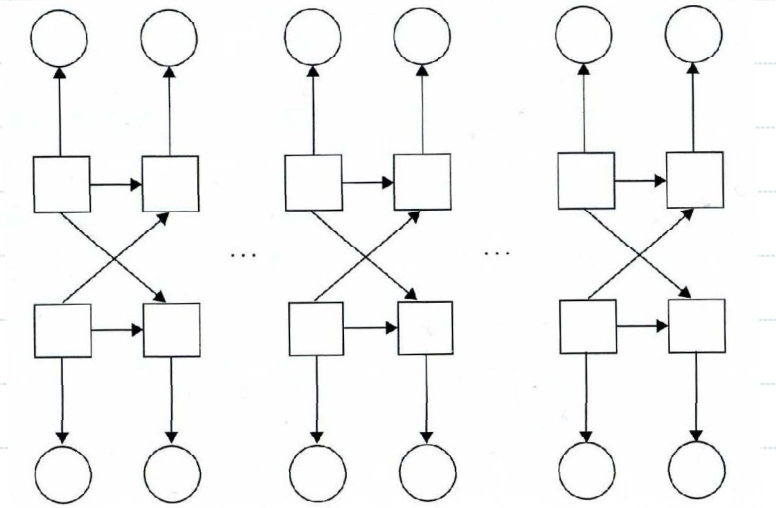



Fig. Comparison results of the three different modeling approaches in terms of precision- recall curves. Solid line: audio modeling alone; dashed line: audio+dominant color modeling; dotted line: audio + dominant color + motion modeling. X-axis: recall; Y-axis: precision

# Methods cont…

## Method 3 Audio - Visual modeling using CHMMs

- Uses Discrete coupled HMM collection
- Two data streams (Audio + video)
- Coupled through transitional probabilities
- Fro time t-1 to t
- Audio label generation - Models built for applause, ball hit, cheers, music speech etc
- Video label generation – uses modified version of MPEG 7 motion activity descriptor
- Captures motion intensity action
- Extracted by quantizing the varaince of the magnitude of the motion vectors between neighboring frames and then classifies them to given lables eg low, very low, medium,high etc
- Information (Audio + Video) used for model training and testing



DCHMM model. Squares –Hidden states, Circles observable states

Table Audio Labels and class names

| *Audio Label* | *Its Meaning* | *Audio Label* | *Its Meaning* |
|---|---|---|---|
| 1 | Silence | 5 | Music |
| 2 | Applause | 6 | Female speech |
| 3 | Ball-hit | 7 | Male speech |
| 4 | Cheering | 8 | Speech with music |

# CHMM Vs HMM Precision-recall comparison curves

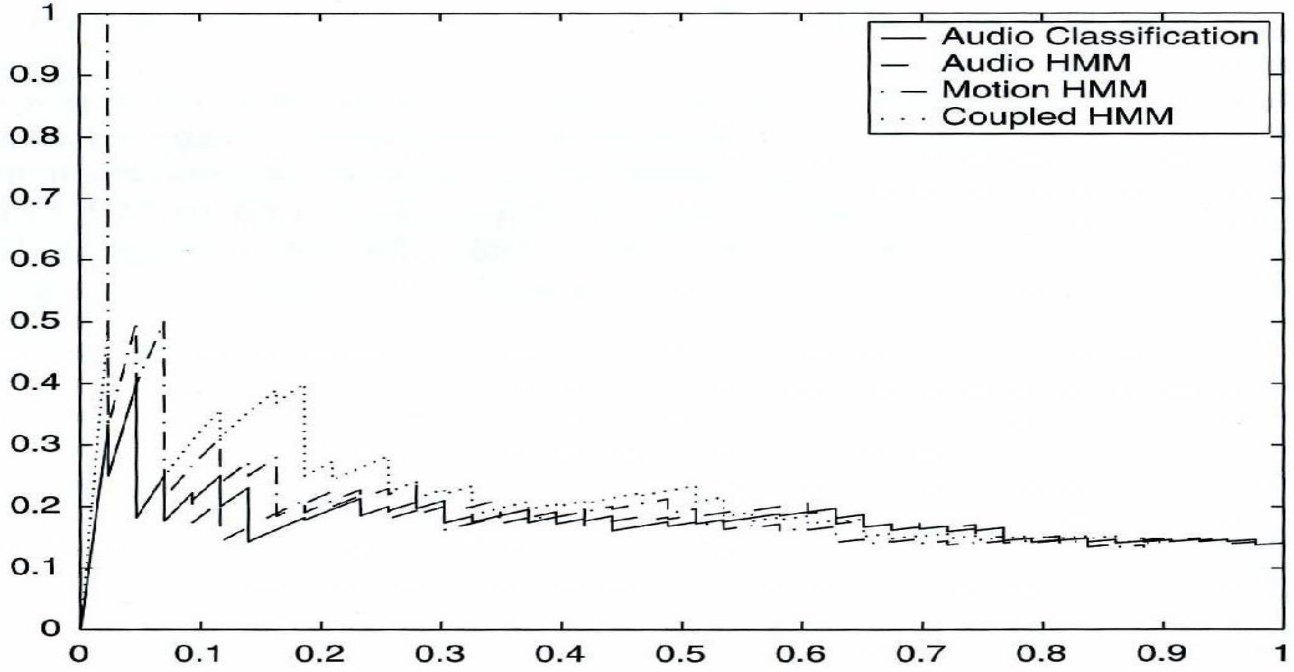



**Figure 3.36** Precision-recall curves for the test soccer game. The four highlights extraction methods compared here are (1) audio classification followed by long, contiguous cheering selection; (2) HMM classification using the models trained from audio labels of the highlight and nonhighlight examples; (3) HMM classification using the models trained from video (motion) labels of the highlight and nonhighlight examples; and (4) coupled HMM classification using the models trained from both audio and video (motion) labels of the highlight and nonhighlight examples. X-axis: recall; Y-axis: precision.

Comments

- CHMM –based approach shows higher precision over single modality HMM
- CHMM has lower false alarm rate for a given recall rate

# Review

- 3 methods presented for finer resolution highlights extraction
- Common features – All operate after highlight candidates have been found, either by audio marker detection e.g in golf or by joint audio –video association (Soccer, Baseball)
- Differences –Complexity and supervision
- Method 1. – simple, but needs parameter tuning (non universal)
- Method 2 & 3 –increase in complexity

# Future improvements

- More work must be done to improve on:
- Some false alarms wrongly classified as highlights
- Though they share some properties with highlights e.g soccer midfield level fighting with high motion activity with a high level of background sound that is classified as cheers.

# Summary

In this presentation, we have looked at:

- Highlight extraction framework based on hierarchical representation tha includes:
- play/break segmentation
- audio-visual marker detection, association and finer- resolution highlight classification.
- also included in the framework classification algorithms were the semantic and subjective concepts of sports highlights.
- The key component in this framework was the detection of audio and visual objects that serve as the bridge between the observed video signal and the semantics.
- The experimental results have conffirmed the effectiveness and advantage of this approach.

# References

1. Z. Xiong, R. Radhakrishnan, A. Divakaran, Y. Rui, T. S. Huang, ”A unified framework for video summarization, browsing and retrieval, ” Elsevier academic press, 2006

Thank you!

Discussions and questions welcome