行政院國家科學委員會專題研究計畫　成果報告

以場景分析為基礎之棒球運動節目事件偵測系統

計畫類別：個別型計畫
計畫編號：NSC94-2213-E-216-023-
執行期間：94 年 08 月 01 日至 95 年 07 月 31 日
執行單位：中華大學資訊工程學系

計畫主持人：連振昌

報告類型：精簡報告
報告附件：出席國際會議研究心得報告及發表論文
處理方式：本計畫涉及專利或其他智慧財產權，1 年後可公開查詢

中　華　民　國 95 年 10 月 13 日

行政院國家科學委員會補助專題研究計畫 ☑成果報告 □期中進度報告

（以場景分析為基礎之棒球運動節目事件偵測系統）

計畫類別：☑ 個別型計畫　　□ 整合型計畫

計畫編號：NSC 94-2213-E-216 -023 -

執行期間：　94 年 08 月 01 日　至　95 年 07 月 31 日

計畫主持人：連振昌

共同主持人：

計畫參與人員：石家銘，林建程

成果報告類型(依經費核定清單規定繳交)：□精簡報告　☑完整報告

本成果報告包括以下應繳交之附件：

□赴國外出差或研習心得報告一份

□赴大陸地區出差或研習心得報告一份

☑出席國際學術會議心得報告及發表之論文各一份

□國際合作研究計畫國外研究報告書一份

處理方式：除產學合作研究計畫、提升產業技術及人才培育研究計畫、列管計畫及下列情形者外，得立即公開查詢

□涉及專利或其他智慧財產權，□一年□二年後可公開查詢

執行單位：中華大學資訊工程學系

中華民國 95 年 10 月 11 日

1. 中英文摘要

本計劃提出了一個針對棒球節目發展以棒球運動事件為基礎的視訊偵測及檢索系統。傳統之運動節目事件偵測系統通常以視訊畫面(frame-based)或以視訊鏡頭(shot-based)為基礎作視訊特徵之擷取及語義之萃取，進而偵測特定之運動事件。但是，此做法較難得到較完整運動事件之視訊畫面，同時也比較沒有結構化來分析棒球節目。棒球運動節目結構化會讓使用者會更容易得到他想查詢的片段。一開始我們先切割出所有的鏡頭(shot)，接著針對各個鏡頭的關鍵畫面(key-frame)及在此鏡頭(shot)內之連續視訊畫面做靜態及動態之影像特徵擷取。靜態之影像特徵包含球場內之草地與泥土顏色的偵測、物件的切割及擷取、球員之膚色偵測等，動態之視訊特徵主要是場景變換之位移量。接著利用事先規劃好的規則來做帶有語義之場景(semantic scene)偵測及分類。通常一特定之棒球運動事件是由一連串且連續之語義場景所組成，在本計劃中場景分類包含：投球場景(pitching scene)，跑壘場景(running scene)，本壘場景(base scene)，打擊場景(player scene)，內野場景(infield scene)，外野場景(outfield scene)，特寫場景(close-up scenes)以及其他類場景。利用這一連串分類好的場景套入 HMM (Hidden Markov Models) 中做事件的偵測，可以準確的偵測出某一特定之棒球運動事件。在本計劃裡，我們想要偵測之棒球運動事件有下列四種，分別為：三振、接殺，刺殺、以及安打，實驗的數據顯示我們的系統可以正確的偵測出此四種棒球運動事件。如此對於某一特定之棒球運動事件，使用者可以獲得完整之視訊畫面。

關鍵字：場景，運動視訊，隱藏式馬可夫串列，關鍵畫面。

Recently, many researches are addressed on the event detection of sport video. Most of these researches are categorized into the frame-based or the shot-based systems. By extracting the semantics for the successive frames or the segmented shots, the various kinds of sport events are detected. However, the low-degree structuralization of the frame-based or the shot-based systems may make the analysis of the sport video more complex and then the retrieval of the video clips for a sport event inefficient. Therefore, in this project, the new scene-based sport event detection system is proposed. Firstly, the sport video is partitioned into many video shots with the same semantics. Secondly, the key frame and several visual features (soil and grass color percentage, object number, motion vector, skin detection, player's location) for each shot are extracted and analyzed to identify the semantics and then each shot will be classified into various kinds of semantic scenes. Here, there are total eight semantic scenes including pitching scene, running scene, base scene, player scene, infield scene, outfield scene, close-up scene, and other scene are defined. Finally, the semantic scenes are regarded as the observation symbols in the HMM (Hidden Markov Model) event detection system. Four kinds of events including the base hit, strikeout, ground outs, and air outs are detected in the proposed system. Experimental results show that the proposed system may detect the four kinds of sport events accurately.

Keywords: Video shot, Video scene, Sport event detection, HMM

2. 報告內容

Scene-Based Event Detection for Baseball Videos

Cheng-Chang Lien, Chiu-Lung Chiang, and Chang-Hsing Lee
Dept. of Computer Science and Information Engineering
Chung Hua University
Hsin-Chu, Taiwan, R.O.C.
Email: cclien@chu.edu.tw

Abstract

A lot of research has lately been focusing on scene analysis in sport videos. By extracting the semantics of successive frames or segmented shots, various kinds of video scenes may be identified. However, general baseball events, e.g., strikeout and ground outs, are hard to be detected because a general baseball event is composed of a series of video scenes and each scene is further composed of several video shots. Hence, the detection of general baseball events has to be developed in terms of scenes to facilitate the retrieval of the required video clips. To do this, the baseball video is firstly segmented into many video shots. Then, various visual features including the image-based features, object-based features, and global motion are extracted to analyze the semantics for each video shot. Each video shot is then classified into the predefined semantic scenes according to its semantics. Finally, the hidden Markov model (HMM) is applied to detect the general baseball events by regarding the classified scenes as observation symbols. The accuracy analysis for the scene classification and event detection are illustrated with a large amount of video data consisting of several hours of video frames. Experimental results show that the proposed system detects the four kinds of general baseball events with reasonable accuracy.

Keywords: Video shot, Semantic scene, Baseball event, Hidden Markov model

1. Introduction

With the great demand for fast browsing and searching sport video contents, event-based video searching technology is being developed rapidly. In general, event-based searching technology is based on scene analysis. The scene analysis technology may be roughly categorized into the frame-based [1-6] and the shot-based [7-11] methods.

In [1], the important events were modeled by the state transition diagram of "play" and "non-play". The "plays" were detected by applying some low-level visual features. Based on the state transition model, the video content was summarized. However, the detection of some general events was not addressed. In [2], human running behavior in a sport video was detected by using a periodic motion descriptor. In [3], replay video segments in sport programs were automatically detected by identifying the logo in the scene transitions. Furthermore, baseball scenes were also classified by using the maximum entropy method [4] in which image, audio, and text features were utilized. The methods in [5-6] also applied audio and text features to develop event detection systems. The abovementioned frame-based methods are suitable for analyzing the scenes in sport videos but the loose structured video representation makes it difficult for them to identify general baseball events, e.g., base hit, strikeout, ground outs, and air outs.

For the shot-based methods, Ekin *et al.* [7] applied cinematic and object-based features to detect the goal, referee, and penalty-box and then to summarize the soccer videos. Then, the soccer videos are analyzed and summarized. Lu *et al.* [8] proposed a structuring system to classify the basketball scenes using LDA and some low-level features. Especially in [9], shot-based baseball scene classification is proposed. Seven scenes were detected by using some visual features. In this study, this method was improved by considering global motion and object-based features for more scene classifications. Based on these classified scenes, HMM was applied to identify general baseball events.

To detect video events, a state machine was often used. In [10], the state diagram of a hunting event was generated to detect the hunts in a wildlife video. In [11], the shot-based state diagram was also generated for video surveillance. Here, the concept of the state machine was also adopted to develop a scene-based baseball event detection system facilitating the searching of general baseball events.

General baseball events, e.g., strikeout and ground outs, are hard to be detected by using the frame-based or shot-based methods because a general baseball event is composed of a series of video scenes and each scene is further composed of several video shots. Some semantic scenes including infield, outfield, player, and pitching may consist of several video shots. Instead of developing the shot-based method, scene transition was used to design the state machine and detect the baseball events. The scene-based baseball event detection system was developed with the following phases. Firstly, the video was segmented into the various video shots using the knowledge-based shot-boundary detecting method [15]. Secondly, the key-frame was extracted for each video shot and then several visual features (color distribution, motion information) were extracted within the key-frames and shots. Furthermore, the moving objects were extracted by using the multiresolution and flooding based RSST (MFRSST) video segmentation [12] such that the number and relative positions of players were easily acquired. Based on these visual features the various kinds of semantic scenes were classified. Finally, the scene-based baseball event detecting system was developed by applying the hidden Markov models (HMM). The baseball events consisted of base hits, strikeouts, ground outs, and air outs. The block diagram of the proposed system is shown in Fig. 1. In addition, the structure of a baseball video is shown in Fig. 2. A baseball game is composed of many innings and each inning is composed of two half innings.

The accuracy analyses for the scene classification and event detection were using a large amount of video data that consisted of several hours of video frames. Experimental results show that the proposed system can identify the four kinds of baseball events accurately.

2. Shot Boundary Detection

Generally, a baseball event is composed of a lot of successive scenes and each scene is further composed of a series of shots. The precise shot detection [13-15] and scene classification may make the detection of baseball events accurate. Here, the method proposed by Hanjalic et al. [15] was applied to detect the shot boundary. In [15], the prior knowledge of shot-boundary transition was used to improve the detection accuracy. The prior knowledge includes shot-length distribution, visual discontinuity patterns at the shot boundaries, and temporal changing characteristics (average size of shot boundary) around a boundary. Firstly, by computing the block-based motion compensation error, the discontinuity function was defined to measure the discontinuity between successive frames and written as:

$$z(k,k+1) = \frac{1}{N_{Blocks}} \sum_{i=1}^{N_{Blocks}} D(b_i(k), b_{i,m}(k+1)), \qquad (1)$$

where $b_{i,m}(k+1)$ is the motion estimated block for $b_i(k)$ on adjacent frame $k+1$ and $\mathrm{D}(b_i(k), b_{i,m}(k+1))$ is the motion compensated prediction error using the block mean values of Y, U, V components between frames k and $k+1$, which is defined as:

$$\begin{aligned} D(b_i(k), b_{i,m}(k+1)) &= \left|Y_{ave}(b_i(k)) - Y_{ave}(b_{i,m}(k+1))\right| \\ &+ \left|U_{ave}(b_i(k)) - U_{ave}(b_{i,m}(k+1))\right| + \left|V_{ave}(b_i(k)) - V_{ave}(b_{i,m}(k+1))\right| . \end{aligned}$$

Secondly, an adaptive discontinuity threshold [15] for detecting the shot boundary was determined by applying a prior probability function of the shot length and a conditional probability function of shot discontinuity between frames k and $k+1$. The prior probability function of the shot length may be modeled as a Poisson distribution and defined as:

$$P_k^a(S) = \frac{1}{2} \sum_{w=0}^{\lambda(k)} \frac{\mu^w}{w!} e^{-\mu} \quad , \qquad (2)$$

where the parameter μ denotes the average shot length in a video sequence, w is the frame number counted from the previous shot boundary and reset to zero each time a shot boundary is detected, and $\lambda(k)$ is the current shot length at the frame k. Generally hard cuts are detected at the positions where the discontinuity value reaches the local maximum. In [15], a function $\psi(k)$ used to describe the characteristics of the shot boundary of a hard cut is defined as:

$$\psi(k) = \begin{cases} 100 \times \dfrac{z'(k,k+1) - z_{sm}}{z'(k,k+1)} \% & \text{, if } z'(k,k+1) = \max(\forall z(i,i+1), \\ & i \in (k - \dfrac{N-1}{2}, k + \dfrac{N-1}{2} - 1)) \\ 0 & \text{, else} \end{cases} , \qquad (3)$$

where z_{sm} denotes the second largest discontinuity value and parameter N denotes the windows size. Then, a conditional probability function defined as Eq. (4) is used to model discontinuity probability distribution between successive frames k and $k+1$.

$$P_k(S \mid \psi(k)) = \frac{1}{2}(1 + erf(\frac{\psi(k) - d}{\sigma_{erf}})) \quad , \qquad (4)$$

where $erf(x) = \frac{2}{\pi}\int_0^x e^{-t^2} dt$, parameter d is the time interval from the starting frame to the current frame and σ_{erf} is the factor that determines the steepness of the middle curve segment. For various types of boundary transition, the parameter pair (d, σ_{erf}) is adjusted to fit the type of boundary transition. The parameter pair (d, σ_{erf}) is set to (60, 2) such that it fits the characteristics of the hard cut boundary transition. Finally, by combining (3) and (4), the adaptive threshold function is:

$$T(k) = \frac{1 - P_k^a(S) P_k(S \mid \psi(k))}{P_k^a(S) P_k(S \mid \psi(k))} \quad . \qquad (5)$$

If $z(k,k+1) > T(k)$ then there is a shot boundary between frames k and $k+1$. Otherwise there is no shot boundary between frames k and $k+1$.

3. Semantics Extraction and Classification

The semantics of each video shot was extracted according to the image-based features (soil and grass regions, skin color) [9], object-based features (object number, object size, relative positions of players), and global motion information. In total, eight semantic scenes were classified, namely pitching, running, base, player, infield, outfield, close-up, and other scenes. Fig. 3 illustrates the flowchart classifying the semantic scenes for the baseball video shots. First, the key-frame for each video shot was determined using the median filtering [16]. Second, based on the key frame, the scene for each shot was roughly classified according to the area proportion of soil and grass regions. Third, global camera motion was used to detect the running scenes. Finally, the proposed MFRSST video segmentation method [12] was applied to extract the moving objects and then the number and locations of the moving objects (players) were used to classify the other semantic scenes.

3.1 Global Motion Estimation

The main objective of global motion extraction is to acquire a global motion compensated image so that the player regions can be obtained. Here, the method of indirect global motion estimation [18] was applied to calculate global motion. The affine parameters of global motion were estimated by minimizing the following energy function:

$$E_{fit} = \sum_{x_n \in S} w_n \, | d(x_n; a) - d_n |^2 , \tag{6}$$

where S represent the set of pixels used to calculate the global motion, w_n is the weighting coefficient for the pixel x_n, and $\boldsymbol{a}=[a_0, a_1, a_2, b_0, b_1, b_2]^\mathrm{T}$ is the affine parameter vector. In Eq. (6), the motion vector d_n is obtained by using a three-step search [18]. By setting $\partial E_\mathrm{fit}/\partial \boldsymbol{a}=0$, the parameter vector $\boldsymbol{a}$ is obtained as:

$$a = \left(\sum_{n \in N} w_n [A(x_n)]^T [A(x_n)] \right)^{-1} \left(\sum_{n \in N} w_n [A(x_n)]^T d_n \right) , \tag{7}$$

where the matrix $[A(x_n)]$ is defined as:

$$[A(x_n)] = \begin{bmatrix} 1 & x_n & y_n & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & x_n & y_n \end{bmatrix} .$$

3.2 Extraction of Player Information

By careful observation, the number and locations of players may be utilized to classify the semantic scenes with players. For example, when the color distribution of the soil and grass regions is medium, the players' locations are needed to determine whether it is a pitching scene or not. Hence, before detecting player location the object segmentation process is needed to extract the player regions. In this study, MFRSST (multiresolution and flooding based recursive shortest spanning tree) [12] was used to separate the object regions with the change regions between the key frame and its global motion compensated image. Once the objects were separated they were counted. MFRSST speeds up the segmentation process and with the same partition quality as RSST [22].

Two image processing methods were adopted to develop the MFRSST image segmentation algorithm. The first one was flooding in the watershed algorithm [21]. Instead of merging the two vertices of the least weighted link, the neighboring vertices whose link weights closed to the lowest link weight were merged concurrently. Such a merging process is similar to a flooding in the watershed algorithm that fills up the image valleys to a specified level. The second concept is that of multiresolution decomposition. Performing the partitioning in the decimated image can significantly reduce the computation cost. The block diagram of MFRSST method is illustrated in Fig. 4-(a) and the method is described as follows.

1. Perform decimation process to obtain low resolution image.
2. Apply the FRSST [20] method to partition the decimated image into N regions.
3. Apply fusion process to remove small spark regions.
4. Interpolate the segmented decimated image according to the following rules:

```
    If the pixel is located at the region boundary (see Fig. 4-(b)),
        interpolate it with an empty block (2x2) and fill it with the pixels from original image.
        (see Fig. 4-(d))
    Else,
        interpolate it with a block (2×2) having the same value. (see Fig. 4-(c))
5. Regard each reconstructed region as a single node and then construct the new weighted graph.
6. Form the new shortest spanning tree (SST).
7. Cut the $N$-$1$ links to preserve the previous segmented $N$ regions.
8. Cut the next most costly $M$-1 links to partition the $M$ high detailed regions within the
   boundary region.
9. Map the regions onto a segmentation image.
```

Fig. 5-(a) illustrates the segmented objects. In the following, the labeling algorithm [23] was used to remove some small regions and count the object as shown in Fig. 5-(b).

3.3 Soil and Grass Color Distribution

The infield and outfield scenes were classified by using the area ratio of soil to grass regions [9]. The regions of soil and grass were separated by using the color distribution on the HSV color coordinate [9]. The color distributions for the grass and soil regions were described as:

Color distribution for soil: $5<H<35$, $0.2<S<0.5$, $V>100$
Color distribution for grass: $100<H<150$, $0.2<S<0.5$, $V>100$

Fig. 6 illustrates the color distribution for the soil and grass regions. By calculating the grass and soil areas separately, the scenes were roughly classified according to the following rules:

1. If the area ratio of the grass and soil regions was larger than 45%, then the scene was classified as outfield, infield, or player. Further classification (see Fig. 3) was done by evaluating the ratio of grass to soil and the object size.
2. If the area ratio of the grass and soil regions was between 25% and 45%, then the scene would be classified as a pitching scene. The pitching scene was detected (see Fig. 3) by using the additional object-based feature (player's location shown in Fig. 5-(b)).
3. If the area ratio of the grass and soil regions was lower than 25%, the scene would be classified as the close-up, base, running, or other scenes. Further classification (see Fig. 3) was done by evaluating the other visual features, e.g., motion vector, skin color, and number of objects.

3.4 Detection of Player's Relative Position

When a scene was classified as a candidate pitching scene by using the area ratio of grass to soil, the relative positions of the players was used to determine whether it was a pitching scene or not. To find the relative player's position more efficiently, the image was partitioned into 16 equal blocks as shown in Fig. 7-(a). Because the pitcher, batter, and catcher appear in the pitching scene, the relative positions of the players can help to detect the pitching scene. When the segmented objects (players) were located at the blocks 3 to 14 as shown in Fig. 7-(b), the scene was classified as a pitching scene.

3.5 Skin Color Detection

If the soil and grass regions were small and the global motion was, consequently, smaller than a specified threshold, such a scene was classified as close-up, base, or other scenes. Because the players' faces take up a certain percentage in close-up and base scenes, skin color detection was used to classify whether the shot was a close-up, base, or other scenes. Here, the method in [17] was applied to detect the skin color. Given the probability distributions for skin and non-skin color, a pixel classifier could be established via the standard likelihood ratio approach. A pixel was labeled as skin if

$$\frac{\mathrm{p(rgb\,|\,skin)}}{\mathrm{p(rgb\,|\sim skin)}} \geq \theta \qquad (8)$$

where $0 < \theta < 1$ was a threshold value that depended upon the application-specific cost of classification error, as well as on the prior probabilities of skin and non-skin.

3.6 Classification of Close-up, Base, Running, and Other Scenes

When the area ratio of soil and grass regions was low or medium, global motion information, object size, and skin color were used to classify the close-up, base, running and other scenes. Firstly, if the magnitude of the global motion was larger than a specified threshold, then the shot was classified as a running scene. Here, the threshold value for the magnitude of global motion $|\boldsymbol{a}|$ was set to be 7. Secondly, skin color was used to determine whether it belonged to the other scenes or not. If the number of classified skin pixels was less than a predefined threshold T_s (400), then the scene would be classified as the other scene. On the other hand, if the number of skin pixels was more than Ts, the video shot was allocated to base or close-up scenes. Finally, the number of objects was used to differentiate between base and close-up scenes. If the key frame of a shot had more than two players in the field, then the scene would be categorized as a base scene. The block diagram for classifying the close-up, base, running, and other scenes is shown in Fig. 8.

4. General Baseball Event Detection

The main objective of this study was to develop a scene-based baseball detection system in order to detect and retrieve video clips of general baseball events more efficiently. Based on the classified semantic scenes that were regarded as the observation symbols in the HMM, a 4-state ergodic HMM was applied to detect the general baseball events including the base hit, air out, ground out, and strikeout.

4.1 Training Process for Baseball Event Detection

Two kinds of hidden Markov models are frequently used for pattern classification or recognition applications. Fig. 9-(a) illustrates a left-right model in which the state transition proceeds in a left to right direction. Fig. 9-(b) illustrates the ergodic model in which the state transition proceeds from a state to all the other states including itself. Since the semantic scenes in a baseball video may change irregularly, the ergodic HMM is more suitable for identifying the general baseball events.

In HMM, vector quantization or K-means are frequently used to generate observation symbols by clustering the acquired features. Here, the classified semantic scenes were regarded as the observation symbols in the HMM to develop a baseball event detection system. Generally, HMM is expressed by a 3-tuple parameters λ=($\mathbf{A}$, $\mathbf{B}$, π) [19], where $\boldsymbol{\pi}$ is the initial state distribution, $\mathbf{A}=[a_{\text{ij}}]$ is the matrix of state transition probabilities with element a_{ij} being the state transition probability from state i to state j, and $\mathbf{B}=\{b_{\text{j}}(\text{k})\}$ the observation symbol probability distribution for symbols o_k in state j. By training the parameters ($\mathbf{A}$, $\mathbf{B}$, π) for each baseball event, the four kinds of baseball events may be detected.

4.2 Baseball Event Detection

So far, not much research [24] has addressed the detection of general baseball events; while a lot of research has been focused on the scene classification [1-11]. Some semantic scenes including infield, outfield, player, and pitching may consist of successive video shots. Instead of developing a shot-based method, scene transitions were used to design a state machine and to detect the baseball events. Here, four kinds of baseball video events including the base hit, air out, ground out, and strikeout were identified by using the HMM method. Based on the detected semantic scenes serving as observation symbol sequence $\mathbf{O}$, the parameters ($\mathbf{A}$, $\mathbf{B}$, π) for the 4-state ergodic HMM were estimated using the Baum-Welch algorithm [19]. Given the observation symbol sequence $\mathbf{O}$, the observation probability P($\mathbf{O}|\boldsymbol{\lambda}$) for each baseball event could be computed via the forward-backward procedure [19]. Let the observation symbol sequence with length T be denoted as $\mathbf{O}=[\text{o}_1, \text{o}_2,\ldots, \text{o}_\text{T}]$, a forward variable $\alpha_\text{t}(i)$ is defined to calculate the probability of partial observing sequence of state i at time t for the model $\boldsymbol{\lambda}$ denoted as $\text{P}(\text{o}_1\text{o}_2\ldots\text{o}_\text{t},\text{q}_\text{t}=i|\boldsymbol{\lambda})$. P($\mathbf{O}|\boldsymbol{\lambda}$) may be calculated as follows.

1) Initialization: $\alpha_i(i)=\pi_i b_i(o_1)$

2) Induction: $\alpha_{t+1} = [\sum_{i=1}^{N} \alpha_t(i) a_{ij}] b_j(o_{t+1}),\ 1 \le t \le T-1, 1 \le j \le N$

3) Termination: $P(O \mid \lambda) = \sum_{i=1}^{N} \alpha_T(i)$

All the values of initial transition probabilities a_{ij} are equally distributed with value 0.25 and all the values of observation probabilities $b_j(O)$ for all states are also equally distributed with a value of 0.125 since there are 8 semantic scenes in total, i.e., 8 observation symbols. Twenty hours of video frames were used as training data for the HMM event detection system. For each baseball event, the transitions among the semantic scenes are carefully observed.

5. Experimental Results

In the experiments, the training/test videos were recorded from the televised broadcasting programs and digitized into MPEG-I format. The frame size of the training/test video was 352×240 and the frame rate was 30 fps. Firstly, an accuracy analysis for the shot boundary detection was done. Precise video shot segmentation made the classification of semantic scenes more accurate. Secondly, the experimental results for classifying the semantic scenes were illustrated. Furthermore, a comparison of the proposed and the other methods was done. Finally, various kinds of baseball video events were identified using HMM.

5.1 Accuracy Analysis of Shot Boundary Detection

Here, several baseball videos including the CPBL (Chinese Professional Baseball League) and the ABCS (Asian Baseball Championship in Sapporo) were partitioned into large sets of semantic shots by using the method in section 2. Fig. 10 illustrates the discontinuity values during the shot boundary detecting process. The parameter N in Eq. (3) was set to 21. The accuracy analysis of the shot boundary detection is given in Table 1. The total video length is 1 hour 41 minutes 10 seconds and the number of segmented shots is 637. In Table 1, the column "Number of shots" denotes the number of shots observed by the human eye. The accuracy of the shot boundary detection using the method mentioned in section 2 was analyzed using the following formula:

$$\text{Accuacy} = \frac{\text{Number of detected shots - Number of false detected shots}}{\text{Nnumber of shots}}.$$

The accuracy for the shot boundary detection was about 96%. In the following, the key frame for each shot was determined by using the median filtering method and then the objects in the key frame were extracted for the scene classification process. Fig. 11-(a) illustrates the first frame for the shot of a pitching scene and Fig. 11-(b) illustrates the extracted key frame for this shot. Fig. 11-(c) gives that the segmented objects obtained by using the MFRSST and Fig. 11-(d) gives the results after the object labeling process.

5.2 Classification of Semantic Scenes

In this study, there were eight semantic scenes to be detected namely pitching, running, base, player, infield, outfield, close-up, and other scenes. The rules for classifying the semantic scenes in a baseball video are given in Fig. 3. The eight semantic scenes were classified according to classification rules and shown in Fig. 12(a)-(h). Furthermore, the visual features that were used to classify these semantic scenes are illustrated in Table 2. The accuracy analysis for the scene classification is given in Table 3. Here, the PRECISION and RECALL were applied to analyze the accuracy of the scene classification. The RECALL ratio was defined as the ratio of the number of relevant records retrieved to the total number of relevant records in the database, and the PRECISION was defined as the ratio of the number of relevant records retrieved to the total number of irrelevant and relevant records retrieved. Furthermore, several scene classification methods including Hua's method [4], Pei's method [9], and the method proposed in this paper are compared and illustrated in Table. 4. By applying the object-based features and global motion information, the scenes were classified into more categories and the PRECISION and RECALL are improved significantly.

5.3 General Baseball Event Detection

In this study, four baseball events including the base hit, air out, ground out, and strikeout were detected. Based on the classified semantic scenes that served as the observation symbols in the HMM, the 4-state ergodic HMM was used to detect the four baseball events. Here, three baseball games in the Asia Championship 2003 were used namely the Chinese vs. Korea, Chinese vs. Japan, and Chinese vs. China games as the test data to analyze the accuracy of general baseball event detection. The HMM was trained with the baseball videos whose length is about 19 hours 43 minutes. Figures 13-16 show the scene change for the four detected baseball events.

When a sequence of symbols (classified semantic scenes) were observed by the HMMs of the four general baseball event detectors, the observation probability $P(\mathbf{O}|\boldsymbol{\lambda})$ for each baseball event was calculated. When the observation probability $P(\mathbf{O}|\boldsymbol{\lambda})$ for a certain general baseball event was higher than the others and lager than a specified probability threshold, then a general baseball event was detected and the event boundary was determined too. The next general baseball event is detected continuously. The accuracy analysis of the baseball event detection is given in Table 5. The number of mis-detected events denotes the number of events that were not detected in the test video clips. The number of false-detected events denotes the number of events that were detected incorrectly. It is obviously that the detection rate of the “ground out” event was lower than the others because there were similar scene changes in both the “ground out” and “base hit” events. The average precision for detecting the four general baseball events was about 83% and the average recall was about 90%.

Furthermore, Han’s method [24] was compared with the proposed method. In Han’s work several multimedia features including image, audio, and text features were integrated to detect several baseball events (home run, outfield hit, outfield out, infield hit, infield out, strike out, and walk). The accuracy analysis for Han’s work is shown in the Table 6. The average recall ratio and precision for Han’s method were about 71% and 79%, respectively. It is obvious that the proposed method for general baseball events outperforms Han’s method. However, the number of detected baseball events was lager than with the method proposed in this paper. Hence, in future works, the system would be improved to detect more baseball events.

6. Conclusion

Generally, precise event detection is complicated in video processing research. In this study, we tried to use as many visual features as possible to classify the semantic scenes and then applied HMM to identify the baseball events. Four kinds of baseball events including the base hit, strikeout, ground outs, and air outs were detected with the proposed system. Firstly, the sport video was partitioned into many video shots. Secondly, the key frame for each video shot was determined and then some visual features in the video shot, e.g., soil and grass color percentage, object number, motion vector, skin detection, player location etc., were extracted to classify the semantic scenes. Finally, HMM was used to detect the various kinds of baseball events. Experimental results show that the proposed system can detect the four kinds of baseball events accurately. The average PRECISION for detecting the four general baseball events was 83.25% and the average RECALL was 90%.

Acknowledgements

This work had been partially funded by Chung Hua University under project no. CHU-94-TR-02 and National Science Council under project no. NSC 94-2213-E-216 -023.

Reference

[1] B. Li, M. I. Sezan, Event detection and summarization in sports video, Proceedings, IEEE Workshop on Content-Based Access of Image and Video Libraries 2001 (CBAIVL 2001), Dec. 2001, pp. 132 – 138.

[2] F. Cheng, W. J. Christmas, J. Kittler, Detection and description of human running behavior

in sports video multimedia database, Proceedings, 11th International Conference on Image Analysis and Processing, Sept. 2001, pp. 26-28.

[3] H. Pan, B. Li, M. I. Sezan, Automatic detection of replay segments in broadcast sports programs by detection of logos in scene transitions, Proceedings, IEEE International Conference on Acoustics, Speech, and Signal Processing, May 2002, 4, pp. 13-17.

[4] W. Hua, M. Han, Y. Gong, Baseball scene classification using multimedia features, Proceedings, IEEE International Conference on Multimedia and Expo, Aug. 2002, 1, pp. 821-824.

[5] S. Miyauchi, A. Hirano, N. Babaguchi, T. Kitahashi, Collaborative multimedia analysis for detecting semantical events from broadcasted sports video, Proceedings, 16th International Conference on Pattern Recognition, Aug. 2002, 2 , pp. 11-15,

[6] Y. Rui, A. Gupta, A. Acero, Automatically extracting highlights for TV baseball programs, Proceedings, Eighth ACM International Conference on Multimedia, 2000, pp. 105-115.

[7] A. Ekin, A. M. Tekalp, R. Mehrotra, Automatic soccer video analysis and summarization, IEEE Trans. on Image Processing, 12(2003) 796-807.

[8] H. Lu, Y.-P. Tan, Sports video analysis and structuring, Proceedings, IEEE Fourth Workshop on Multimedia Signal Processing, Oct. 2001, pp. 3-5.

[9] S. C. Pei, F. Chen, Semantic scenes detection and classification in sports videos, Proceedings, 16th IPPR Conference on Computer Vision, Graphics and Image Processing, Aug. 2003.

[10] N. Haering, R. J. Qian, M. I. Sezan, A semantic event-detection approach and its application to detecting hunts in wildlife video, IEEE Trans. on Circuits and Systems for Video Technology, 10(2000), 857-868.

[11] G. L. Foresti, L. Marcenaro, C. S. Regazzoni, Automatic detection and indexing of video-event shots for surveillance applications, IEEE Trans. on Multimedia, 4(2002), 459-471.

[12] C. H. Chen, C. C. Lien, The multiresolution and flooding based RSST (MFRSST) image segmentation method, Chung Hua Journal of Science and Engineering, 1(2003), 9-16.

[13] W. J. Heng, K. N. Ngan, Shot boundary refinement for long transition in digital video sequence, IEEE Trans. on Multimedia, 4(2002), 434-444.

[14] U. Gargi, R. Kasturi, S. H. Strayer, Performance characterization of video-shot-change detection methods, IEEE Trans. on Circuits and Systems for Video Technology, 10(2000), 1-13.

[15] A. Hanjalic, Shot-boundary detection: unraveled and resolved?, IEEE Trans. on Circuits and Systems for Video Technology, 12(2002), 90-105.

[16] S. X. Ju, M. J. Black, S. Minneman, D. Kimber, Summarization of videotaped presentations: automatic analysis of motion and gesture, IEEE Trans. on Circuits and Systems for Video Technology, 8(1998), 686-696.

[17] M. J. Jones, J. M. Rehg, Statistical color models with application to skin detection, Proceedings, IEEE International Conference on Computer Vision and Pattern Recognition, 1(1999), 23-25.

[18] Y. Wang, J. Ostermann, Y. Zhang, Video Processing and Communications, Prentice-Hall, 2002.

[19] L. Rabiner, B.-H. Juang, Fundamentals of speech recognition, Prentice-Hall 1993.

[20] S. H. Kwok, A. G. Constantinides, A fast recursive shortest spanning tree for image segmentation and edge detection, IEEE Trans. on Image Processing, 6(1997), 328-332.

[21] L. Vincent, P. Soille, Watersheds in digital spaces: An efficient algorithm based on immersion simulations, IEEE Trans. on Pattern Analysis and Machine Intelligence, 13(1991), 583-598.

[22] O. J. Morris, M. J. Lee, A. G Constantinides, Graph theory for image analysis: an approach base on the shortest spanning tree, Proc. Inst. Elect. Eng., 133(1986), 146-152.

[23] E. R. Davies, Machine vision- theory, algorithms, practicalities 3rd Edition, Elsevier, 2005.

[24] M. Han, W. Hua, W. Xu, and Y. Gong, "An integrated baseball digest system using maximum entropy method," Proc. of ACM Multimedia, pp. 347-350, 2002.

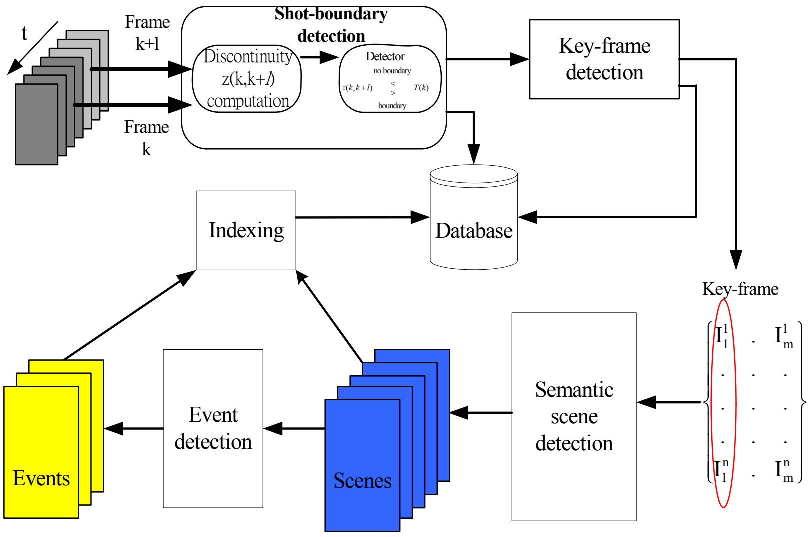


Fig. 1 Block diagram of the baseball event detection system.

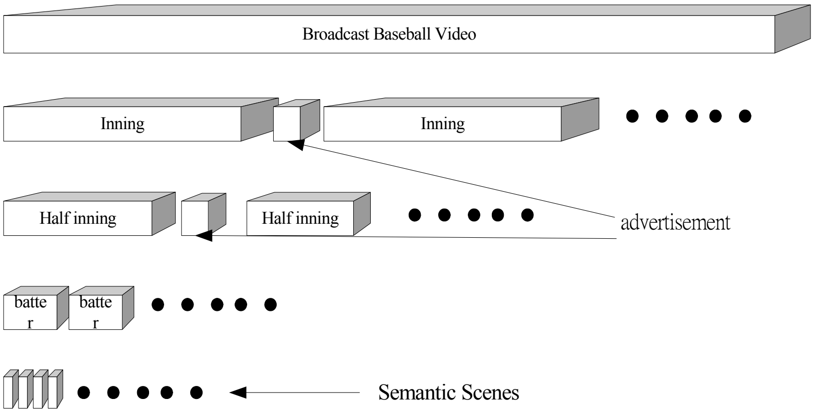


Fig. 2 The structure of a baseball video.

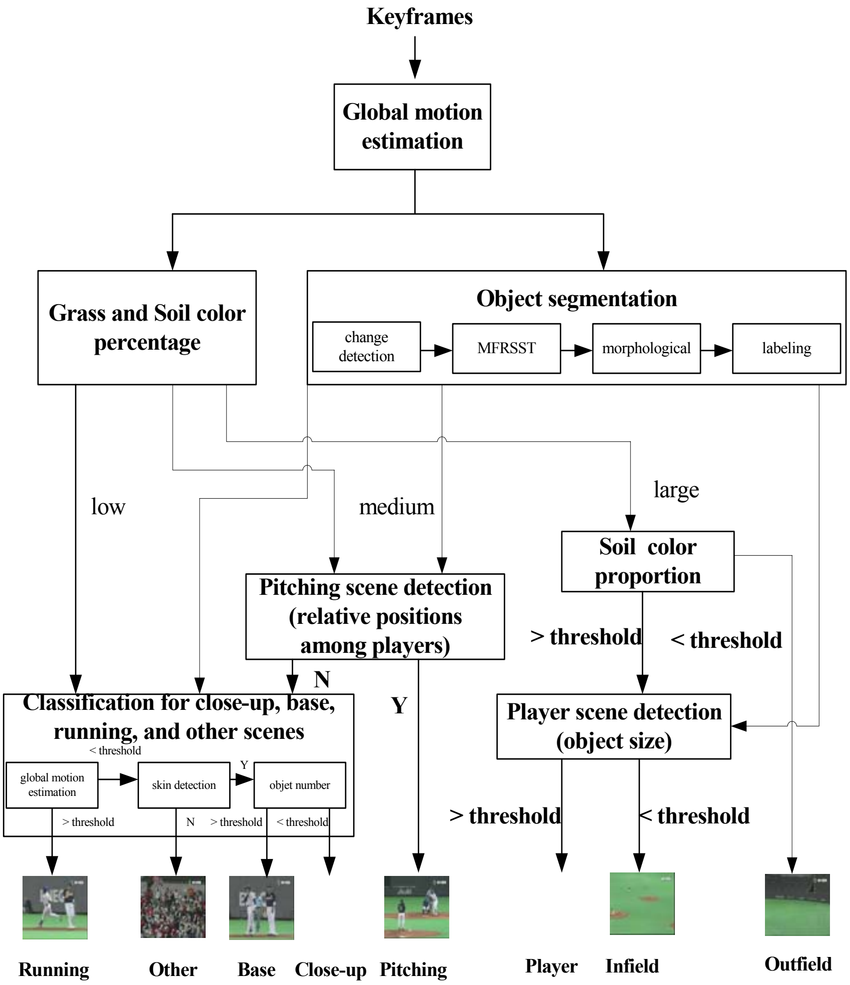
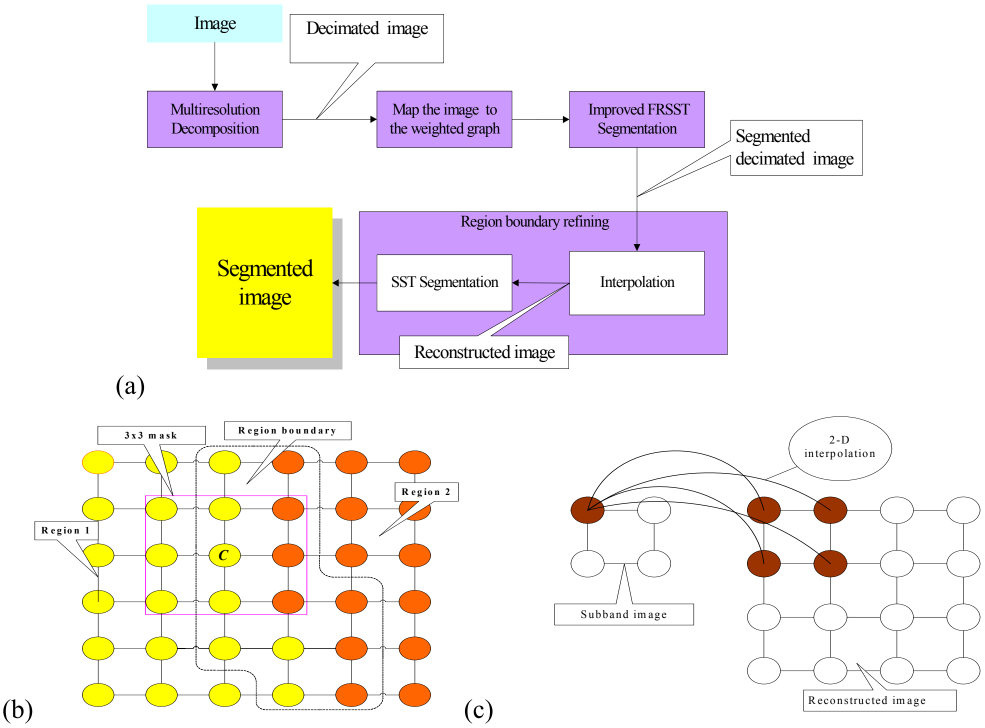
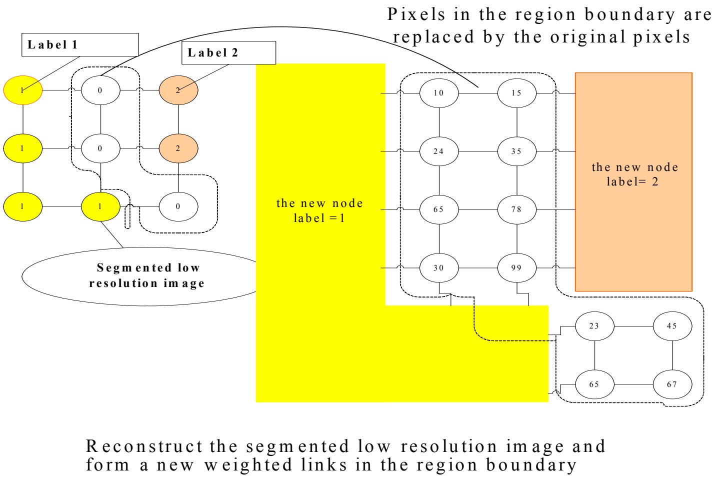


Fig. 3 Flowchart of classifying the semantic scenes.




Pixels in the region boundary are replaced by the original pixels
Label 1
Label 2
0
2
1
1
1
1
0
0
2
0
Segmented low resolution image
the new node label = 1
10
15
24
35
65
78
30
99
the new node label= 2
23
45
65
67
Reconstruct the segmented low resolution image and form a new weighted links in the region boundary

(d)

Fig. 4 (a) Block diagram of the MFRSST image segmentation system. (b) Region boundary detection. (c) 2-D interpolation process. (d) Each reconstructed region is regarded as a single node and the pixels in the region boundary are replaced by the original image pixels. After that, a new weighted graph is constructed.

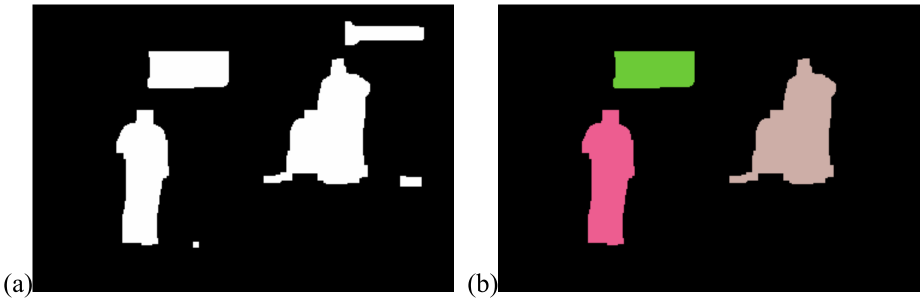
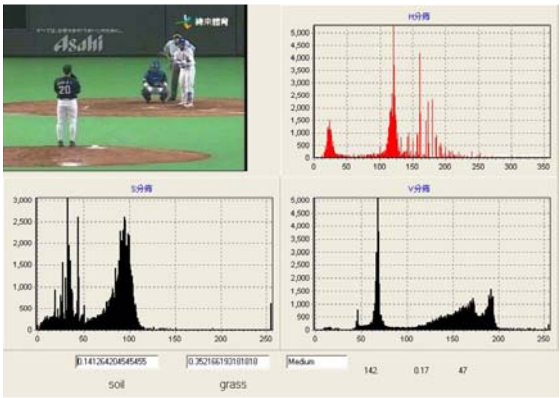


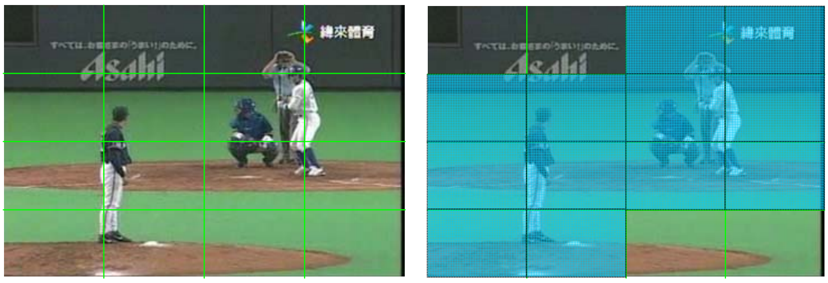
Fig. 5 (a) Objects are segmented by using MFRSST method. (b) Objects are extracted by using the labeling method.



0.141264204545455 0.352166193181818 Medium 142 0.17 47

soil grass

Fig. 6 Color distributions for the soil and grass regions.



(a) (b)

Fig. 7 (a) 16 partitioned blocks used to extract the relative players' position. (b) Relative positions of the players in the pitching scene.

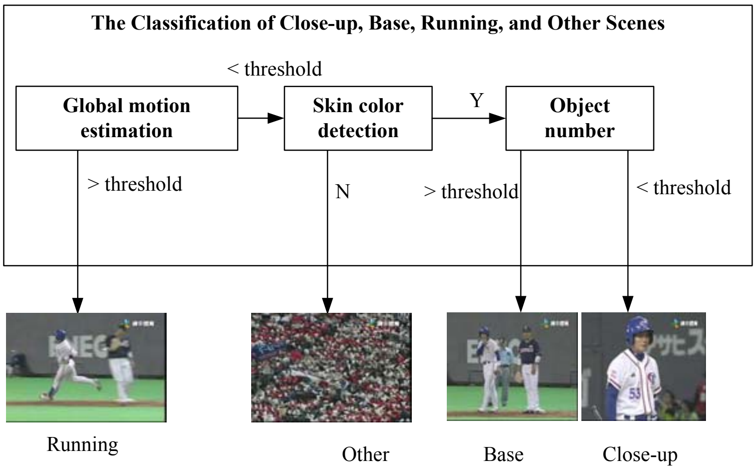


Fig. 8 Block diagram for classifying the close-up, base, running, and other scenes.

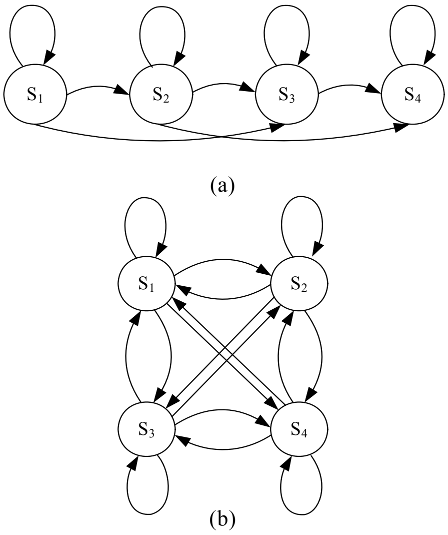


Fig. 9 Illustration of two hidden Markov models. (a) Left-right model. (b) Ergodic model.

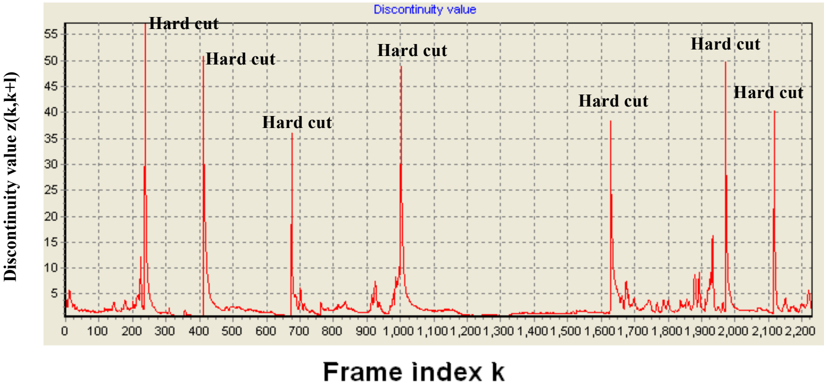


Fig. 10 Discontinuity values at the shot boundary positions.

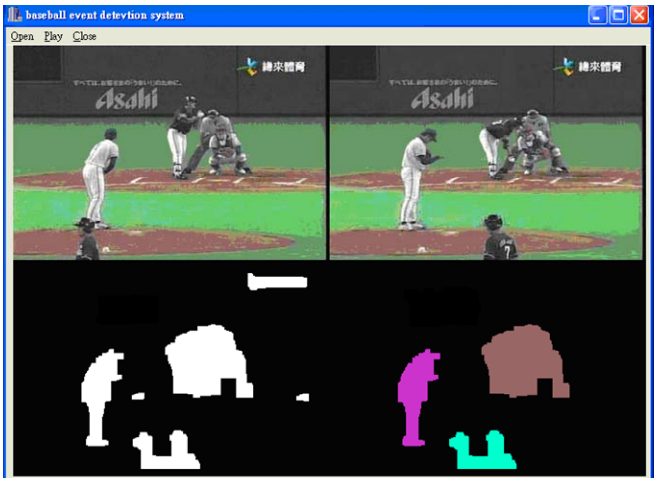


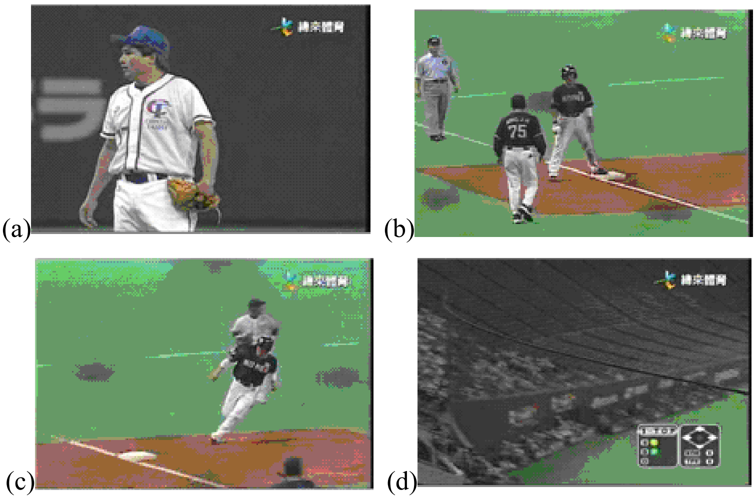
Fig. 11 (a) The first frame for the shot of a pitching scene. (b) The extracted key frame for this shot. (c) The segmented objects obtained by using the MFRSST. (d) The results after the object labeling process.

(a)

Table 1 Accuracy analysis for the shot boundary detection

	Video length	Number of shots	Number of detected shots	Number of miss-detected shots	Number of false detected shots
Clip 1	11’39”	68	67	1	0
Clip 2	6’09”	40	38	2	0
Clip 3	8’24”	49	49	0	0
Clip 4	9’”09	34	31	3	0
Clip 5	13’48”	84	82	5	3
Clip 6	19’28”	120	116	6	2
Clip 7	12’10”	90	92	1	3
Clip 8	20’33”	152	150	7	5
Total	101’10”	637	625	25	13

(c)



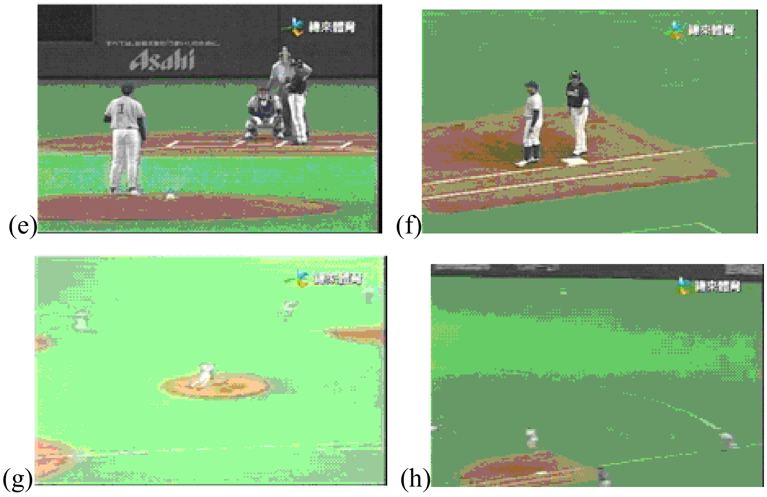


Fig. 12 Eight semantic scenes. (a) Close-up scene. (b) Base scene. (c) Running scene. (d) Other scene. (e) Pitching scene. (f) Player scene. (g) Infield scene. (h) Outfield scene.

Table 2 Visual features for classifying the semantic scenes shown in Fig. 12 (a)-(h).

	Grass field percentage	Soil field percentage	Field percentage	Motion magnitude $\|a\|$(Eq.8)	Skin color	Object number	Play's location	Object size
(a)	1.0%	2.4%	Low	3	Yes	1		
(b)	32.8%	7.5%	Medium	3	Yes	2	No	
(c)	30.0%	12.0%	Medium	7			No	
(d)	11.5%	0.0%	Low	1	No			
(e)	38.7%	15.1%	Medium				Yes	
(f)	64.2	13.2%	Large					12%
(g)	83.6	5%	Large					
(h)	85.7	1.3%	Large					

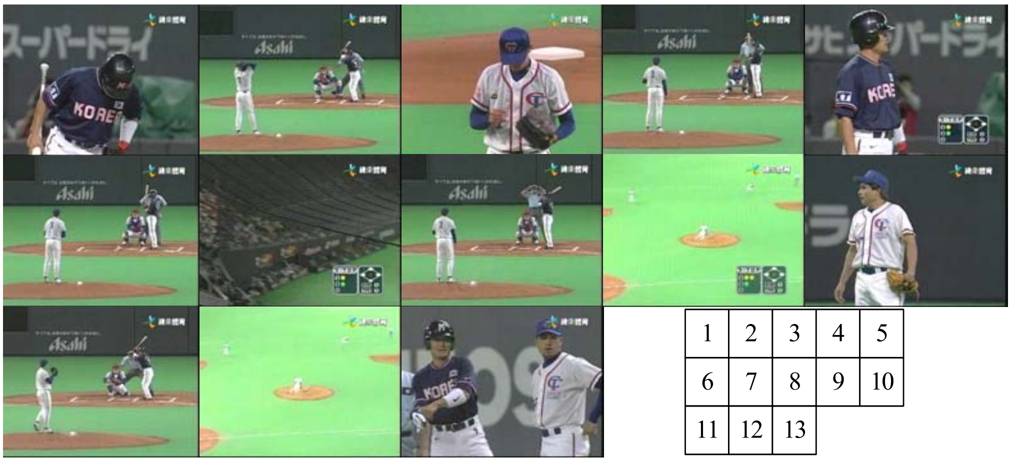
Table 3 Accuracy analysis for the semantic scene classification.

	Number of observed scenes	Number of detected scenes	Number of mis-detected scenes	Number of false-detected scenes	Recall ratio	Precision
Close-up	115	107	10	2	91%	90%
Base	31	30	2	1	94%	91%
Running	12	9	3	0	75%	75%
Other	30	40	0	10	100%	75%
Pitching	219	224	0	5	100%	98%
Player	17	12	5	0	71%	71%
Infield	37	42	0	5	100%	88%
Outfield	24	23	1	0	96%	96%

Table 4 The comparison for the scene classification among several methods

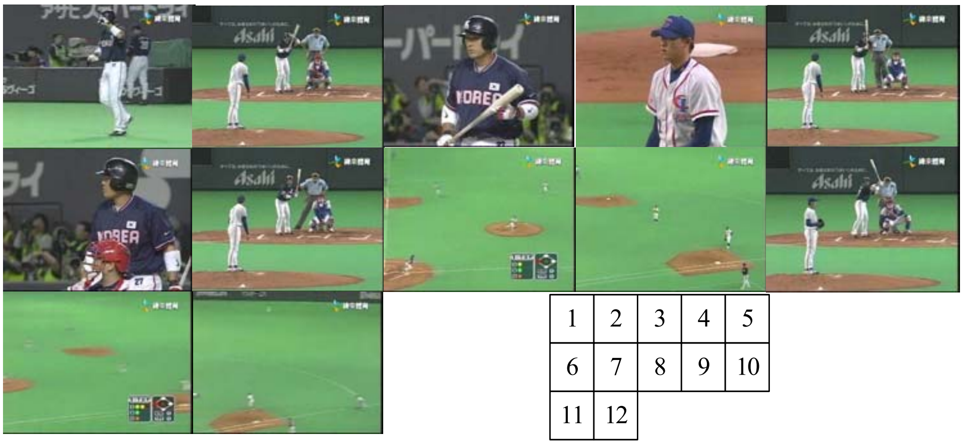
	Hua's method[4]		Pei's method[9]		The proposed method	
Scenes	Recall ratio	Precision	Recall ratio	Precision	Recall ratio	Precision
Close-up	39%	51%	60%	86%	91%	90%

Base	34%	42%			94%	91%
Running	35%	34%			75%	75%
Other					100%	75%
Pitching	92%	89%	94%	94%	100%	98%
Player					71%	71%
Infield	85%	88%	90%	92%	100%	88%
Outfield	72%	70%	85%	90%	96%	96%



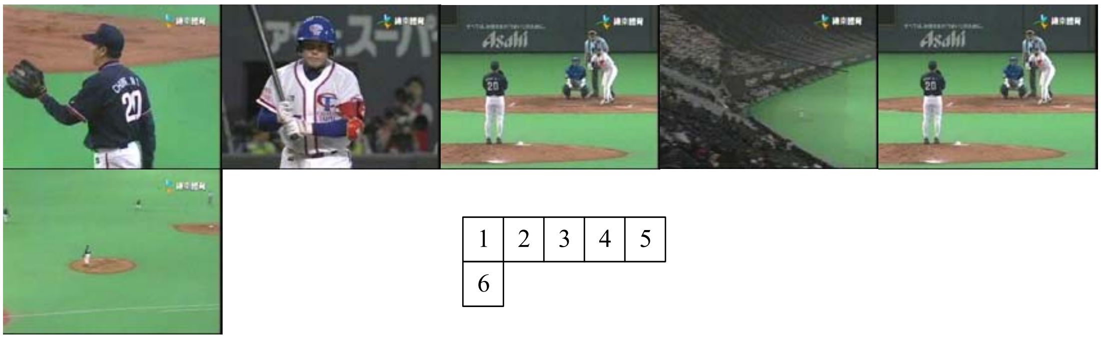
1	2	3	4	5
6	7	8	9	10
11	12	13		

Fig. 13 Scene changes for the detected base hit event.



1	2	3	4	5
6	7	8	9	10
11	12			

Fig. 14 Scene changes for the detected air out event.



1	2	3	4	5
6				

Fig. 15 Scene changes for the detected ground out event.

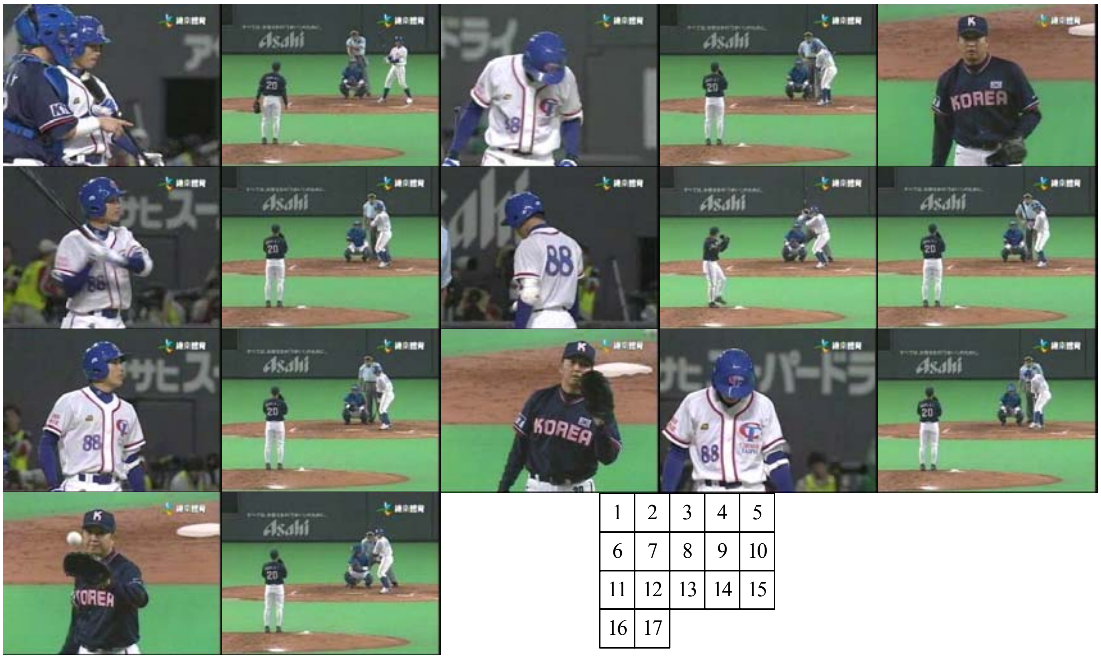


Fig. 16 Scene changes for the detected strikeout event.

Table 5 Accuracy analysis for the proposed baseball event detection.

General baseball events	Number of events	Number of detect events	Number of mis-detected events	Number of false-detected events	Recall	Precision
base hit	11	10	1	0	91%	91%
ground outs	13	11	3	1	77%	71%
air outs	12	11	1	0	92%	92%
strikeout	15	19	0	4	100%	79%

Table 6 Accuracy analysis for Han's work [24].

	Total	Correct	Misclassified	Missed	False alarm	Recall	Precision
home run	3	2	0	1	0	67%	100%
outfield hit	13	11	2	0	2	85%	73%
outfield out	17	14	1	2	0	82%	93%
infield hit	6	3	1	2	3	50%	43%
infield out	30	26	2	2	5	87%	79%
strike out	16	11	0	5	2	69%	85%
walk	9	5	0	4	1	56%	83%

3. 參考文獻

[1] **Cheng-Chang Lien** and Chiu-Lung Chian, "Scene-Based Event Detection for Baseball Videos," to appear in Journal of Visual Communication and Image Representation. **(SCI, EI) (NSC 94-2213-E-216 -023 -)**

[2] Cheng-Chang Lien and Chiu-Lung Chian, "A Scene-Based General Baseball Event Detection System," International MultiConference of Engineers and Computer Scientists 2006 (IMECS 2006), June, Hong Kong, pp. 507-512. **(NSC 94-2213-E-216 -023 -)**

4. 計畫成果自評

The research of this project (NSC 94-2213-E-216 -023 -) is to be published in the Journal of Visual Communication and Image Representation indexed by SCI and EI.