數位家庭應用中內容分析與展示技術之研發(III)

計劃主持人：吳家麟　台灣大學資訊網路與多媒體研究所教授

一　摘要

(中文摘要)

本計劃以提昇數位家庭應用為目標，共分成兩大方向：(1)影片事件偵測(video event detection)與(2)複合式影音展示(composite audiovisual presentation)。分別對數位家庭中媒體接收分析與整合展示兩方面著手，適切地符合使用者對於新一代數位應用的需求。

【影片事件偵測】技術之主要目的是利用專有知識(domain knowledge)與製作原則(production rules)來輔助數位內容的分析。有別於以往的研究只能對數位資料做大致的分類或切割，本計劃希望能確切地偵測影片內事件的種類與發生的時間。本技術將從針對棒球影片開始，建立完整的分析架構。利用以規則為基礎(rule-based decision)與以模型為基礎(model-based decision)的判斷方式來實作出各種事件的偵測，並進行自動產生比賽過程的記錄與摘要。

【複合式影音展示】主要希望整合多個媒體共同展示，為不同的個別媒體增加附加價值。本技術從照片特性與音樂節奏的分析開始，建立一套架構自動產生以音樂為導向的幻燈秀。其中在照片展示方面將結合圖片分類、使用者注意模型、區域重要性模型、自動編輯與多媒體同步等功能，務求最後的展示結果能結合各照片的關連與音樂特性，以期有別於傳統賞圖軟體只能依序一張張的幻燈秀。

(英文摘要)

This project aims to enhance the digital applications at home. It focuses on two phases: (1) video event detection and (2) composite audiovisual presentation. They respectively tackle with content analysis and multimedia integration, and therefore match the needs of next-generation digital home applications.

The goal of “video event detection” is mainly on how to exploit domain knowledge and production rules to facilitate digital content analysis. Unlike conventional researches, which roughly classify or segment digital content, this project tries to exactly detect what happened and the timestamps of events. We will start with analyzing baseball videos and construct a complete framework. With rule-based decision and model-based decision methods, we implement exactly event detection and further develop the techniques of automatic scoreboard generation and game summarization.

“Composite audiovisual presentation” integrates multiple media and adds the values to isolated media. It analyzes photos characteristics and music tempo, and then builds a framework to automatically generate music-driven slide shows. We integrate the techniques of image classification, user attention model, region-based importance model, automatic editing, and multimedia synchronization. We will present the results with correlations between photos and music characteristics, instead of sequentially showing photos one-by-one.

二　計畫緣由與目的

生活數位化的時代已隨著各種數位內容產品的普及與數位電視的全面推廣而快速來臨。在面對大量數位資訊與各種媒體來源時，許多問題因為實際的應用與使用者的需求而逐漸衍生出來。為了讓大量的多媒體資料能有效率地被運用，進而提昇數位內容的應用價值，有效的多媒體檢索與數位內容

分析是急需被發展的技術。以商業附加價值大的運動影片為例，若能應用數位內容分析技術將比賽中的事件偵測出來，並自動剪輯對應的片段或進行影片摘要，將可提供使用者更多元的觀賞模式。

目前數位內容分析的研究主要面臨的課題在於如何跨越語意鴻溝(semantic gap)。以往對於數位資料的分析只著重於特徵值萃取(feature extraction)，然而，低階特徵值的特性往往跟人的感覺不同。常用的特徵值包括圖片中顏色統計(color histogram)、顏色分佈(color layout)、邊緣統計(edge histogram)等；影片中的移動(motion)、鏡頭種類(camera type)等；聲音中的過零率(zero-crossing rate)、能量(energy)、音高(pitch)、節奏(tempo)等。由於這些特徵值不直接具有人為解釋的意義，因此由這些特徵值描述而得的結果常常不如預期。此種低階特徵值與高階語意(semantics)之間的落差稱為語意鴻溝(semantic gap)。

再以運動影片為例，目前的研究成果主要在於利用顏色與邊緣資訊進行畫面分類(shot classification)(如棒球或網球影片)，或者是利用球的軌跡(ball trajectory)與移動(motion)進行比賽持續或暫停(play-break)的分析(如足球影片)。然而，這些初步的成果與使用者真正需要的相去甚遠。以球迷的觀點來說，他們關心的是場上球員的表現：是否有安打？是否有全壘打？投手今天有幾次三振？此場比賽的射門與得分畫面等等。目前的研究成果只能大致找出比賽中“可能比較精采的片段”，而不是“比賽中確切發生了什麼事情”。許多受歡迎且有意義的應用雖然已經開始出現(如美國大聯盟網站上的線上影片)，它們仍需依靠人為的剪輯，需要大量的時間與人力。因此，在數位媒體大量被創造的同時，發展一套能有效偵測影片裡事件的技術已經是刻不容緩。

此外，從數位電子產品的普及來看，一般的使用者越來越有能力錄製或創造自己的數位內容。這些數位資料雖然可以很輕易地被創造與儲存，但由於一般使用者常疏於管理或因為拍攝品質不佳，降低了這些數位資料的價值與實用性。在這項課題當中，以往的研究主要提供使用者一個大致的分類結果，例如，以場景的相似性將類似的照片歸為同一類，或以相機或攝影機附帶的時間資訊做時間上的分類。這樣的做法在近年來也開始往內容加值的方向進行。例如，與其觀賞一段長達一兩個小時的家庭影片，不如將其重要的片段配上音樂，剪輯成一段五六分鐘的影音摘要。同理，數位相機拍得的大量照片也可在偵測畫面中物件的重要性與場景分類之後，配上適當的音樂製作成動態的幻燈片展示(slide show)。此種摘要式的後製資料更能被人們接受，其相對的觀賞價值也跟著提高。

為提昇數位家庭應用，本計劃將目標集中在兩大方向：影片事件偵測(event detection)與複合式影音展示(composite audiovisual presentation)。

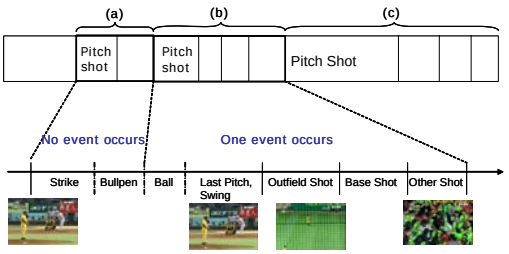
三 研究方法與成果

本計畫共以三年時間完成，第一年著重低階特徵值萃取與以規則為基礎的事件偵測。第二年進一步發展以模型為基礎的事件判斷模組與圖片、音樂分析模組。第三年整合多項模組完成影片摘要與幻燈秀的實作。以下簡介本計畫所完成之目標與整合成果。

A【影片事件偵測】

A.1 棒球視訊特徵值萃取

根據我們的觀察，棒球比賽中的事件一定發生在兩個投手畫面之間。如圖一所示的例子而言，投手投出一球但打者沒打、畫面轉往其他地方，這段期間就沒有事情發生。下一個畫面再切回投球畫面、打者把球打出去、畫面切到外野、打者站上一壘、然後畫面再切回投球畫面，這段期間就發生了對比賽狀態有確實影響的事件。



圖一、棒球事件進行範例

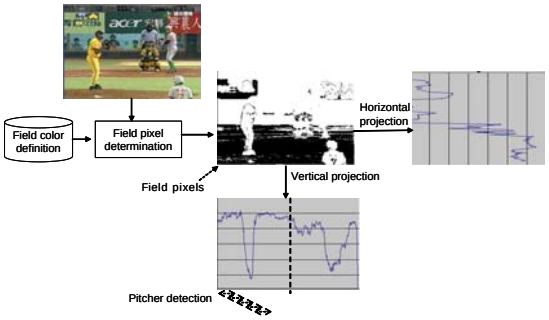
根據這樣的觀察，我們利用加註在投手畫面

的 caption information 的變化來推斷在連續兩個投手畫面之間發生了什麼事情。藉由偵測兩連續投手畫面中的字幕，我們可以監看比賽狀況的變化(包含出局數、佔壘數與得分)，並進而依棒球規則推測何種事件發生。

舉個例子, 如果第 *i* 個投手畫面上是無人出局、一壘有人、無得分；到了第 *i*+1 個投手畫面無人出局、無人在壘、但是卻多了兩分， 那麼我們就可以推測這段期間發生了全壘打。因為根據出局變化我們知道有安打產生。根據分數變化跟佔壘變化，這樣的情況必定是得到兩分的全壘打。基於這些觀察，我們在事件偵測之前必須先辨別畫面種類，然後辨識在投手畫面中的字幕資訊。以下分別就兩項工作做詳細說明。

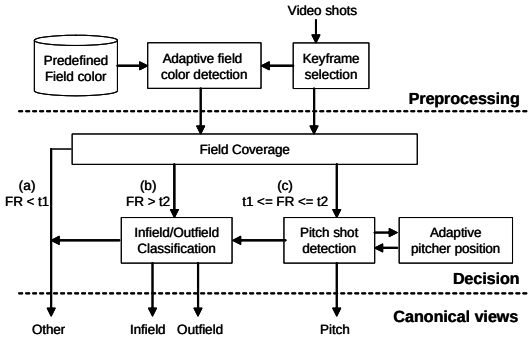
A.1.1 畫面分類

分析的第一個步驟是將視訊資料分成一個個場景(shot)。在此我們利用兩張畫框(video frame)之間的 histogram 差距來判斷是否為新場景的依據。在做完場景界限偵測(shot boundary detection)之後，我們對每個場景取出一張關鍵畫面(keyframe)。接著，根據畫面中場地顏色的分佈與位置資訊來判斷此畫面是否為投手畫面。如圖二所示，投手畫面中場地顏色主要集中在下半部。此外，由於轉播視角的關係，畫面左半部份的場地顏色分佈通常有一個由投手站位造成的凹陷。



圖二、投手畫面偵測

另一方面，如果是內野或外野，場地顏色將佔畫面的大部份。由於外野通常包含較多由觀眾或球場建築物造成的結構，因此我們可進一步根據畫面中的邊緣資訊(edge information)將內野與外野區分開。若場地顏色所佔顏色非常低，那通常是球員特寫或場邊等不相關的畫面造成。由圖三的流程所示，基於這兩個顏色與空間的資訊，我們可將任一畫面分類成投手、內野、外野及其他畫面。

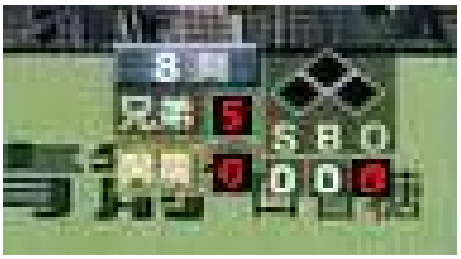


圖三、畫面分類流程

A.1.2 字幕資訊辨識

在投手畫面中我們將從字幕上的資訊來了解比賽狀態的變化。我們所關切的資訊包括出局數、分數與佔壘情況，如圖四所示。首先，因為字體出現的地方通常有較高的亮度值，我們依照像素的亮度將字塊區域內的像素做二元化(binarization)。接著針對此二元化區取 Zernike moments [1]當作此區域的特徵值。由於此特徵值具有 rotation invariant 與 translation invariant 的特性，因此很適合用於字體辨識。

基於此特徵值與字庫中的字體比對，我們可辨識出分數與出局數。此外，佔壘情況可直接根據壘包區域的亮度變化來判斷(亮度高代表壘包上有人)。

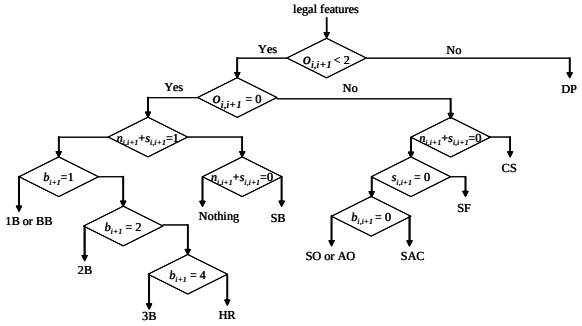


圖四、字幕資訊

A.2 以規則為基礎的棒球事件偵測模組

由於棒球比賽的結構明確，且比賽的進行嚴

格遵守棒球規則，因此我們可利用規則與比賽狀態變化來推測出事件的種類。我們可將棒球比賽規則轉換成為 decision tree，如圖五所示。在計算出兩投手畫面之間的資訊變化後，我們可順著此 decision tree 判斷出大部份的棒球事件。舉例來說，如果兩投手畫面間沒有出局變化，沒有分數變化，但在第二個投手畫面中顯示二壘有人，則此段時間我們可推測有個二壘安打產生。

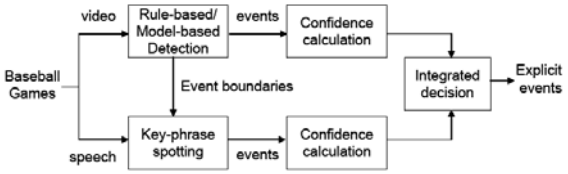


圖五、事件偵測的 decision tree

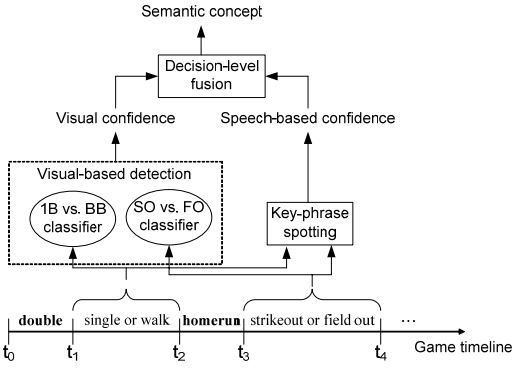
A.3 比賽規則與慣例於以模型為基礎(model-based)的事件判斷模組

利用比賽規則與慣例的方法雖然可偵測球賽中的大部份事件。但有些事件光靠字幕資訊與規則仍然無法明確區隔，例如一壘安打與四壞球保送；三振與內外野出局等。因此，我們現階段再進一步地加入畫面切換與語音的資訊來區分這些混淆的狀況。

如圖六所示，我們首先分別從視覺跟語音資訊做偵測，之後再結合這兩方面的資訊以得到一個整合偵測的結果[1]。如圖七所示，在時間 t_1 到 t_2 之間，經由字幕變化與規則判斷，我們僅能大略知道這段期間出現一壘安打或四壞球。在這種情況下，此區段的資料將分別進行以視訊為基礎的偵測(Visual-based detection)與以語音為基礎的偵測(Speech-based detection)。以視訊為基礎的偵測模組以“是否有投手畫面緊接內外野畫面”、“在打擊出去前投手總共投了幾球”、以及“攝影機移動程度(motion magnitude)”等為特徵值，利用 K-nearest modeling 來描述一壘安打或四壞球的特性。以語音為基礎的偵測模組利用“關鍵字詞偵測(key-phrase spotting)”[2]來評估某個概念發生的機率。



圖六、視覺跟語音資訊的偵測整合

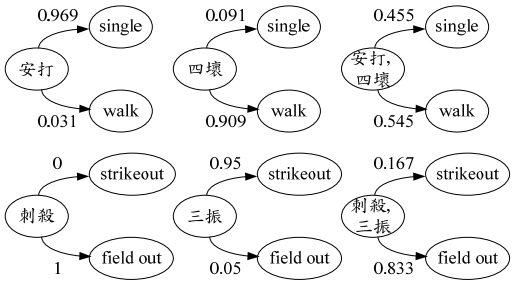


圖七、結合語音關鍵字句的結果來進行混淆概念的區分

由於特定概念的發生將伴隨著特定的主播語音說明，如表一所示，因此我們可根據資料庫中的比賽來評估各個關鍵字詞出現時某特定概念出現的機率。據本研究所統計的結果，關鍵字詞與相關概念的機率關係如同圖八所示。另外，三振與內外野出局的區分狀況亦同。

Concepts	Corresponding Key-phrases
Single (C_1)	R_1={安打(hit), 一壘安打(single)}
Walk (C_2)	R_2={觸身球(hit by pitch), 保送(walk), 四壞球(four balls)}
Strikeout (C_3)	R_3={三振(strikeout), 三振出局(strikeout)}
Field out (C_4)	R_4={刺殺('touch out' or 'out before reaching bases'), 接殺(catch out)}

表一、常見的關鍵字詞與相對應的概念



圖八、關鍵字詞與相關概念的機率關係

在分別進行完兩種不同的偵測之後，我們應用結合不同分類器的概念將兩者的意見融合起來。我們可選擇利用以下規則來整合多個分類器的結果。相加規則(sum rule)：

$$\text{assign } Z \to C_j \text{ if } \sum_{i=1}^{2} P(C_j \mid \mathbf{x}_i) = \max_{k=1}^{2} \sum_{i=1}^{2} P(C_k \mid \mathbf{x}_i)$$

相乘規則(product rule)：

$$\text{assign } Z \to C_j \text{ if } \prod_{i=1}^{2} P(C_j \mid \mathbf{x}_i) = \max_{k=1}^{2} \prod_{i=1}^{2} P(C_k \mid \mathbf{x}_i)$$

取大值規則(max rule)：

$$\text{assign } Z \to C_j \text{ if } \max_{i=1}^{2} P(C_j \mid \mathbf{x}_i) = \max_{k=1}^{2} \max_{i=1}^{2} P(C_k \mid \mathbf{x}_i)$$

取小值規則(min rule)：

$$\text{assign } Z \to C_j \text{ if } \min_{i=1}^{2} P(C_j \mid \mathbf{x}_i) = \max_{k=1}^{2} \min_{i=1}^{2} P(C_k \mid \mathbf{x}_i)$$

經過這些步驟之後，本研究所發展的系統可以將棒球概念準確地偵測出來。此研究的成果可精確偵測十三種不同的棒球概念而非僅只於大約找出比賽中的精采片段。這十三種棒球概念分別是：一壘安打(single, 1B)、二壘安打(double, 2B)、三壘安打(triple, 3B)、全壘打(home run, HR)、盜壘(stolen base, SB)、盜壘阻殺(caught stealing, CS)、內外野飛球接殺(fly out, AO)、三振(strikeout, SO)、四壞球(base on ball, Walk, BB)、犧牲觸擊(sacrifice bunt, SAC)、犧牲高飛(sacrifice fly, SF)、雙殺(double play, DP)、以及三殺(triple play, TP)。表二列出五大類最常出現於棒球比賽中的概念，以 CPBL(Chinese Professional Baseball Leagues)[3] 與 MLB(Major League Baseball)的比賽為例，其偵測結果具有非常高的準確度。

Game		Hit/BB	Double	Home Run	Out	Sacrifice	Double Play
CPBL1	Pre.	1 (15/15)	1 (6/6)	1 (2/2)	1 (35/35)	1 (5/5)	1 (3/3)
	Rec.	1 (15/15)	1 (6/6)	1 (2/2)	0.85 (35/37)	1 (5/5)	1 (3/3)
CPBL2	Pre.	1 (15/15)	1 (3/3)	1 (2/2)	1 (34/34)	1 (4/4)	0.75 (3/4)
	Rec.	0.83 (15/18)	1 (3/3)	1 (2/2)	0.89 (34/38)	1 (4/4)	1 (3/3)
CPBL3	Pre.	1 (17/17)	1 (3/3)		0.98 (43/44)	1 (2/2)	1 (2/2)
	Rec.	0.89 (17/19)	1 (3/3)		0.91 (43/47)	1 (2/2)	1 (2/2)
MLB	Pre.	1 (18/18)	1 (3/3)		1 (25/25)	0.67 (4/6)	0.75 (3/4)
	Rec.	0.95 (18/19)	1 (3/3)		0.81 (25/31)	1 (4/4)	0.75 (3/4)
Total	Pre.	1 (65/65)	1(15/15)	1 (4/4)	0.99(137/138)	0.83(15/17)	0.85(11/13)
	Rec.	0.92(65/71)	1(15/15)	1 (4/4)	0.90(137/153)	1(15/15)	0.92(11/12)

表二、四場不同比賽的概念偵測結果，以 precision 與 recall 表示

另外，對於混淆概念的區分(discrimination)，由

Games	Decision	Precision / Recall	F1
Lions vs. Bears	Visual	0.88 / 0.82	0.85
	Visual + speech	0.96 / 0.89	0.92
Bulls vs. Lions	Visual	0.76 / 0.68	0.70
	Visual + speech	0.85 / 0.74	0.79
Lions vs. Whales	Visual	0.77 / 0.73	0.75
	Visual + speech	0.93 / 0.88	0.90

表三可以看到，在大多數的情況下，同時整合視訊與語音偵測的結果會有最好的區分效能。在此圖中我們以 F1 值來表示效能，它代表的是同時考慮 precision 與 recall 的一種效能指標：

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision+Recall}}$$

Games	Decision	Precision / Recall	F1
Lions vs. Bears	Visual	0.88 / 0.82	0.85
	Visual + speech	0.96 / 0.89	0.92
Bulls vs. Lions	Visual	0.76 / 0.68	0.70
	Visual + speech	0.85 / 0.74	0.79
Lions vs. Whales	Visual	0.77 / 0.73	0.75
	Visual + speech	0.93 / 0.88	0.90

表三、概念區分的整體效能分析

A.4 影片事件偵測技術整合與延伸應用

基於第一年與第二年之研究成果，我們可準確地偵測概念，並在第三年中發展許多更有用也更準確的延伸應用，如自動產生比賽記錄表或比賽摘要

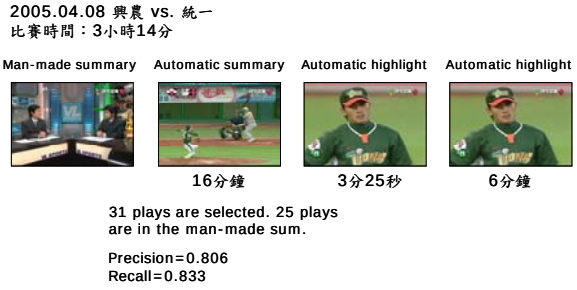
等。在自動產生摘要時，我們將概念的重要性分成五類：

- Rank 1：使比賽勝負狀態改變的概念。包括讓 A 隊領先 B 隊、A 隊追平 B 隊、或 A 隊被 B 隊趕過的情況。
- Rank 2：帶有打點的安打或盜壘。雖未造成比賽勝負狀態改變，但得分依舊是比賽中重要的概念。
- Rank 3：未帶有打點的安打、盜壘、雙殺或三殺。雖未造成得分，但這些概念的發生仍間接影響比賽結果。
- Rank 4：帶有殘壘的三振、內外野出局或犧牲觸擊。壘上有人時，觀眾通常會對接下來的打者帶有期待。若他未能打出安打，則相對來講對於比賽結果與觀眾感受的影響較大。
- Rank 5：未帶有殘壘的三振與內外野出局。棒球比賽中最普通的情況。

根據這樣的分類，利用本研究發展出的演算法選擇適合選為摘要的片段。此演算法如圖九所示：

- Level 1: Only the concepts with rank 1 and rank 2 are collected. This level of summary contains the most compact results.
- Level 2: Basically, only concepts with ranks 1–3 are collected. Rank-3 concepts and rank-4 concepts are considered to be discarded or added through checking concept context:
  - Rank-1 and rank-2 concepts are definitely picked as the summary.
  - Check each rank-3 concept i.
    If both the ranks of the (i+1)-th and (i+2)-th (r_{i+1} and r_{i+2}) concepts are less than 4, pick them all as the summary.
    If $r_{i+1} < 3$ and $r_{i+2} = 5$, just pick the i-th and (i+1)-th concepts as the summary.
    If $r_{i+1} = 4$ and $r_{i+2} = 5$, ignore all the i-th, (i+1)-th, and (i+2)-th concepts.

圖九、自動摘要演算法



圖十、自動摘要與自動精采畫面產生之結果與電視台製作之結果比較

我們將自動摘要的結果與電視台專業記者剪輯出來的摘要作比較，發現本研究的作法不管是 precision 或 recall 都可達到八成以上的準確度，如圖十所示。

與自動摘要類似的還有另外一種應用：自動選取精彩片段。在這個部分我們除了考慮各個片段代表的概念之外，還考慮到它發生的時間與相對應時間內聲音變化的情況。因為通常比賽越到後半段張力越高，而且當有精彩事件發生時，觀眾與主播的聲音將明顯變化。我們將每段視訊的重要性以底下的式子表示：

$$S(E_i) = S_r(E_i) \cdot S_t(E_i) \cdot S_a(E_i)$$

它代表概念的意義、時間、與聲音變化造成的重要性的加乘。在此之後，再根據圖十一所表示的精彩片段選取演算法(highlight selection algorithm)產生出最後的結果。這種選取方式可根據使用者的需求產生出不同長度的 highlight，如圖十所示。

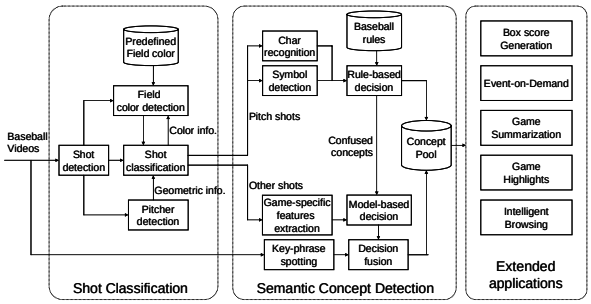
Input: the user-defined highlight length T and the set of events E in the game.
Output: the set of highlighted events A.

```
HIGHLIGHT_SELECTION(T, E)
1 A ← ∅
2 sort E into nonincreasing order by significance degrees
3 for each e_i ∈ E
4     do if length of (A ∪ {e_i}) < T
5           then A ← A ∪ {e_i}
6                SMOOTH(A)
7 return A
```

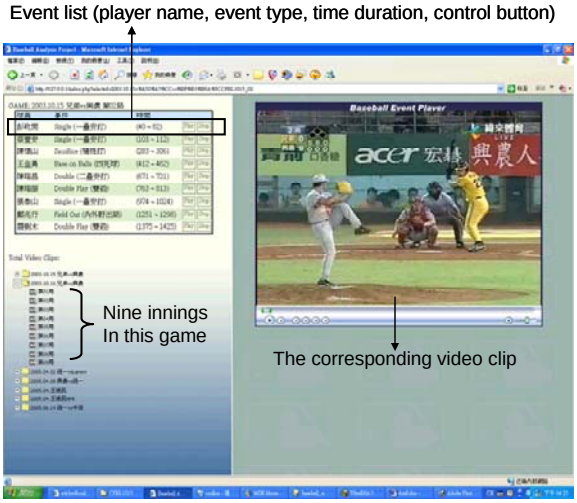
圖十一、精彩片段選取演算法

本研究計劃的實驗結果顯示出此種方法能有效且精確地幫助我們檢視棒球影片內容，也成為跨越語意鴻溝(semantic gap)研究中一個很典型的例子。整個棒球語意分析的過程可用圖十二來表示。它包括場景分類(shot classification)，語意概念偵測(semantic concept detection)、以及延伸應用(extended applications)。有了這樣架構之後，我們可得到精確且全面的概念偵測結果，也因此可發展許多有趣的應用。除了之前所提到的自動摘要與精彩片段選取之外，我們亦可提供線上的隨選視訊服務，如圖十三所示。類似的服務已經被廣大的棒球

球迷所接受，也已經成為新一代運動轉播公司所倚重的重要收費服務之一。本研究所得到的結果可有效地幫助服務提供者進行視訊處理，也可幫使用者產生更符合需求的視訊內容。



圖十二、棒球語意分析完整流程圖



圖十三、隨選事件的使用者介面

B【複合式影音展示】

B.1 圖片特徵值萃取

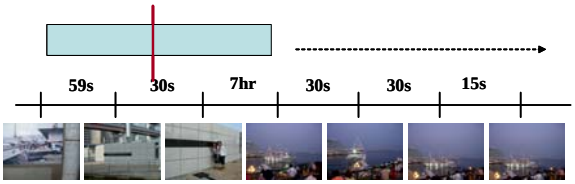
我們根據 MPEG-7 標準[4]所建議的特徵值為資料庫的影像建構了簡短的數位描述。這些數位描述主要由色彩描述子所構成，它們包括主要色彩(Dominant Color)、色彩配置(Color Layout)等等。主要色彩就是影像中主要的構成色彩，譬如，一張由藍天大海和沙灘構成的照片，它的主要色彩很明顯是藍色和褐色。主要顏色的抽取主要是先將影像轉換至適合的色彩空間(Color Space)，如 RGB、HSV 等，然後透過 Generalized Lloyd 演算法將色彩量化並群集，最後演算法收斂所得的結果就是照片的主要色彩構成。

色彩配置則是將影像劃分成數個區塊影像(Block Images)，計算區塊影像的平均(mean)及方差(variance)，這些區塊影像的統計特性可以表示原本影像的色彩空間分佈。因此，我們不僅可以利用主要色彩和色彩配置在廣大的影像資料庫找出我們所需的數位影像，同時也可以有效地幫助整理散亂的照片。

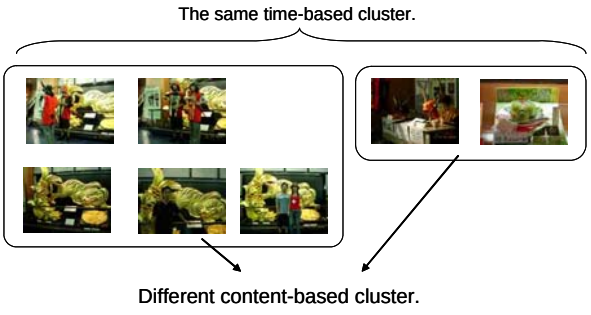
B.2 圖片叢集(photo clustering)

基於圖片的特徵值，我們發展自動叢集(clustering)功能，以便於整理大量的圖片資料。在目前的成果中，我們主要以旅遊照片為處理的對象。圖片叢集的目的在於把大量的資料做分類，以便於往後進行管理與呈現。

時間資訊在旅遊照片中扮演很重要的資訊，通常人們會在特定景點在短時間內拍攝大量照片。照片的時間資訊可由隱含於資料中的 metadata 取得[4]。因此，我們利用此特性，利用一個 sliding window 觀看兩兩照片的時間差距。若有相鄰兩張照片的時間差大於一個動態界限值(dynamic threshold)時，就將其歸類於不同 cluster，如圖十四所示。



圖十四、基於時間的圖片叢集

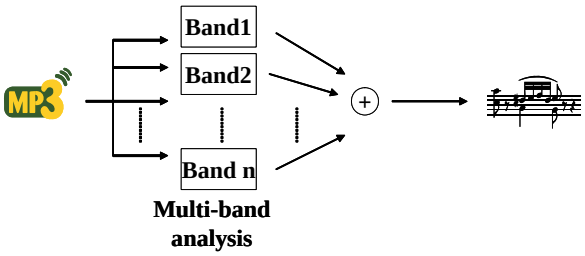


圖十五、基於內容的圖片叢集

有時一個時間叢集(time-based cluster)會過於龐大(包含太多張照片)，因此我們可基於前一小節所提到的主要色彩與色彩配置等特徵值對他們再細分，如圖十五所示。在此我們採用 k-nearest neighbor 的作法將一個 time-based cluster 中的照片分成多個 content-based cluster。

B.3 音樂特徵值萃取

為了往後在配合音樂與視覺畫面能更協調，我們希望能偵測音樂的節奏，以便未來依據音樂的拍子進行畫面切換與安排。如圖十六所示，我們主要依據 Scheirer 所提的方法[6]，將訊號分成不同的頻帶(frequency bands)來分析。在偵測各個頻帶的能量變化之後，整合各頻帶的變化整合成整首音樂的拍子資訊。

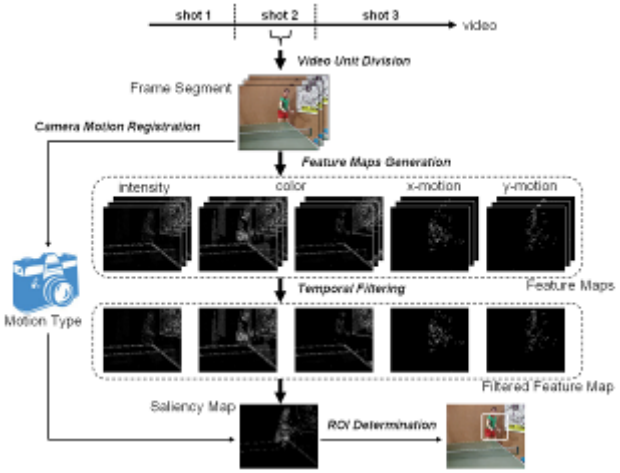


圖十六、音樂拍子偵測

B.4 使用者注意模型建置

我們將提出一個以使用者注意模型 (user attention model) 為基礎的自動視訊興趣區決定架構(如圖十七所示)。在這個研究中，視訊的注意特徵值 (attentive features) [5] 及應用媒體美學的知識都被同時考慮且利用。使用者興建區可用於往後建造複合式影音展現的重要依據。

在自動決定視訊興趣區的過程中，首先我們將欲分析之原始視訊以一使用空間顏色敘述子(spatial color descriptor) 為核心之場景變化 (shot detection) 演算法將其分為數個場景。接著在每段場景之中，我們以固定長度數量的訊框組成互不重疊之訊框切片 (frame-segment)，以每一訊框切片取代單一之訊框做為視訊興趣區分析的基本單位。接著在每一訊框切片中，我們對每一張訊框分別以使用者注意模型取出三類不同的視覺注意特徵值，包含亮度 (intensity)、顏色 (color)及運動(motion)，並分別得到其對應之特徵值映圖 (feature map)。對於不同種類的特徵值映圖，我們分別以時間平均過濾器 (temporal mean filter) 將其過濾為唯一之已過濾特徵值映圖 (filtered feature map)。不同的已過濾特徵值映圖即用以表示在該訊框切片中其對應之某類注意特徵值的空間分布情形。另一方面，我們也對每一訊框切片找出其所屬之運鏡種類，將不同之已過濾特徵值映圖合併為單一之顯著映圖 (saliency map) 時，此運鏡種類資訊將用以決定每個特徵值映圖之合併參數。在得到該訊框切片之顯著映圖後，即可決定出該訊框切片之興趣區數量，同時決定出每個興趣區在訊框中之大小及位置。整部原始視訊即可以此種方式分段決定出所有的使用者興趣區。

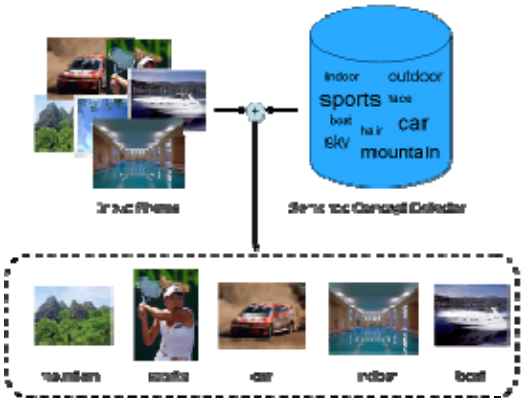


圖十七、視訊興趣區決定之所提架構

B.5 實作自動圖片分類模組

在人手一台相機的時代，照片和我們的生活緊緊相連，但大量的照片不僅整理不易且無從瀏覽，因此，自動圖片分析技術也日漸重要。我們根據第一年所作的圖片叢集(photo clustering) 的分析，實作完成圖片自動分類的模組，我們利用第一年所提的方法萃取出圖片的特徵值，除了利用時間資訊來作使用者的圖片叢集以及利用內容資訊的圖片叢集外，我們另外發展一種分類模組可供使用者選擇，此模組為採用將網路上收集到的大量照片

配合機器學習的方法，如 SVM、HMM 等，得到圖片的語義概念集(Semantic Concept Set)。最後，使用訓練好的分類器將使用者照片依序分類，實作的圖片分類模組的流程如圖十八所示。



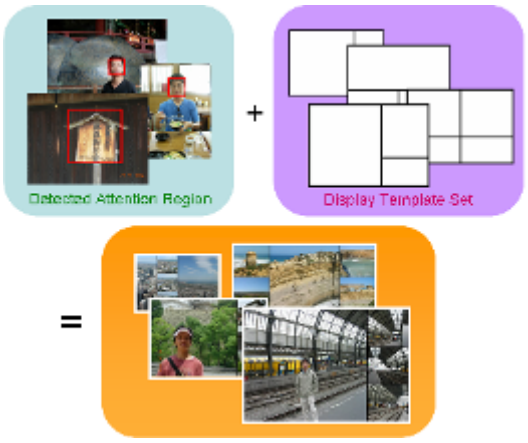
圖十八、圖片分類流程

圖片分類完成後，使用者即可使用關鍵字查詢或瀏覽想要的照片，這種分類方式幫助使用者更有效地管理個人照片。同時分類完的照片主題相關性更強，更適合複合式影音的呈現。

B.6 實作使用者注意模型分析模組

目前影像處理中最待解決的問題，即是同一張影像在不同裝置上的呈現，如個人電腦及手機等。直接的影像大小的縮放，經常會導致照片中的人物或景物變形失真，我們利用第一年中所分析的使用者注意特徵值[5]，實作出可以有效決定照片中使用者感興趣的視覺中心的模組，並在指定的區域大小(例如手機的螢幕比例等)為限制下，我們可進一步得到適合指定區域大小的照片區域。同時，我們的方法可確保照片中的人事物保持正常的比例，使用者可以在指定區域中(例如手機螢幕)，清楚地瀏覽自己的照片。

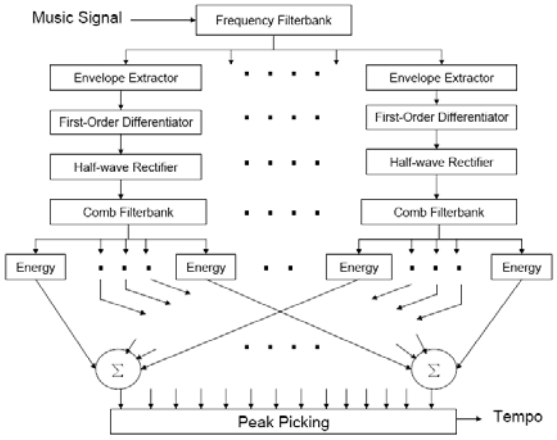
利用相同的技術，我們還可以將使用者的照片，配合預先設計好的拼貼範本，產生創意十足的照片拼貼，結果如圖十九所示。



圖十九、使用者注意模型分析與應用

B.7 實作音樂節奏分析模組

基於第一年在音樂特徵值萃取的研究方法，我們實作出音樂節奏分析的模組，其演算法如圖二十所示。首先我們將音樂訊號分成六個不同的頻帶(frequency band)來分析，為了避免在聲音上突然的變化所造成的干擾，這些不同頻帶中的信號會進一步轉變為主要能呈現出信號趨勢的包絡型式(envelope form)。緊接著對於不同頻帶的包絡信號，利用一階微分器以取得聲音振幅(amplitude)的最大改變量的位置，並透過半波整流(half-wave rectified)以作為進一步的週期性脈波分析。

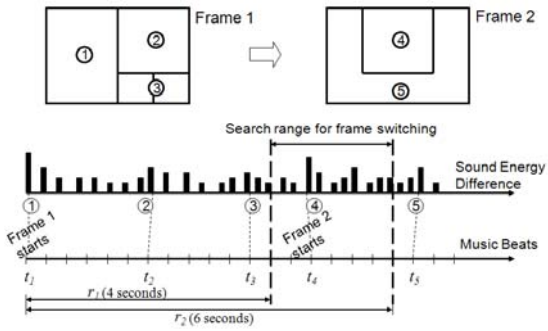


圖二十、音樂節奏偵測演算法

與傳統的方法不同，我們採用梳狀濾波器(comb filter)作進一步的分析[6]，如果梳狀濾波器的週期與輸入訊號相同，則會有比週期不相同的輸入信號還大的能量輸出值。接著將對於不同的頻帶

中，相同週期的能量輸出加總後，有最大能量加總輸出值的週期，即為音樂訊號的節奏資訊。因此將輸入的音樂信號作節奏的偵測與分析之後，我們便可以利用音樂的節奏資訊來輔助取得整首音樂的節拍位置。

利用音樂的節拍資訊便可進一步作為拼貼畫面切換的依據。由於音樂長度有限，為了能讓音樂和照片能更緊密的結合，我們將以音樂為主並將音樂分成幾個段落，而後根據音樂的分段數來選擇合適的照片來作展示播放的搭配。同時，我們也將考慮音樂節拍的強弱，選擇強拍發生的時機來做畫面的切換，藉由此方式加強複合式影音的瀏覽效果，照片和音樂節拍搭配的流程如圖二十一所示。



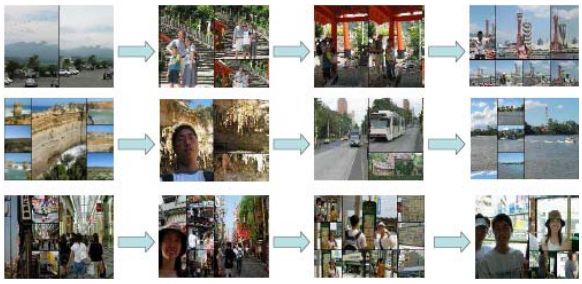
圖二十一、音樂節奏分析與應用

B.8 拼貼幻燈秀整合與實作

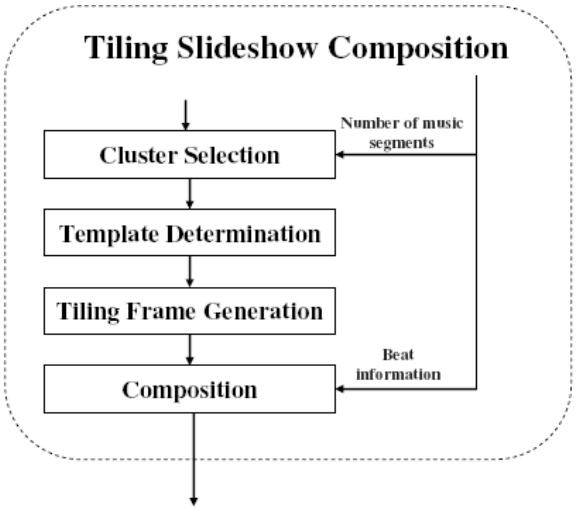
複合式影音展示之目標在於以幻燈秀之型式將屬於同一叢集的照片展示於同一個影格(frame)之中，並以照片內容為基礎，精巧地將照片配置於每一塊拼貼區域(tile)。在拼貼的過程中會遇到幾個難題:

- 在時間長度有所限制的配樂片段中，我們無法以適當的照片播放速度播放完所有照片，必須從中挑選部份播放。例如，當每個影格停留 4~6 秒鐘時，一首長度 4 分鐘的音樂最多只能播放 60 張照片。
- 對於屬於同一叢集的照片，必須合理地將照片配置於畫面中。例如，較重要或較引人注意的照片應顯示於較大的拼貼區域中，而較相似的照片應被配置於較相鄰的拼貼區域中。根據上述原則，我們必須設計一演算法，從所有定義好的拼貼版面(layout)中選取最適合的版面以及位置來放置照片。
- 一旦決定好哪些照片該配置在哪些拼貼區域後，仍需適當地裁切照片，並調整照片尺寸後才可放入拼貼區域。因此須設計一最佳化(optimization)演算法決定照片的縮放比例與裁切的區域。

針對上述難題，我們整合第一年與第二年的研究成果，開發如圖二十二所示之幻燈秀，圖二十三為生成幻燈秀之流程圖，我們將針對每一模組詳加介紹。



圖二十二、複合式影音展示之幻燈秀系統



圖二十三、實作拼貼幻燈秀之流程圖

B.8.1 選擇照片叢集(Cluster Selection)

本計劃第二年所研發之音樂節奏分析成果可依據節拍資訊將使用者所選擇之背景音樂切割成數個較短的時間區段(約4~6 秒)。屬於同一個照片叢集(Photo Cluster)的照片將在同一個時間區段中播放，然而使用者輸入的照片數目龐大，而一首包

含有限(N_s)個時間區段的音樂僅能播放N_s個照片叢集，因此我們依據每個照片叢集的重要性，選出前 N_s 個照片叢集來播放。

- **定義照片叢集之重要性**

每個照片叢集的重要性由兩個特徵定義而得：每秒單位的照片數(PPM: photos per minute)，以及照片一制性(PC: photo conformance)。以時間為基礎所找出的某個照片叢集Ψ，可再以內容為基礎將其分為m個照片叢集C_k, $k=1,2,...,m$，我們將每秒單位的照片數PPM定義為：PPM(C_k)=N(Ψ)/*Time Duration*(Ψ)，其中N(Ψ)為Ψ中所包含的照片數目。而照片一制性(PC)可定義為:

$$PC(C_k) = 1 - \frac{1}{n_k(n_k-1)} \sum_{P_i \in C_k} \sum_{\substack{P_j \in C_k \\ P_i \neq P_j}} d(P_i, P_j)$$

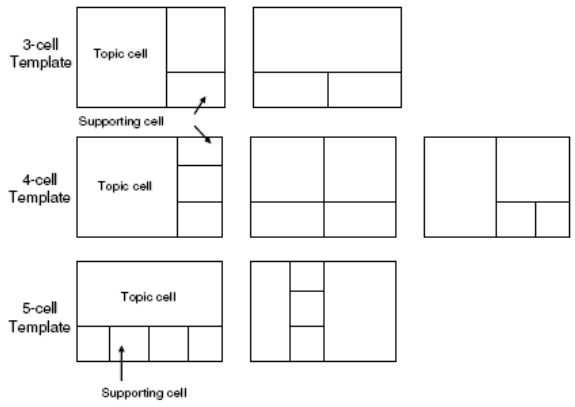
此定義在概念上類似內部叢集距離(within-cluster distance)的定義。結合上述兩種特徵，我們以一個特徵向量來描述一個照片叢集C_k:

$$\bar{x}=(PPM(C_k),\ PC(C_k))$$

而C_k 的重要性CI_k可用一非線性結合函式定義:

$$CI_k = E(\bar{x}) + \frac{1}{2(m-1)+m\lambda} \sum_{i=1}^{m} |\bar{x}_i - E(\bar{x})|$$

其中E(・)為特徵向量的平均值，λ為事先定義之常數。



圖二十四、包含 3~5 個巢的版面樣本範例

B.8.2 生成拼貼影格(Tiling Frame Generation)

決定該播放哪些照片叢集後，我們為每個照片叢集生成一拼貼影格。欲將含有h張照片的照片叢集配置到畫面中，系統會自動生成含有h-巢(cell)的版面樣本，圖二十四為包含3~5個巢的版面樣本範例。為了讓幻燈秀更富多樣性，系統設計了多個包含h-巢的版面樣本{$T_{h;1}$, $T_{h;2}$,..., $T_{h;s}$}，我們必須從中挑選一個適當的版面樣本，並決定每張照片在版面中的位置。

- **定義重要性向量(Importance Vector)**
  - **定義版面樣本重要性向量**

如圖二十四所示，每個版面樣本會包含至少一個主題巢(topic cell)與多個配角巢(supporting cell)。主題巢占有較大的面積，通常較能引起觀眾的注意力。對於一特定樣本版面T所包含的每個巢Tc_i，可定義其重要性為巢Tc_i在版面T中所占的面積比率:

$$Ic_i = Area(Tc_i)/Area(T)$$

將Ic_i由大至小排列後，每個樣本版面T可得一對應之樣本重要性向量(template importance vector):

$$TV=(Ic1, Ic2, ..., Ick)$$

  - **定義照片叢集重要性向量**

在 B.8.1 中，我們以叢集為基礎(cluster-based)，為每個照片叢集定義一重要性值，在此我們以叢集中的每張照片為基礎(photo-based)，為每個叢集定義一重要性向量(PV)。此重要性由兩個特徵組成:臉部區域比例(FR: face region)與注意力值(AV: attention value)。照片P_i 的臉部區域比例FR可定義為:

$$FR(P_i) = \sum_{j}^{n_f} Area(Face_j)/Area(P_i)$$

，其中n_f 為照片中所偵測到的臉部個數。而照片P_i 的注意力值AV可定義為:

$$AV(P_i) = \sum_{x} \sum_{y} Sa(x,y) \times G(x-m_1, y-m_2)$$

其中$S_a(x,y)$為點(x,y)的顯著值(saliency values)，(m_1,m_2)為顯著圖(saliency map)的中心點。我們將兩個特徵值FR與AV做加權線性結合得到照片P_i的重要性:

$$PI_i = W_{face} \times FR(P_i) + W_{attention} \times AV(P_i)$$

由於臉部區域比例較注意力值更具語意概念，我們

給予其較高的權重。一個照片叢集中的所有照片重要性(PI_i)由大至小排列後，可組成此照片叢集之重要性向量(PV : importance vector)。

B.8.3 決定版面樣本(Template Determination)

對於一個含有h張照片之照片叢集，我們比較此照片叢集之重要性向量(PV)與所有含h-巢的版面樣本之重要性向量($TV_{h;i}$)，並選擇向量相似度最高的版面樣本:

$$T_{h,i} = \arg \max_{i=1,2,\ldots,s} \left(\frac{PV \cdot TV_{h,i}}{||PV|| ||TV_{h,i}||}\right),$$

而每張照片在此版面樣本中的位置可由每張照片的重要性決定，擁有較高重要性的照片會被放置在較大的巢中。

B.8.4 合成幻燈秀(Composition)

將照片放至對應的巢中，就像是將磁磚貼至牆上所屬的位置，因此我們稱此系統為拼貼幻燈秀。然而每個巢的長寬比，通常與所要放置的照片長寬比不同，且照片的解析度往往超過1600x1200個像素，遠大於每個巢的像素。為了不讓照片內容嚴重失真，對於照片叢集中的每張照片都須藉由切割(cropping)與調整尺寸(resizing)等技術來選擇一適當的照片顯示區域做為拼貼的素材。我們將此問題導為一個有條件限制的最佳化問題。

- **選擇照片顯示區域之最佳化**

給定一張照片，從中找尋一塊內容值(content value)最大的區域 R，且 R 的長寬比和所要放置的巢(R_c)之長寬比相同:

$$\max_R C(R), \text{ such that } g(R) = g(R_c),$$

其中 g(・)為該區域之長寬比，C(・)為該區域所含之內容值。根據第二年所研發之使用者注意模型，我們可算出每一塊區域的內容值(content calue)，並由上式找尋一塊使資訊損失量最少但又符合適當長寬比的區域。最後截取此區塊，將其解析度調整為巢的解析度，即可拼貼入巢中。

四　結論與討論

本計畫的成果主要分兩部份。第一部分是以棒球影片的事件分析為目的，棒球比賽中的事件是球迷了解比賽過程與球員表現的重要依據，不同事件所造成的效果對於球員與觀眾有不同的意義。因此，有別於以往只大約偵測精采畫面或場景分類的研究，本系統已能確切地偵測棒球比賽中發生的各種事件。

第二部份是實作複合式影音展現之拼貼幻燈秀，不同於以往一張幻燈片僅播放一張照片的呈現方式，拼貼幻燈秀結合了音樂與照片內容分析之成果，讓使用者一次瀏覽多張照片，卻又能使視覺畫面與聽覺刺激有明顯的調和，不失欣賞與回味的樂趣。

- 計劃成果自評

【影片事件偵測】的第一年成果以畫面分類、字幕資訊辨識以及基於棒球規則的事件偵測模組為主；第二年進一步實作出以模型為基礎的事件判斷模組，完成準確的事件偵測；第三年整合事件分析系統並發展如自動摘要、精彩影片選取、線上隨選視訊服務等延伸應用。此影片事件偵測技術的貢獻可簡短摘要如下：

- ✧ 多層架構(multilevel framework)：引進中介資訊，根據不同數位內容決定與發展對應函式(mapping function)。
- ✧ 棒球影片中的語意分析：以字幕資訊為中介資訊，詳盡利用棒球規則於語意分析。另外結合畫面與語音資訊準確完成所有棒球概念的偵測。此技術模組完成了內容的分析與檢索(indexing)。
- ✧ 延伸應用：基於分析結果發展實際且有用之應用。自動產生摘要與精彩片段完成內容重新組織的目標。

【複合式影音展示】方面，第一年與第二年利用圖片特徵值萃取技術將照片自動分類，並研發使用者注意模型與音樂節奏分析模組。有了自動圖片分類的模組之後，我們才能將相近的照片同時展現，藉以提昇整個視覺的強度與資訊的廣度。有了音樂節奏分析模組，我們才能配合音樂的節拍作圖

片的展現，藉以同時滿足視覺與聽覺上的感官享受。第三年進一步利用圖片分類模組、使用者注意模型和音樂節奏分析模組，整合視訊與音訊，完成自動幻燈秀的實作。

在【影片事件偵測】與【複合式影音展示】中所研發之模組與系統皆受到國際著名會議與期刊之肯定並刊登(詳列於**六.相關論文之發表情況**)，其中所研發之拼貼幻燈秀系統也已申請為專利，足見本計劃之學術與應用價值不容小覷。

五 未來研究方向

在【影片事件偵測】方面：本計畫係以棒球影片為主要之分析對象。未來對具有類似特色之球賽影片，如網球電視影片，應可以相似方法進行重要事件之偵測。唯各種不同之球賽影片有其特有之影片特性，在分析方法及技巧上也有所不同，應是值得投入的研究方向之一。此外，事件偵測的結果除了本計畫所展示的「自動摘要」,「精彩影片選取」,及「線上隨選視訊服務」等應用外，是否有其他更具商業價值的應用，將也是未來研討的對象之一。

在【複合式影音展示】方面：本計畫目前所完成之影音幻燈秀系統係數位內容呈現之利器，未來將考量如何將更具人性特色的因素，如情緒，納入整合系統以期所呈現之內容及可滿足人類視覺與聽覺的享受外更可與使用者的情緒相契合。

六 相關論文之發表情況

1. W.-H. Cheng, C.-W. Wang, and J.-L. Wu, "Video Adaptation for Small Display Based on Content Recomposition," IEEE Transaction on Circuits and Systems for Video Technology, vol. 17, no. 1, pp. 43-58, January 2007.
2. J.-C. Chen, W.-T. Chu, J.-H. Kuo, C.-Y. Weng, and J.-L. Wu, "Tiling Slideshow," Proceedings of ACM Multimedia Conference, pp. 25-34, 2006.
3. W.-T. Chu and J.-L. Wu, “Explicit Semantic Events Detection and Development of Realistic Applications for Broadcasting Baseball Videos,” revised for Multimedia Tools and Applications, 2006.
4. W.-T. Chu, C.-W. Wang, and J.-L. Wu, “Extraction of Baseball Trajectory and Physics-Based Validation for Single-View Baseball Video Sequences,” accepted by IEEE International Conference on Multimedia & Expo, 2006.
5. W.-T. Chu and J.-L. Wu, “Development of Realistic Applications Based on Explicit Event Detection in Broadcasting Baseball Videos,” Proceedings of International Multimedia Modelling Conference, pp. 12-19, 2006.
6. W.-T. Chu and J.-L. Wu, “Integration of Rule-based and Model-based Methods for Baseball Event Detection,” Proceedings of IEEE International Conference on Multimedia & Expo, pp. 137-140, 2005.
7. W.-H. Cheng, W.-T. Chu, and J.-L. Wu, “A Visual Attention based Region-of-Interest Determination Framework for Video Sequences,” IEICE Transactions on Information and Systems Journal, vol. E-88D, no. 7, pp. 1578-1586, 2005.

七 參考文獻

[1] Kitter, J., Hatef, M., Duin, R.D.W., and Matas, J., “On combining classifiers,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 3, pp. 226-239, 1998.

[2] Chen, B., Wang, H.-W., Chien, L.-F., and Lee, L.-S., “A*-admissible key-phrase spotting with sub-syllable level utterance verification,” Proceedings of IEEE International Conference on Spoken Language Processing, 1998.

[3] Chinese Professional Baseball League, http://www.cpbl.com.tw

[4] S.-F. Chang, T. Sikora, A. Purl, “Overview of MPEG-7 standard,” IEEE Transactions on

Circuits and Systems for Video Technology, vol. 11, no. 6, pp. 688-695, 2001.

[5] L. Itti, C. Koch, and E. Niebur, “A model of saliency-based visual attention for rapid scene analysis,” IEEE Trans. PAMI, vol. 20, no. 11, pp. 1254-1259, Nov. 1998.

[6] E.D. Scheirer, “Tempo and beat analysis of acoustic musical signals,” Journal of Acoustical Society of America, vol. 103, no. 1, pp. 588-601, 1998.

[7] 其他參考文獻請參考計畫書