利用特寫鏡頭偵測與主角辨識技術來自動建立電影摘要

范世鎮 劉志俊
中華大學資訊工程學系
ccliu@chu.edu.tw

摘要

電影是現代人類最重要的文化資產之一。隨著數位化電影資料日漸成為人們日常生活的一部份，電影資料的內涵式分析成為目前重要的研究主題。在本文中，我們提出一種電影摘要自動合成技術以及一種角色自動分析技術。我們利用鏡頭樣板來判別電影鏡頭是否為特寫鏡頭。再將特寫鏡頭依照膚色特徵來作叢集分析，來找到一部電影各個主要演員的戲份。最後我們根據演員的戲份比重，由特寫鏡頭叢集中挑選代表性特寫鏡頭來合成電影摘要。

關鍵字

MPEG-4、視訊摘要(video summarization)、電影摘要(movie summarization)、特寫鏡頭(close-up shots)、電影資料庫(movie databases) 、內涵式查詢(content-based retrieval)

1. 序論

多媒體資料的摘要 (multimedia summarization)為目前多媒體資料庫領域的重要研究主題之一。而電影是現代人類最重要的文化資產之一，隨著數位化電影資料日漸成為人們日常生活的一部份，所以電影資料的自動化摘要成為許多視訊/電影應用系統的重要技術。

一部電影在要上映之前的第一步就是為這部電影作宣傳，而最直覺宣傳的工具當然就是這部電影的電影摘要。典型的電影摘要，如電影海報跟電影劇照的設計，目的是為了讓所有使用者能對此部電影的大致內容能一目了然。尤其是電影海報，在宣傳時期可以說到隨處可見，就連是公車站牌都能看見正當宣傳期的電影海報。除此之外當然還有最動人心弦的電影預告片，可以說是一部電影最精華的部分。從預告片中可以看到主要演員在重要場景中的演出，以及最具代表性的鏡頭(shots)。

不論是在電影海報、電影劇照或電影預告片中，都可以清楚的看到其中使用了大量的特寫鏡頭。因此我們可以藉由特寫鏡頭的偵測及擷取，來自動合成電影摘要。我們也可將自動合成的電影摘要作為索引，用於電影資料庫內涵式查詢。如此一來使用者可在大量的電影資料中，很快的搜尋到有興趣欣賞的電影。

MPEG-4 為新一代的多媒體資料壓縮標準[1]。由於 MPEG-4 以物件作為壓縮單位，不會將單調的背景一起壓縮進去，如此便能降低一定程度的資料量。MPEG-4 視訊中最重要之概念就是所謂的視訊物件平面(Video Object Plane;VOP) [1] [2]。為了達到與過去視訊壓縮標準的相容性(如 MPEG-1、MPEG-2、H.261 以及 H.263 等)，視訊物件平面仍然是以區塊為主(Block-based) ，但也容許有任意形狀的物件，因此不但有更多的彈性，也能達到與過去標準的相容性。此外亦可提供對畫面上不同的物體，依頻寬傳送不同解析度的資料流[6] (不論是空間上或是時間上)，這對於像是網路傳送視訊摘要的應用非常實用。所以，未來電影資料非常可能是以 MPEG-4 格式來儲存。

在多媒體的摘要的相關研究方面，視訊摘要的應用是屬於比較成熟的部分，但是視訊摘要的做法與定義看法十分分歧。在[4]論文中說明目前視訊摘要主要有兩種方法，第一種是將鏡頭中的重要物件形狀或場景取出，作為一段視訊摘要的；第二種則是將”重要的”或”有趣的”場景取出，作為一段視訊摘要。而在[3]

*本論文研究為國科會補助之研究成果，計劃編號NSC 91-2213-E-216-003

的論文中，則是提出使用人類心理學作為視訊內容的基礎，改變了以往使用框架(Frame)架構的方式，將視訊摘要融入心理結構的成分。

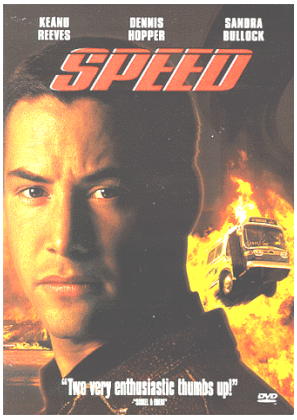
[11]一文中提出一種電影摘要方式。提出以"米字形"物件形狀特徵值來偵測並擷取特寫鏡頭，然後將特寫鏡頭合成電影摘要。但是特寫鏡頭有很多都是在兩人對話時出現，此時的特寫鏡頭會讓人物物件靠左或靠右，如此便會影響特寫鏡頭偵測的準確度。在本文中，我們提出另一種電影摘要自動合成技術以及一種角色自動分析技術。我們利用特寫鏡頭樣板(close-up templates)來比對電影鏡頭是否為特寫鏡頭。再將特寫鏡頭依照膚色特徵來作叢集分析，來找到一部電影各個主要演員的戲份，然後將特寫鏡頭合成電影摘要。

本論文的結構說明如下。在第 2 節中我們將介紹現行電影摘要的種類與涵義；在第 3 節中說明我們所提出的電影摘要系統的整體架構；鏡頭的種類與定義以及特寫鏡頭的偵測方法將在第 4 節中說明；如何利用膚色特徵將特寫鏡頭做叢集分析來進一步辨識演員戲份的技術將在第 5 節中說明；第 6 節中說明電影摘要合成方式；第 7 節說明主要的實驗結果；最後第 8 節為本文的結論。

2. 目前電影摘要的種類

目前電影摘要主要有電影海報(posters)、劇照(stage photos)、電影預告片(previews)、劇情簡介(synopsises)以及電影寫真集[14]等。

2.1 電影海報及劇照



a.特寫版面海報　　b.演員群版面海報

c.摘要版面海報

d.綜合版面海報

圖 1. 電影海報種類

電影的宣傳方式大致分為靜態與動態兩大類，其中以電影海報為靜態宣傳的主要代表。電影的海報分類如圖(1)所示，圖(1.a)為特寫版面海報，內容為將一部電影中的一個主要角色，以特寫方式搭配電影場景作為海報；圖(1.b)為演員群版面海報，擷取電影中重要角色的鏡頭畫面，利用合成技術合成電影海報，這類電影海報是最常見的一類，而使用的鏡頭又以特寫鏡頭占絕大多數；圖(1.c)為摘要版面海報，從電影的劇照中挑選具帶表性的精采劇照，結合成為電影海報；圖(1.d)為綜合版面海報，它讓設計者可以自由排版，自由度相當

高，沒有侷限應用的鏡頭種類，包含電影場景也可成為海報主題。

2.2 電影預告片

電影動態宣傳方式的主要為電影預告片，它是將電影中最具代表性，也最吸引人的電影鏡頭結取出來，然後合成一段電影短片。

2.3 劇情簡介

劇情簡介是以第三者的身分，對一部電影做文字的簡介與劇情說明，國家電影資料庫(http://www.ctfa.org.tw/)目前的電影摘要就是以劇情簡介為主。

2.4 電影寫真集

電影寫真集裡有許多精采劇照都是使用特寫鏡頭，除此之外它記載電影幕前幕後花絮[14]，其中包括電影的劇本、電影裡的服裝設計介紹、排演過程等等，配合著電影的上映，進一步的促進該部電影的推廣。

3. 電影摘要的系統架構

本文所提出之電影摘要系統整體架構如圖 2 所示。我們假設要建立電影摘要的電影為 MPEG-4 格式，因為其中的物件已經標明，若為其他格式的電影資料，則須經過物件辯識的程序取出電影中的物件。經由鏡頭偵測模組(Shot Change Detection) [12]，以鏡頭為單位對一部電影進行切割。再將電影鏡頭做鏡頭前製處理(Pre-Processing)，取出關鍵視訊物件平面後，對視訊物件平面做特寫鏡頭偵測(Close-Up Shot Detection) 。

自動判斷出特寫鏡頭之後，將特寫鏡頭輸出到鏡頭叢集分析器(Shot Cluster Analysis)做鏡頭叢集分析，進而判定演員的戲份。最後將鏡頭叢集輸出到電影摘要合成模組(Movie Summary Synthesis)中，在電影摘要合成模組中，我們可以從演員的戲份決定各個演員的重要性，依照重要性調整其在合成電影摘要時所佔之比重。

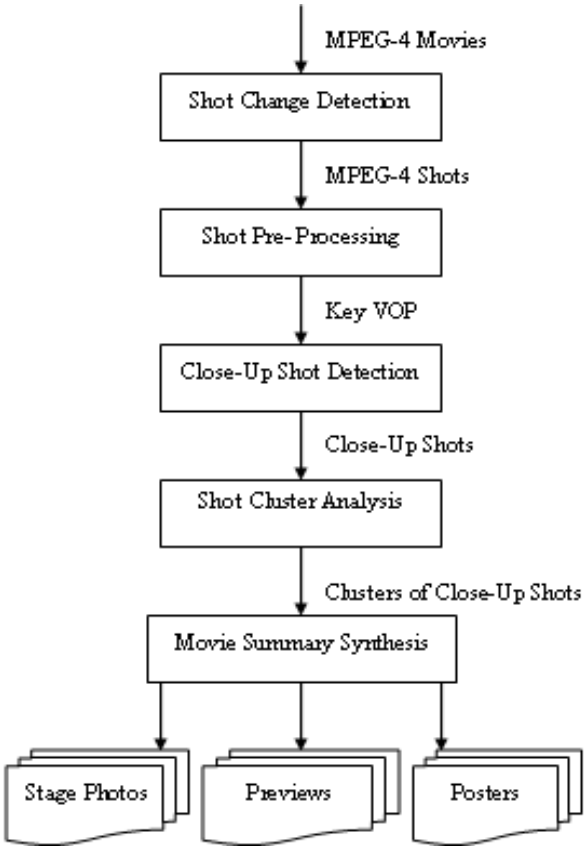


圖 2. 電影摘要系統架構圖

4. 特寫鏡頭辨識

4.1 鏡頭的種類

表 1. 電影鏡頭分類表[9][10][13].

鏡頭種類	鏡頭解釋
特寫鏡頭(Close-up)	經由近距離所拍攝的人或物的放大或細節描繪的鏡頭，如拍攝對象是人，則指肩部以上的拍攝範圍。
近景鏡頭(Close shot)	介於中景與特寫之間的鏡頭，如拍攝對象為人，指頭頂至胸腹的範圍。
中景鏡頭(Medium shot)	在視界與視覺角度方面介於特寫鏡頭與遠景鏡頭之間的一種鏡頭。以中景鏡頭來表示一個人時，最典型的視覺範圍是從該人的膝蓋以上來拍攝。
全景鏡頭(Full shot)	拍攝對象佔滿整個螢幕的鏡頭。若被攝體是一個人，則他或她的身體會全部容納在鏡頭中。
大特寫鏡頭	一種構圖非常緊密的特寫鏡頭，能將一件微小的物體，或物體和人物的某部分誇張放

(Extreme close-up)	大，譬如一張人臉的鏡頭，只顯出眼睛、鼻子或嘴唇部分。

不論是電影海報、劇照或是電影預告片，特寫鏡頭都是其主要內容來源。所以本文主要是提出使用鏡頭樣板自動偵測並擷取特寫鏡頭，再利用特寫鏡頭進一步合成視訊摘要。我們可以從表 1 的電影鏡頭分類[9][10][13]得知一般電影鏡頭大致分為特寫鏡頭(Close-up)、近景鏡頭(Close shot)、中景鏡頭(Medium shot)、全景鏡頭(Full shot)、大特寫鏡頭(Extreme close-up)等類型。各類鏡頭的範例如圖 3 所示，資料取自電影"哈利波特 2-消失的密室"。

a.特寫鏡頭

b.近景鏡頭　　c.中景鏡頭

d.全景鏡頭　　e.大特寫鏡頭

圖 3. 各類型電影鏡頭實例

4.2 使用鏡頭樣板偵測特寫鏡頭

由表 1 我們可以發現鏡頭的種類，是與視訊物件在鏡頭中出現的大小跟位置相關，因此我們便可以利用視訊物件在鏡頭裡的大小跟位置來判斷是否為特寫鏡頭，偵測方法如下：

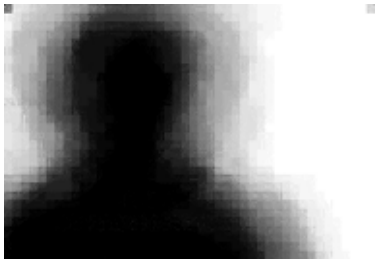
1. 首先我們找出各 256 x 3 = 768 個鏡頭，分別有 256 個視訊物件置中的特寫鏡頭、256 個視訊物件偏右的特寫鏡頭以及 256 個視訊物件偏左的特寫鏡頭。利用 768 個特寫鏡頭合成三種初始鏡頭樣板，如圖 4 所示。
2. 將初始鏡頭樣板做影像處理，將樣板做極色化(posterize)的處理，使得極色化的樣板分三區，分別為全黑區域(值為 0)、全白區域(值為 255)以及中間地帶(值為 128)如圖 5。
3. 對三個區域分別設定權重，也就是當物件落到該區域時的得分。經由實驗結果，發現將全黑區域得分設為 2，中間地帶得分設為 1，全白區域得分設為 -6，可以得到較佳實驗結果。
4. 我們將以鏡頭為單位的視訊物件平面分別與我們的三個樣板做位置比對，然後再將所得分數(S)代入正規化公式得到正規值(R)，公式如下：

$$\text{左樣板}\begin{cases} R = S/41469, S \geq 0 \\ R = 0, S < 0 \end{cases}$$

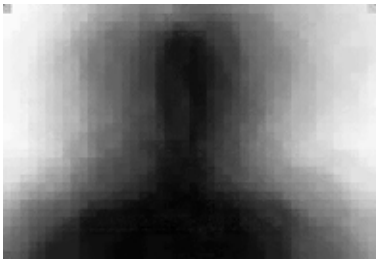
$$\text{中樣板}\begin{cases} R = S/38746, S \geq 0 \\ R = 0, S < 0 \end{cases}$$

$$\text{右樣板}\begin{cases} R = S/47462, S \geq 0 \\ R = 0, S < 0 \end{cases}$$

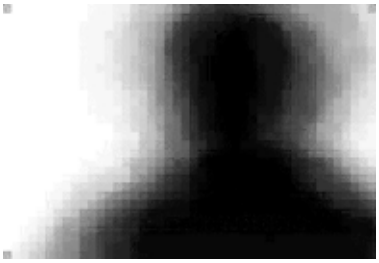
5. 視訊物件平面分別在三個樣版得分中，我們取最高的正規值作為正規值的代表，然後設定一個門檻值(Threshold)，若是正規值大於門檻值則判定其為特寫鏡頭。經由實驗的結果，我們發現將門檻值設為 0.85 可得到較佳的準確度(Precision)跟回覆率(Recall)。



a.左樣板



b.中樣板



c.右樣板

圖 4. 三種特寫鏡頭比對用初始鏡頭樣板

a.左樣板

b.中樣板

c.右樣板

圖 5. 三種特寫鏡頭比對用極色化後樣板

5. 演員戲份分析

5.1 特寫鏡頭的叢集分析

由前節結果中，我們可以得到整部電影裡的特寫鏡頭。我們希望能進一步對這些特寫鏡頭以演員為分類單位做叢集處理(假設每個鏡頭的演員身份皆為未知)。最終得到的結果是多個特寫鏡頭叢集，每個叢集可以視為同一個演員的特寫鏡頭集。

在[7]提出了一個在視訊影像的膚色識別方法。此方法首先將畫面切為 5x5 的區塊(patch) ，在將顏色由 RGB(Red, Green, Blue)領域轉換為 HSV(Hue, Saturation, Value) 領域。一開始讓使用者指派一個起始的區塊，然後以色度跟飽和度為基礎並計算其直方圖(histogram)相似度，可以得到一個不錯的膚色識別結果。

我們將[7]所提出方法運用在對演員的臉部做膚色叢集處理上。由於三種樣板比對法可以提供特寫鏡頭中演員臉部的大略位置的資訊，因此我們可以不用一開始讓使用者指派一個起始的區塊，而在每個樣版的額頭位置取一 5*5 像素的區塊作為起始的區塊，如圖 6 之紅色區域，如此便能有效的自動識別出特寫鏡頭的臉部部位。我們再以鏡頭的臉部部位的直方圖統計作為特徵值做叢集處理，每一個特寫鏡頭在初始化的狀態下會被視為是各自獨立的特寫鏡頭叢集，對每兩個特寫鏡頭叢集。我們嘗試將其合併為一個特寫鏡頭叢集，當兩個特寫鏡頭叢集的直方圖 相似度大於我們設定的門

檻值時，我們則將這兩個叢集合併為一個特寫鏡頭叢集。我們就可以得到叢集分類的結果。

下面我們舉一範例說明之。假設有 13 個特寫鏡頭，初始化時將每個特寫鏡頭設定為各個獨立的叢集，接著加入特寫鏡頭 2，由於特寫鏡頭 1 和特寫鏡頭 2 皆互相相似，特寫鏡頭 1 和特寫鏡頭 2 會被歸納在同一個叢集(叢集 1)中。再來加入特寫鏡頭 3，它會試著加入叢集 1。因為特寫鏡頭 1 和特寫鏡頭 2 其中有一個與特寫鏡頭 3 不相似，則特寫鏡頭 3 無法加入叢集 1，它會自己成為一個獨立的叢集。同理，我們就可以得到叢集分類的結果。如圖 7 所示。

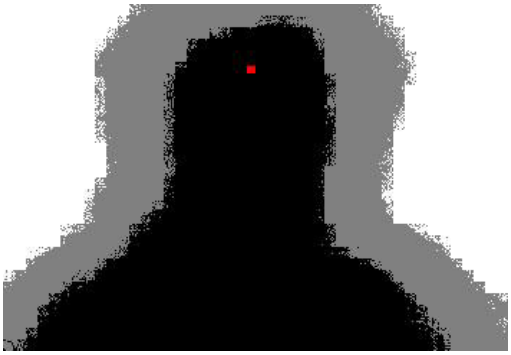


圖 6. 樣板額頭位置起始膚色區塊圖

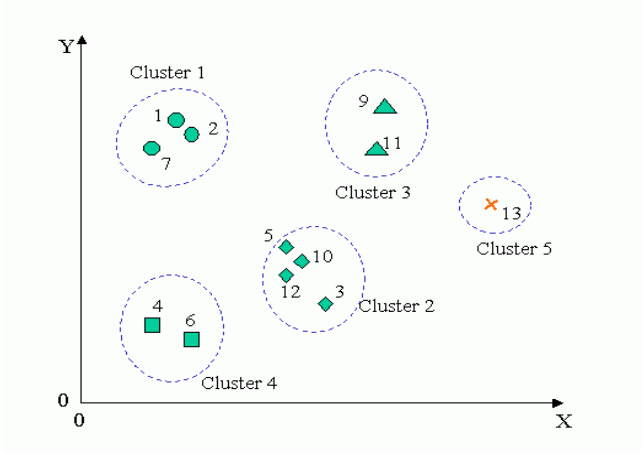


圖 7. 特寫鏡頭叢集分析範例

5.2 演員戲份的統計

我們統計每一個叢集的特寫鏡頭總數，並依照特寫鏡頭的數量由多到少排序，叢集 1 為最大的叢集，叢集 2 次之，叢集 3 再次之，依此類推。因為一個叢集代表一個演員，所以當叢集愈大則代表此演員的戲份愈重，藉由戲份的統計如表 2 所示，我們便可作角色自動分析，找出電影中的主要角色，並以主要角色的特寫鏡頭來製作電影摘要。

表 2. 哈利波特主角戲份排行表

主角排行	叢集 1	叢集 2	叢集 3
代表鏡頭			
戲份	45.7%	12%	6.8%

6. 製作電影摘要

6.1 合成電影劇照

由特寫鏡頭叢集所佔的特寫鏡頭總數來判斷此演叢集所佔的戲份，由戲份便可判定出主要演員的色寫鏡頭叢集之後便可按戲份比重來挑選主要演員的特寫鏡頭進行電影劇照合成，圖 8 為取自電影”哈利波特 2-消失的密室”中的合成劇照。

圖 8. 電影“哈利波特 2-消失的密室”的合成劇照

6.2 合成電影海報

在電影海報合成方面，我們先制定出電影海報的樣板，然後在依照演員的戲份在樣板上進行排列。圖 9 為我們所以使用的樣板之一，樣版中的劇照一為第一主角放置位置，我們從最大的特寫鏡頭叢集中挑出一個特寫鏡頭當第一主角，再從其他次大的從集中挑出特寫鏡頭當配角放置於劇照二、劇照三與劇照四等位置，再挑選一張無角色出現的鏡頭當背景，如此便能合成電影海報如圖 10 所示。

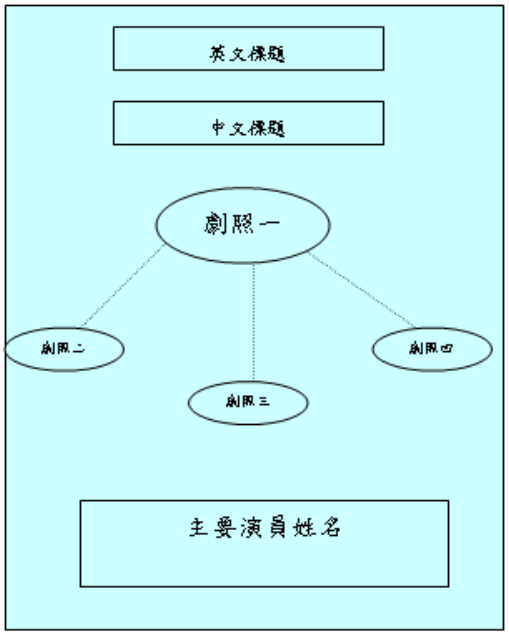


圖 9. 電影海報合成用的樣版

圖 10. 電影“哈利波特 2-消失的密室”合成海報實驗

6.3 實驗環境

為了與米字形方法[11]判斷方法做比較，所以我們採用相同的實驗樣本。將實驗樣本分為兩部分，第一部份為 MPEG-4 所提供的標準範例為實驗樣本，第二部分使用以下四部電影分別代表四種電影類型：

- 臥虎藏龍：武俠片。
- 駭客任務：動畫特效片。
- 鐵達尼：文藝愛情片。
- 神鬼戰士：戰爭格鬥片。

6.4 實驗結果

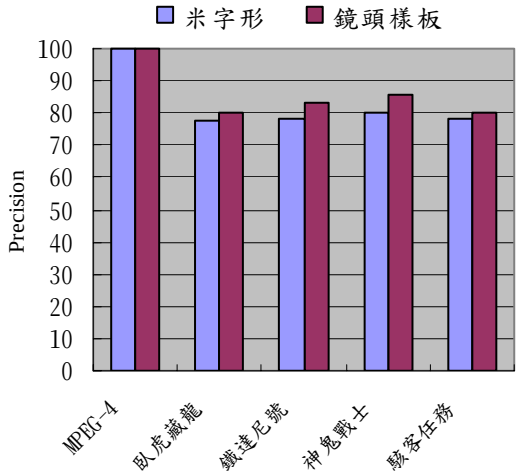


圖 11. 特寫鏡頭辨識之準確率

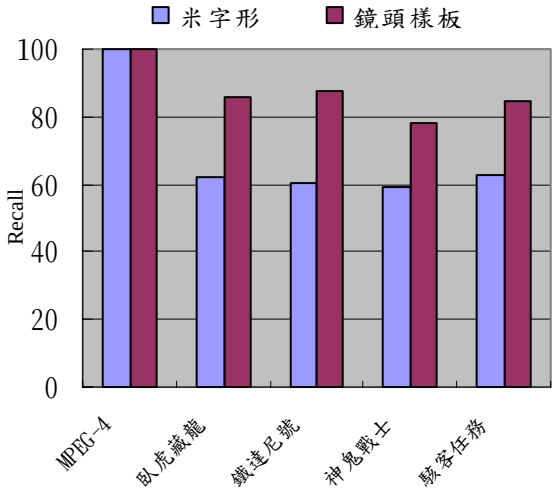


圖 12. 特寫鏡頭辨識之回復率

由我們的實驗結果可以看出使用鏡頭樣板偵測的實驗結果比使用米字形較佳，尤其在回覆率(Recall)的提昇上有明顯改進。

6.5 實驗結果分析

圖 13. 誤判樣本

圖 13 為一近景鏡頭。由於其中只包含一個視訊物件，由於近景鏡頭的物件大小與樣板加分區相近，所以當其位置剛好落在在加分區時，可能會將鏡頭誤判為特寫鏡頭，不過我們可以經由加分區域的加分調整，將其誤判降低。

7. 結論及未來的工作

本文提出一種以樣板比對為基礎的特寫鏡頭偵測方法，能夠有效的偵測到特寫鏡頭，並將其做叢集處理來進行主角的戲份比重分析，並進一步自動合成電影摘要。希望能藉由電影摘要的自動合成，能夠讓使用者快速瞭解到一部電影的內涵，充分發揮電影資料庫的典藏功能。

我們未來的首要工作為提昇鏡頭叢集技術的準確率，並進行大規模實驗。經由提昇叢集技術我們希望除了可以識別演員戲份外，更能進一步判別男演員與女演員，如此便能識別出男主角與女主角，使得電影摘要的合成能更適合電影真正內涵。若能進行大規模實驗，就能藉由實驗的結果修正我們的錯誤，進一步增加我們的準確率與回復率，相對的電影自動化摘要的效果和可靠性也會相對地提高。

8. 參考文獻

[1] "ISO/IEC 14496-2 Information Technology-Generic Coding of Audio-Visual Objects," ISO/IEC , 1998.

[2] Berna Erol, Faouzi Kossentini. "Video object summarization in the MPEG-4 compressed domain", *IEEE International Conference on Acoustics, Speech, and Signal Processing* , Vol.6, pp.2027 -2030, 2000.

[3] Tsuyoshi Moriyama and Masao Sakauchi, "Video Summarization Based on the Psychological Content in the Track Structure." *ACM Multimedia Workshop*, pp.191-194, 2000.

[4] J.H. Oh and Hua K.A."An Efficient Technique for Summarizing Videos Using Visual Contents," *IEEE International Conference on Multimedia and Expo*, Vol.2, pp.1167-1170, 2000.

[5] K. Changick and Jenq-Neng H., "An Integrated Scheme for Object-Based Video Abstraction," in *Proc. of ACM Multimedia*, 2000.

[6] Mei-Juan Chen, Yuan-Pin Hsieh, Yu-Pin Wang, "Multi-Resolution Shape Coding Algorithm For MPEG-4," *IEEE Transactions on Consumer Electronics*, Vol.46, No3, AUGUST 2000.

[7] Saxe, D. and R. Foulds. , "Toward Robust SkinIdentification in Video Images." *In Proceedings of the Second International Conference on Automatic Face and Gesture Recognition*, 379-384,1996.

[8] S. Ahmad, "A usable real-time 3d hand tracker." *Conference Record of the Twenty-Eighth Asilomar Conference on Signals, Systems and Computers*, pp.1257-1261,1994.

[9] Konigsberg, I, *The Complete Film Dictionary*, 2 ed., Penguin Reference, 1997.

[10] Katz, E., *The Film Encyclopedia*, 4 ed., Harper Collins, 2001.

[11] 陳信修、劉志俊"一種利用特寫鏡頭對數位電影資料進行自動化摘要合成之技術" *第一屆數位典藏技術研討會*, pp. 9-16, 2002.

[12] 劉志俊、傅佳源、王志浩、喻仲平, "一種利用物件形狀來進行 MPEG-4 鏡頭變化偵測之技術," *第一屆數位典藏技術研討會*, pp. 17-24, 2002

[13]"電影辭典", 國家電影資料館 1997

[14]"臥虎藏龍寫真集", 東販出版社, 2000.