論文題目：針對新聞影音資料之有效摘要工具

An Effective Summarization Tool for News Videos

作者一：方子維
服務單位：台大通訊與多媒體實驗室
E-mail:ian@cmlab.csie.ntu.edu.tw
作者二：郭晉豪
服務單位:台大通訊與多媒體實驗室
E-mail:david@cmlab.csie.ntu.edu.tw
作者三：吳家麟
服務單位：台灣大學資訊工程學系
E-mail: wjl@csie.ntu.edu.tw

摘要

在本篇論文中，我們提出一個以重要性標示為基礎的影音摘要架構。 在這個架構之下，我們可以從一段數位多媒體影音當中，淬取出一些有意義的資訊，並藉由這些資訊標示出該段影音對於人類視覺的重要性。 利用這些標示出來的重要性，我們就可以依據使用環境/需求/頻寬等因素，在任何時間長度的設定條件下，對該段影音做出有效的摘要。

我們相信這個以影音重要性標示為基礎的架構，可以應用於任何類型的影音資料上。在這個數位多媒體逐漸蔓延的時代，對影音做出摘要已經變得越來越重要。 除了應用於搜尋多媒體資料外， 影音摘要對於數位電視及窄頻多媒體相關應用，亦存在著相當大的助益。

關鍵詞： 摘要 / 新聞 / 重要性 / 標示

第一章 簡介

1.1 什麼是影音摘要

所謂的摘要就是針對一些資料，將其中所包含的意義，以簡短的方式呈現出來，而又不失去其完整的意義，稱之為摘要。對於電腦來說，了解與歸納並不容易，因此最簡單的方式是---取出原有資料所包含最重要意義的一部份，藉以表達全部資料的摘要。 而影音摘要顧名思義就是要從一段影音資料中，取出一部份重要的影音，藉以在較短時間及較少資料量的情況下，完整表達出整段影音原始意義的技術。

1.2 研究動機

隨著多媒體的數位時代來臨，傳統管理媒體資料的方式已不再有效。 於是，以數位內容管理為主題的多媒體國際標準 MPEG-7，在這幾年逐漸風起雲湧，成為非常重要的研究課題，而其管理目的就在於讓人類能夠更有效率地使用多媒體資料。 但是，影音資料包含了非常多的冗餘資料，以至人類必須花費極高的時間與代價來了解，因此，唯有將壓縮過後的影音資料直接呈現給使用者觀賞，方能有助於人類有效率的了解該影音。 不過，這裡所謂的壓縮技術，並非傳統的編碼壓縮技術， 而是前面提到過的影音摘要。 傳統編碼壓縮技術產生的資料，並不是人類所能夠理解的訊號。 唯有透過相關於人類語意(Semantic)的影音壓縮技術，方能有助於人類的理解與吸收。 一旦語意相關的壓縮技術能夠成熟， 影音就可以依照人類使用環境/需求/頻寬等因素，在任何時間

長度的設定條件下做出有效的摘要。 如此, 人類的時間得以節省, 頻寬的限制得以消除, 人類知識傳播的效率與速度將更加的無遠弗屆。

1.3 章節說明

在接下來的章節中, 我們首先在第二章中介紹近十年間, 有關摘要相關技術的研究情況; 在第三章中, 介紹我們所提出來的影音摘要架構; 實作新聞影音摘要系統則在第四章中介紹; 實驗的經驗與結果呈現在第五章; 而第六章則說明了我們的結論與未來努力的方向。

第二章 相關研究

2.1 關鍵視訊頁(Key-Frame)的選擇

理論上, 構成同一個場景的所有視訊頁, 因所處的場景相同, 具有非常相似的特質, 所以, 人們可以利用其中最具代表性的一個視訊頁, 相當程度表達出整個場景視訊的意義。 這個最具代表性的視訊頁, 被稱之為關鍵視訊頁(Key-Frame)。

在早期的相關文獻[2][7]中, 關鍵視訊頁的選擇多半以條列式的規則為基礎(rule-based)。 條列式的規則由於過於死板缺乏彈性, 顯然並無法完全適用於比較多樣性的視訊中。 因此, 另一類以分類技術為基礎的方式[4][5][1][14]被發展出來。 但是, 分類技術在處理上必須計算所有視訊頁兩兩之間的差值, 所以, 計算複雜度明顯會提高很多。

2.2 關鍵場景(Key-Shot)的選擇

當人們意圖對視訊做摘要, 且期望最後的結果能夠以視訊的方式呈現時, 第一個步驟就是關鍵場景的選擇。 因為, 摘要視訊呈現的時間遠遠低於原始視訊, 而視訊又具有不易加速的特質, 所以, 唯一的方式就是找出關鍵而重要的場景單元予以呈現, 而忽略其他不重要的場景單元片段。

在 1999 年, FX Palo Alto 實驗室發表了文獻[10], 其中, 定義了一個非常重要的場景重要性計算方程式, 這個方程式至今仍被廣泛應用:

$$I_j = L_j \log \frac{1}{W_k} \quad (1)$$

式中, I_j表示第j個場景的重要性; L_j表示第j的場景的長度; 而第j個場景屬於第k個分類, W_k表示權重。 權重的公式如下:

$$W_k = \frac{S_k}{\sum_{i=1}^{C} S_i} \quad (2)$$

式中, C表示分類的總數; S_k表示所有屬於第k個分類的場景長度總和。

簡言之, [10]的方程式可選出兩類場景: 一. 長度長的場景; 二. 稀有的場景。 這個計算公式之所以重要, 在於它精確的捕捉到視訊摘要的兩個重要想法:

1. 當人們願意在某個場景花更多的時間來描述, 意味著這個場景的重要性越高。 (I 與 L 成正比。)
2. 視訊摘要必須避免重複性的場景不斷出現, 以降低冗餘資料的量, 因此, 人們傾向選擇與其他相似度低的場景來顯示。 (I 與 $-\log(W_k)$成正比。)

由上述兩點可知, 這個方程式已經與低階視訊特徵完全無關, 而是試圖去捕捉人類對視覺重要性的認知, 進入了語意層級(Semantic Level)的領域。

2.3 影音摘要

摘要工作是人類淬取資料, 以加速了解

資料的重要過程。 在 1995 年, Carnegie Mellon University (CMU)發表了一個真正的影音摘要系統 InfoMedia ,從此引爆了影音摘要技術的相關研究。 在 1998 至 1999 年間,影音摘要進入另一個階段, FX Palo Alto 實驗室在[10]中,定義了 2.2 節中描述的場景重要性計算方程式,並以此選擇重要的場景作為摘要。自此,影音摘要逐步擺脫低階影音特徵的摘要,進入了語意層級(Semantic Level)的摘要。

2002 年, Hari Sundaram 在其博士論文[8]中,將影音資料的語意辨識技術可概略分為兩大類:

1. 由下而上: 以低階影音特徵的組合,找出語意層級的資訊。

2. 由上而下: 藉由對於語意層級的了解,反向尋找哪些特徵表達了這些資訊。

由下而上語意辨識技術的問題在於不容易找到足夠的特徵以完整表達出單一語意。而由上而下的語意辨識技術,則存在著語意項目太多,而無法一一找到對應辨識模型的困擾。顯然,不論是由下而上或由上而下的語意辨識,現階段均不容易完整找出所有語意。 因此,唯有切割出一小部分現階段真正需要的語意加以辨識,才是解決之道。

基於上述理由,本篇論文利用由上而下的方法,試圖藉由對於“重要”這個語意的了解,進行影音摘要。 在稍後的第三章,會有更詳盡的說明。

第三章 摘要架構

本論文目標是利用由上而下的方式,以對於“重要”語意的了解,分析影音中各個視訊頁及場景的重要性,並藉由這些資訊做為影音摘要的基礎。 因此,最重要的步驟就是對於所有視訊頁及場景,依照分析的結果,分別給予一個重要性標示(Importance Measurement)。在本章的各節中,我們將先描述整個架構,然後再分別描述視訊頁及場景重要性標示的方法。

3.1 主架構

對影音資料而言,摘要的第一個步驟是先將視訊及音訊串流分離,以便之後分別進行重要性分析。 假使,視訊與音訊資料具有同步的關係,則影音之間的相關性必須予以考慮(如圖 3-1 所示)。

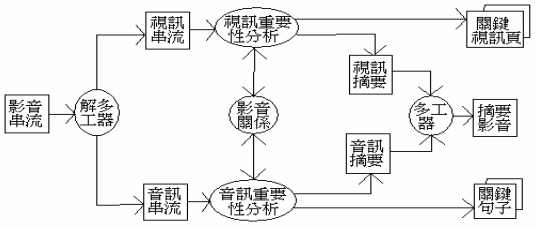

圖 3-1 影音摘要的主架構

以影音摘要而言,圖 3-1 中最重要的步驟是視訊重要性分析及音訊重要性分析兩大模組。 其中,視訊重要性分析的目標是產生關鍵視訊頁及視訊摘要;而音訊重要性分析的目標則是產生關鍵句子及音訊摘要。 最後,再利用多工器將前述的視訊摘要及音訊摘要合併,即可產生影音摘要。

3.2 以視訊頁為基礎的重要性標示及關鍵視訊頁的選擇

由於採用由上而下的方式對“重要”語意進行詮釋,所以首要步驟就是將“重要”語意予以具象化,這個具象化動作的優劣,會直接影響之後重要性標示結果的成功與否至關重要。

所謂具象化就是將語意層次的概念,直接映射於實際的事件上,讓這些事件的發生與語意概念產生直接的聯繫。 一般來說,具象化過程所採用的事件,與欲進行摘要的影音特質

有關(Domain-Dependence)。 對未分類影音具象化的最簡單方法,就是直接去追蹤在影音製作過程中,相關人物對於該影音所表達出來的重要性意念。 由於這些意念是由影音製作的參與者直接抒發出來,因此對於“重要”語意的詮釋也就相當完美。(在稍後第四章中,將對具象化的細節進行說明。)

一般來說,語意經過具象化後產生的事件可分成兩大類:

1. 連續事件(Continue Event): 表示可以在一段時間內,連續多個視訊頁中被檢測到的事件。 例如: 人臉事件。

2. 突發事件(Instant Event): 表示該事件只能夠在某瞬間被人類檢測到,或是人類對於該事件只有指定某瞬間重要性的能力。 例如: 鎂光燈事件。

其中,突發事件是相當弔詭的。 照理來說,在同一個場景中,事件應該是連續的,而不應該有突發的情況發生,之所以產生突發的情況,並不是因為事件本身的特質如此,而導因於人類對於影片中事件檢測能力不足所致。

以前面提到的鎂光燈事件為例,人類可以輕易檢測到鎂光燈閃爍,因而判定某一個視訊頁出現了鎂光燈事件,並藉以對該視訊頁指定一個重要性標示。 但是,由於事件應該連續發生,因此,對於該場景中的每一個視訊頁,我們應該給予如圖 3-2b 的重要性標示,而非圖 3-2a 所顯示者。 也就是說,當檢測到突發事件後,必須先經過一個模擬的程序,將圖 3-2a 轉換為 3-2b,才能夠充分表達出事件在該場景中的重要性標示。

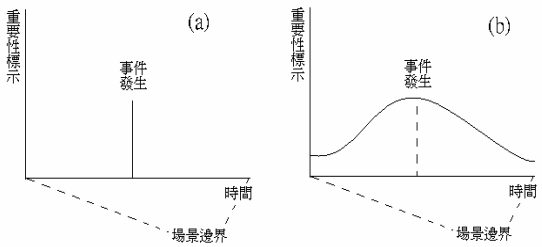

圖 3-2 突發事件的重要性標示

所謂的轉換過程使用的是一個 Mask 函數:

$$M = \exp(\frac{-\Delta t^2}{2\sigma}) \quad (3)$$

由上式可知,經由上述 Mask 函數的處理,重要性標示的值將會隨著距離事件發生時間點越遠,而以指數函數的型態衰減,而衰減的速率則由σ值決定:

$$\sigma = \frac{Const.}{Dyn} \quad (4)$$

式中, Const.為常數, Dyn 則表示鏡頭移動的快慢。 在本論文中,同一場景使用一個固定的 Dyn 值,而該 Dyn 則由視訊頁間的 GMV(Global Motion Vector)決定:

$$Dyn = \sum \frac{1}{N}\sqrt{Tx^2 + Ty^2 + (\frac{Zoom}{Const.})^2} \quad (5)$$

上式中,(Tx,Ty,Zoom)表示 Zoom Model 的 GMV 參數,而 N 則表示場景中的視訊頁個數。

有了以上對連續及突發事件本質的了解後,我們就可以對視訊頁的重要性標示進行定義:

$$FI = w_c FI_c + (1 - w_c) FI_i \quad (6)$$

式中, FI_c及FI_i分別表示對於某個視訊頁連續及突發事件所計算出來的重要性標示, w_c及$(1-w_c)$則表示權重。 而FI_c及FI_i由則下式分別定義:

$$FI_c = L * \sum_i C_i \quad (7)$$

對某個視訊頁而言,上式中的C_i分別表示對於連續事件i的重要性標示值;至於L則表示該視訊頁所在場景的時間長度,這個值是為了定義場景重要性標示而設的,在下一節中將再次提及。

$$FI_i = \sum_t \left(\sum_j b_j L_j I_j \right) * \exp(\frac{-(t-t_{cur})^2}{2\sigma}) \quad (8)$$

上式中,t_{cur}表示目前正在計算重要性標示值的視訊頁在時間軸上的時間;t則表示其他在同一場景中各視訊頁在時間軸上的時間。 b_j是布林函數,分別表示突發事件j是否發生;而I_j則表示我們賦予各個突發事件的重要性標示值;至於L_j則表示時間長度,這個值同樣是為了定義場景重要性標示而設的,在下一節中將詳細描述。

綜合上述對兩類事件的相關定義,在同一場景中所檢測出來的事件,將會被轉化為對每一個視訊頁的重要性標示,流程如圖 3-3 所示。

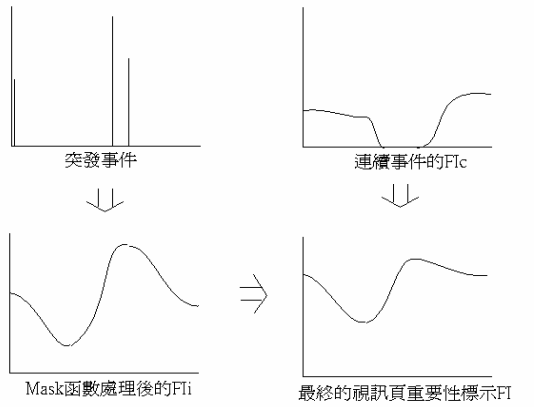

圖 3-3 視訊頁重要性標示處理流程

最後,在每個場景當中,重要性標示值 FI 最大的一個視訊頁,即被選擇成為該場景的關鍵視訊頁。

3.3 以場景變換為基礎的重要性標示及摘要視訊的產生

在第二章曾經介紹過,[10]發表了一個非常重要的場景重要性定義公式:

$$I_j = L_j \log \frac{1}{W_k} \quad (9)$$

因為這個公式與語意層級的概念十分吻合,因此本篇論文決定以上述公式為基礎,做一個細部的修正。 而在 3.2 節中,我們已經對於各個場景中視訊頁定義了重要性標示,因此,上述公式可以被演化如下:

$$SI = FI * \log(\frac{1}{W_k}) \quad (10)$$

上式中,SI為場景的重要性標示值;FI為該場景的關鍵視訊頁所擁有的FI值;而W_k則保持原來的定義(權重)。 比較值得注意的是---為了保留原公式的意義,時間長度的資訊被隱含在FI當中,這也就是在 3.2 節中,FI_c及FI_i公式中必須包含時間長度(L)的理由。

經由上述公式的計算,場景重要性標示值 SI 被標示完成後,即可由重要性最高的場景開始選擇,直到時間限制(Time Constraint)到達為止,再將選出來的所有場景合併,可得最終的視訊摘要。 當然,在選擇的過程中,必須考慮到視訊的可壓縮性/影音之間的關係及摘要的時間限制,不過由於這三項特性與影音的型態有關(Domain-Dependence),故在第四章中一併做說明。

第四章 新聞影音摘要之實作

4.1 新聞的“重要”語意具象表徵

欲具象化新聞影片的“重要”語意,下

列四類人物對於“重要”的表達方式,必須予以追蹤:

1. 攝影師: 在新聞影音中,攝影師相當程度就扮演著導演的角色。

2. 播報員: 播報員扮演著聲音化妝師的角色。

3. 後製人員: 新聞影片最後會經過後製人員的摘要及剪輯處理,並為視訊加上文字旁白。

4. 新聞現場的人物: 所謂新聞現場的人物包含真正的新聞主角,亦包含其他正在現場的人物。 這些人物的臨場反應,往往能適切的表達出新聞的主要意含。

藉由對上述四類人物的反應及行為追蹤,可以歸納出下列事件的發生,具有非常強烈的重要性意義:

1. 突發事件: 鎂光燈事件(Flash Event)/景深變換事件(Zoom Event)/靜態事件(Static Event)/平移事件(Pan Event)/開始結束事件(Start-End Event)

2. 連續事件: 人臉事件(Face Event)/文字事件(Text Event)

4.2 新聞影音摘要的產生

本論文所提出的新聞摘要系統,基本上是以第三章設計的架構為主體發展而來,但鑒於新聞語音具有其非常特殊的性質,因而必須略作調整。 新聞摘要系統的主架構即是圖 3-1,但事實上有兩個模組的內容被依照新聞的特性簡化了:

1. 在視訊與音訊無法明確同步的狀況下,影音關係模組被大幅弱化。

2. 一則新聞影音一般都是以一段主播畫面加上數段外場影音合併而成。 其中,主播畫面中所搭配的聲音,幾乎就是該則新聞最完美的音訊摘要。 基於已經擁有人類辛苦所製作出來的摘要音訊,因此,音訊重要性分析模組也被大幅簡化。 圖 4-1 即簡化後的音訊重要性分析模組。

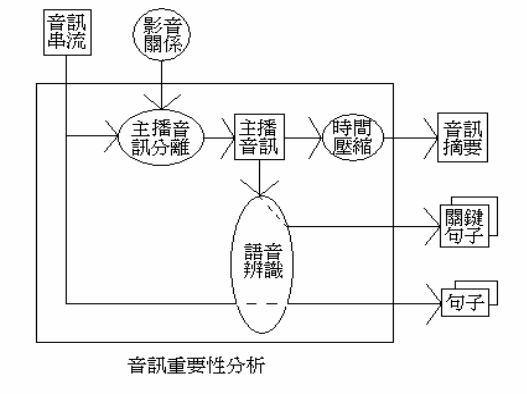

圖 4-1 新聞音訊重要性分析架構

圖 4-1 中,時間壓縮在本系統中採用線性加速 1.5 倍。 換言之,最終的影音摘要時間長度限制(Time Constraint),將是主播場景時間長度的 1/1.5 倍。

視訊重要性分析模組的處理如下:

1. 由於影音不需同步,在關鍵場景的選擇上,就不需要考慮在相同時間點上的音訊情況,只需確定所有選取的場景經壓縮後,所需的呈現時間符合上述時間長度限制(Time Constraint)即可。

2. 基於影音不需同步,視訊壓縮率就不須受限於音訊的兩倍以下壓縮率限制,而可以達到二至五倍的壓縮效果。 至於視訊壓縮率究竟應該設計為多少才合理,則由場景中鏡頭的移動速度決定,意即加速倍率(Factor)為:

$$Factor = \frac{Const.}{Dyn} \quad (11)$$

式中,Const.為常數,Dyn 則定義於第三章。

整個新聞摘要架構的示意圖如圖 4-2 所示。

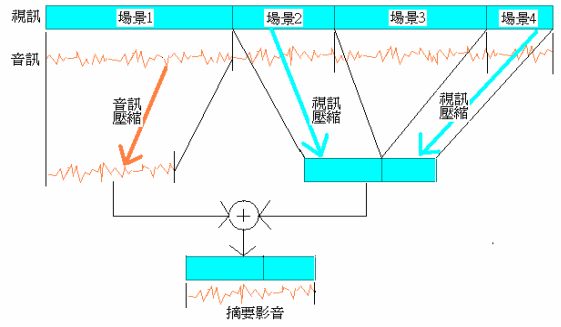

圖 4-2 新聞摘要系統的示意圖

4.3 瀏覽系統

新聞瀏覽系統被設計成兩層的結構，用來呈現前面章節描述方法所產生的影音摘要/關鍵視訊頁/關鍵句子…等資訊。

1. 第一層：摘要瀏覽(如圖 4-3 所示)

圖 4-3 新聞瀏覽系統第一層

2. 第二層：進階摘要與使用(如圖 4-4 所示)

圖 4-4 新聞瀏覽系統第二層

第五章 實驗與結果

實驗的條件如下：

1. 時間：20~25 分鐘
2. 人數：10 人
3. 影音資料：中視 2002 年 9 月 19 日新聞十一則（二百四十九個場景）

5.1 關鍵視訊頁的結果

關鍵視訊頁的實驗程序：

1. 先顯示系統對於某個場景所選擇的關鍵視訊頁，再播放該場景的影音資料。
2. 允許使用者重複觀看上述關鍵視訊頁及場景影音資料。
3. 要求使用者對系統選擇給予極佳/良/可/不良/低劣/無法判定六種評價之一。

表 5-1 關鍵視訊頁的實驗結果

評價	極佳(100)	良(80)	可(60)	不良(40)	低劣(20)	計分
總計	253	281	224	132	24	73.28

分析上述被給予不良及低劣評價的場景，原因多數在於語意問題---意即對於同一個畫面，不同人會有不同的關注點。這類問題並無完美的解決方案，只能利用學習機器(Learning Machine)試圖找出多數人的喜好。

5.2 影音摘要的結果

影音摘要的實驗程序：

1. 先播放摘要影音給受測者觀看，再播放原始影音給受測者觀看。
2. 要求使用者對摘要影音給予極佳/良/可/不良/低劣五種評價之一。

新聞	極佳(100)	良(80)	可(60)	不良(40)	低劣(20)	計分
總計	24	39	37	9	1	73.82

表 5-2 影音摘要的實驗結果

分析上述實驗的結果，大致可以知道---當摘要影音包含原始影音的場景比例提高時，使用者往往會給予較高的評價. 此外，如果原始影音本身就比較不具吸引力時，使用者將受

到影響而傾向給予較低的評價.

第六章 結論與未來

6.1 結論

本篇論文提出了一個由上而下, 以“重要”語意為基準的影音摘要架構。 由於這個架構被設計成與應用領域相關(Domain-Dependence), 因此它具有極高的可塑性。在第四章中, 我們以此架構對“新聞”領域的影音實作了一個摘要系統。 由實作的結果來看, 這個架構的確可以相當程度反映出人類對於“重要”的看法。 而這種語意層級概念的檢測, 正是過去許多相關研究較不完備的部分。也許這樣的結果仍不夠完備, 但相信已經提供了一個相當有意義的研發方向可供參考。

6.2 未來研究方向

本篇論文仍有許多值得探討與未來努力的地方:

1. 本篇論文以重要性事件來驅動摘要行為, 因此, 唯有發展出更多的事件檢測能力, 才能夠更完整的了解“重要”語意的所在, 而藉以發展的摘要系統, 才能夠逐漸提昇其表達重要語意的能力。

2. 在突發及連續事件的重要性標示上, 目前以直覺的方式直接指定一個常數值, 用以成為這些事件重要性標示值計算的基礎。這些常數值未來應該以學習機器(Learning Machine)重新加以定義。

誌謝

1. 感謝國科會提供研究經費上的協助。

2. 感謝台灣大學李琳山教授提供語音辨識模組。

3. 感謝 Intel 提供 OpenCV 人臉辨識模組。

4. 感謝台灣大學曹智富同學提供文字檢測模組。

參考文獻

[1] X. Gao and X. Tang, "Automatic Parsing of News Video Based on Cluster Analysis," in *Proc. of 2000 Asia Pacific Conference on Multimedia Technology and Applications*, Taiwan, Dec. 2000

[2] Maybury, M. and Merlino, A., "Multimedia Summaries of Broadcast News", *International Conference on Intelligent Information Systems*, 1997

[3] Yu-Fei Ma, Lie Lu, Hong-Jiang Zhang, Mingjing Li, “An Attention Model for Video Summarization”, *ACM Multimedia* 2002

[4] O'Connor, N., Marlow, S., Murphy, N., Smeaton, A., Browne, P., Deasy, S., Lee, H. and Mc Donald, K. “Fischlar: an On-line System for Indexing and Browsing of Broadcast Television Content”, In *Proceedings of ICASSP* 2001

[5] N. O’Connor, C. Czirjek, S. Deasy, S. Marlow, N. Murphy, A. Smeaton, “NEWS STORY SEGMENTATION IN THE F´ISCHL´ AR VIDEO INDEXING SYSTEM”, *IEEE ICIP* 2001

[6] L. Rau, R. Brandow, and K. Mitze, “Domain-independent summarization of news”, In *Summarizing Text for Intelligent Communication*, 1994

[7] M. Smith and T. Kanade, “Video Skimming for Quick Browsing based on Audio and Image Characterization”, *tech. report, Carnegie Mellon University*, 1995

[8] Hari Sundaram, “Segmentation, Structure Detection and Summarization of Multimedia Sequences”, *PhD thesis, Dept. Of Electrical Engineering, Columbia University*, NY NY, Oct. 2002

[9] Shingo Uchihashi and Jonathan Foote, “Summarizing Video using a Shot Importance Measure and a Frame-Packing Algorithm”, In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, 1999

[10] S. Uchihashi, J. Foote, A. Girgensohn, and J. Boreczky, “Video Manga: Generating Semantically Meaningful Video Summaries”,

In *Proceedings ACM Multimedia*, 1999

[11] N. Vasconcelos, & A. Lippman, “A Spatiotemporal Motion Model for Video Summarization”, *CVPR*, 1998

[12] H. Wactlar, “Informedia - Search and Summarization in the Video Medium”, In *Proceedings of Imagina* 2000

[13] Li, Ying; Zhang, Tong; Tretter, Daniel, “An Overview of Video Abstraction Techniques”, *HP Labs Technical Reports*, 2001

[14] Y. Zhuang, Y. Rui, T. S. Huang, and S. Metrotra, “Aaptive key frame extraction using unsupervised clustering”, In *Proceeding of IEEE Int. Conf. on Image Processing*, 1998