行政院國家科學委員會補助產學合作研究計畫成果完整報告

媒體內容工程：MPEG-4/7 相關技術之研發(III)

計畫類別： 產學合作研究計畫
計畫編號： NSC 92－2622－E－002－002
執行期間： 92 年 08 月 01 日至 93 年 07 月 31 日

計畫總主持人：吳家麟 台灣大學資訊工程所教授
共同主持人：項 潔 台灣大學資訊工程所教授
子計畫主持人：陳文進 台灣大學資訊工程所教授
吳家麟 台灣大學資訊工程所教授
歐陽明 台灣大學資訊工程所教授
陳炳宇 台灣大學資訊管理所助理教授
黃肇雄 台灣大學資訊工程所教授
周承復 台灣大學資訊工程所助理教授

處理方式：完整報告內容因涉及專利、技術移轉案或其他智慧財產權，不予公開。

執行單位：國立臺灣大學資訊工程系 通訊與多媒體實驗室
訊連科技股份有限公司
太極影科技股份有限公司

中 華 民 國 93 年 09 月 15 日

媒體內容工程：MPEG-4/7相關技術之研發(III)

產學合作計畫成果報告(總計畫)

成果報告

計畫編號：NSC91-2622-E-002 –002

全程計畫：民國 90 年 08 月 01 日至民國 93 年 07 月 31 日

本年度計畫：民國 92 年 08 月 01 日至民國 93 年 07 月 31 日

計畫總主持人 ：吳家麟　台灣大學資訊工程系教授
共同主持人 ：項　潔　台灣大學資訊工程系教授
子計畫主持人 ：陳文進　台灣大學資訊工程系教授
吳家麟　台灣大學資訊工程系教授
歐陽明　台灣大學資訊工程系教授
陳炳宇　台灣大學資訊管理系助理教授
黃肇雄　台灣大學資訊工程系教授
周承復　台灣大學資訊工程系助理教授

一、摘要

（中文摘要）

"媒體內容工程: MPEG-4/7 相關技術之研發" (Content Engineering: Research on MPEG-4/7 Multimedia Technologies) 研究計畫的目的，在於發展一個完整的媒體內容整體架構，涵蓋媒體內容生命週期的每個階段，包括了建構(creation)、儲存(storage)、搜尋(search)、處理(manipulation)、管理(management)、傳遞(delivery)、呈現(presentation)以及互動(interaction)等過程，並符合 MPEG-4 及 MPEG-7 國際標準之規範。

繼 MPEG-1 與 MPEG-2 國際影音壓縮標準成功地帶動整個影音消費市場的成長後；MPEG 標準制定群於 1994 年正式展開下一個國際標準 MPEG-4 的制定，並已經在 1998 年 12 月完成此標準。在電腦(computer)、通訊(communication)、消費性電子(consumer electronic) 以及媒體內容(contents) 整合應用的趨勢之下，MPEG-4 國際標準主要著眼於：物件導向的場景編排及編碼、真實與虛擬的複合媒體編碼、大量媒體資料的遠端控制及傳輸、異質傳輸環境中的單一傳輸界面、以及高度的人機互動等課題，並以發展出可與各式媒體物件進行互動的各種多媒體應用為目標。

另一方面，由於數位影音技術的快速發展，使得數位相機、DVD 放影機、MP3 隨身聽等數位視聽設備成為消費市場的寵兒，大量的數位化媒體內容與日常生活緊密地結合；面對這波媒體內容數位化的趨勢，亟需要一個能夠描述媒體內容的共通標準，使得各種不同的應用系統以及媒體裝置，能夠根據描述媒體內容 metadata 所提供的資訊，正確而有效地存取(access)、交換(exchange)以及再用(re-use)所需的媒體內容。有鑑於此，MPEG 標準制訂群於 1996 年 10 月正式展開了制定 MPEG-7 國際標準的工作，此標準又稱為"多媒體內容描述介面" (Multimedia Content Description Interface)，並已於 2001 年 9 月制訂完成。

本計畫將研發多項與媒體內容相關的技術模組，包括：MPEG-4/7 互動媒體伺服器、MPEG-4 互動媒體場景編輯器、MPEG-4 互動媒體播放器、三維 MPEG-4 場景動畫建構工作台、三維物件搜尋器、影音特徵值粹取工具集，內容分析與摘要模組、資訊串流模組、傳送 MPEG-4 資料的 MPEG-2 傳輸流模組以及 MPEG-4/7 保證服務品質的媒體內容代理伺服器等。並利用上述技術模組，開發兩項與媒體內容工程相關的應用系統： "互動資訊媒體服務系統（Interactive Information Media Service System）" 與 "家庭媒體中心（Home Media Center）"。

本計畫將以個人電腦(PC)為發展平台，分三年完成上述技術模組的研發以及兩項與媒體內容相關的系統。

(英文摘要)

The project, "Content Engineering : Research on MPEG-4/7 Multimedia Technologies", aims at develop a complete multimedia content framework which covers every phase of the content manipulation lifecycle, including the creation, storage, search, manipulation, management, delivery, presentation and interaction of multimedia content. The framework will conform to the MPEG-4 and MPEG-7 standards.

Due to the great success of MPEG-1 and MPEG-2 in the A/V market, the ISO MPEG committee continued to develop the next multimedia standard - MPEG-4. The standard has been completely designed in December 1998. Under the trend of the integration of computer, communication, consumer electronic, and contents, MPEG-4 targets on the following issues: authoring and encoding of object-oriented scene, encoding of natural and synthetic media, media streaming, uniform transmission interface in heterogeneous network environments, and enhanced human-machine interactions. Therefore, MPEG-4 is designed to be applicable for a wide spectrum of applications.

In addition, resulting form the great technical achievement audio-video technologies has made digital camera, DVD player, MP3 player, dominate the consumer market. To satisfy the customer needs, a tremendous amount of multimedia contents have been produced around world. In this trend, we need a standard way to describe the multimedia content which can make the access, exchange, and re-use of contents more effective and more efficient. In October 1996, MPEG committee started a new project to develop the MPEG-7 international standard, called "Multimedia Content Description Interface", which has been completed by September 2001.

In this project, a number of techniques will be developed. The techniques include: MPEG-4/7 interactive server, MPEG-4 interactive scene editor, MPEG-4 interactive player, audio/video features extraction tools, content analysis and summarization nodule, content streaming module, 3D workbench for MPEG-4 scene/animation construction, query by example 3D, MPEG-2 transport stream for MPEG-4 data and MPEG-4/7 Qos-supported proxy server and so on. Base on these techniques, two applications, "Interactive Information Media Service System" and "Home Media Center", will be developed.

The project is to be developed on PC platform. It is scheduled to be completed in three years.

二、計畫緣由與目的

本研究計畫--"媒體內容工程：MPEG-4/7相關技術之研發"將以媒體內容工程(Content Engineering) 的相關議題為研究主軸，探討在媒體內容的生命週期中，建構、分析、編輯、傳送以及呈現等階段所面臨的各種問題，輔以MPEG-4以及MPEG-7這兩套國際標準為經緯，發展必要的核心技術並尋求解決方案。我們預期開發多項技術模組，並將其整合成兩個應用系統："互動資訊媒體服務系統(Interactive Information Media Service System)"及"家庭媒體中心(Home Media Center)"。

時至今日，"文件(document)"一詞給人們的固有印象，早已被大量的數位電子書、影音光碟、串流視訊、虛擬實境等等現代化的"電子文件(electronic document)"所取代。這些數位化之後的資訊，透過網路無遠弗屆地傳送到世界各地，數位化的媒體內容具備容易修改、複製、搜尋、傳遞與儲存等特性，並擁有多樣化的呈現方式，甚至包括雙向互動等傳統文件所欠缺的功能。身處在這樣一個資訊爆炸的時代，如何有效地管理媒體內容？如何正確地對媒體內容做分類？如何將分類後的媒體內容傳送到不同環境的使用者眼前？如何以符合使用者需求的方式呈現媒體內容？如何有效地進行知識管理（knowledge management）的工程？面對這些亟需解決的議題，我們必須針對媒體內容價值鏈的每個階段，研發必要的多媒體技術並尋求整體的解決方案，這個解決方案我們稱之為"媒體內容工程(Content Engineering)"。

媒體內容工程是近年新興的一門學問，涵蓋的相關技術與專業領域十分廣泛，包括了影音訊號數位化、文件電子化、視訊串流技術、內容分析、資訊檢索、多媒體資料庫、數位浮水印、電子認證以及虛擬實境等等，並橫跨了多項與媒體內容息息相關的媒體產業，如電視/電影工業、媒體工作者、後製作公司、有線/無線系統業者，以及開發數位影音設備，行動裝置與影音軟體的高科技軟硬體產業；此外，媒體的收費機制更與電子商務、網路安全等產業有著密不可分的關係。

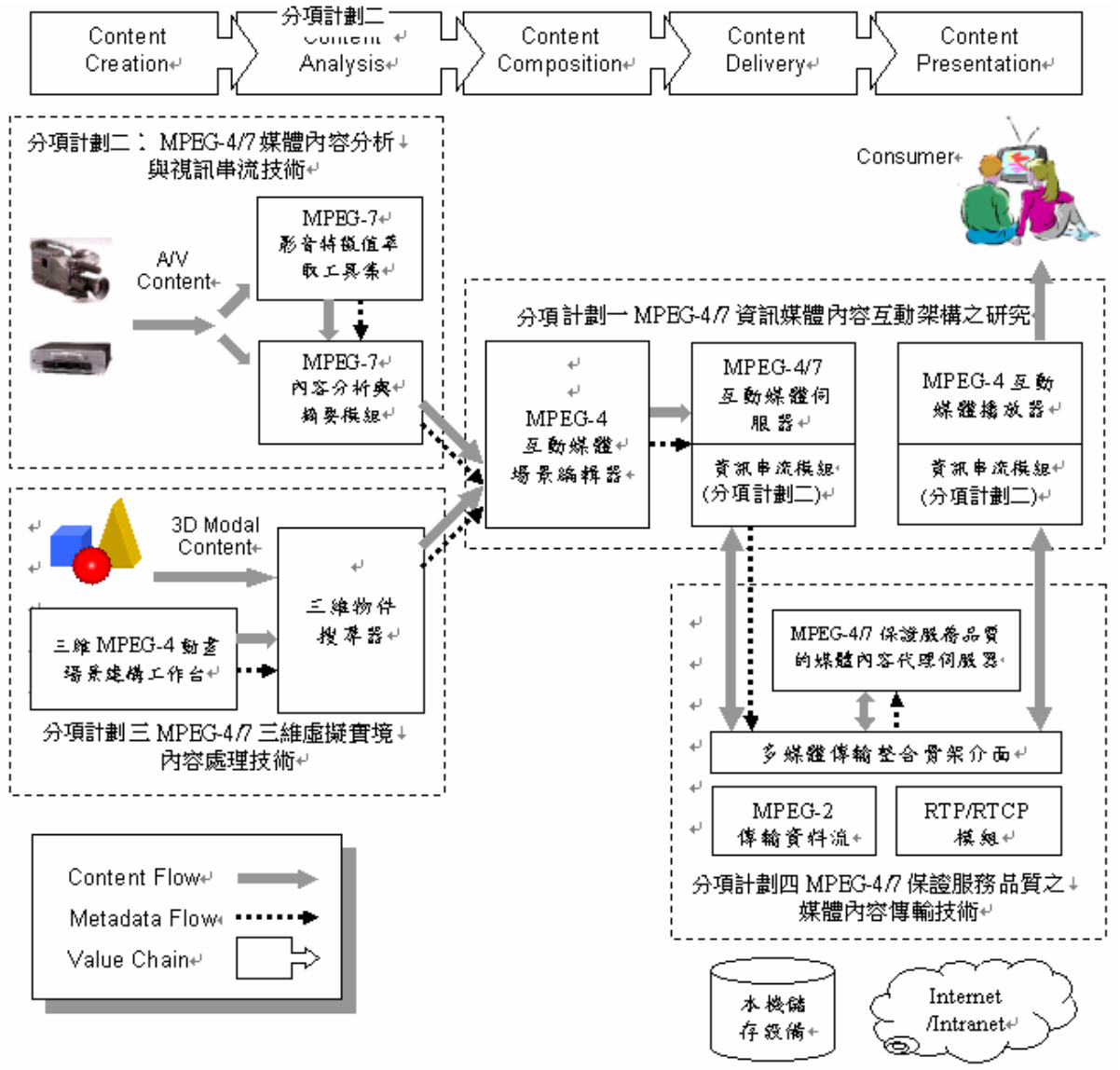
自MPEG-1、MPEG-2標準制定之後，已成功且廣泛地應用在數位通訊及娛樂消費市場中；而和MPEG-1/2不同的是，MPEG-4不再只是以單張畫面為單位來處理視訊資料，而是以物件導向(object-oriented)的概念為主，將每個畫面場景視為不同媒體物件的組合，如：音訊(audio)、視訊(video)、二維/三維圖形(2D/3D graphics)等。MPEG-4中強調真實與虛擬複合媒體(synthetic and natural hybrid media)的合成，除了可經由混合編碼方式達到低位元率的資料傳輸率外，並可提供更多的互動性。MPEG-4的相關技術已經日趨成熟，進入市場的契機也逐漸明朗化。去年六月，IEEE 電路與系統學會(Circuit and Systems Society)與MPEG-4產業聯盟(MPEG-4 Industry Forum)共同於聖荷西(San Jose)市舉辦之第三屆"Workshop and Exhibition on MPEG-4"。參展廠商的研發成果無論是質與量都極為可觀，此為近期內最受重視之有關 MPEG-4 技術與產業現況的成果發表會。本計畫將累積前期產學計畫在研發MPEG-4技術上的經驗，針對MPEG-4標準與媒體內容工程間的相關課題進行更深入的研究，相信對提升國內MPEG-4相關技術的水平，將有最直接的貢獻。

而MPEG-7則是針對媒體內容描述方式而制訂的標準，藉由定義一系列的影音(Audio-Visual)述詞(Descriptor)與描述結構(Description Schema)，MPEG-7將多媒體內容所能表示的metadata加以標準化，一旦metadata之間能夠流通並加以運用，更多更豐富的媒體內容應用將走進日常生活之中。此外，MPEG-7的描述定義語言(Description Definition Language)，對於定義metadata的方式提供了擴充的彈性，這將使得MPEG-7能因應新型態的媒體內容而加以更新。與MPEG-7相關的研究，牽涉的技術層面十分廣泛而重要，本計畫將著重在影像(Video)媒體的分析與摘要，再配合聲音(Audio)以及圖像(Image)分析的輔助，以期建立一符合人類知覺(human perception)、結構化的連續影像物件模型。此外，本計畫將就三維物件的分類與搜尋進行研究，這正是MPEG-7較為欠缺的部分；我們將研究適當之三維模型述詞(Descriptor)與描述結構(Description Schema)，並朝進入MPEG-7 標準而努力。相信對提昇國內MPEG-7與媒體內容分析的研究水平，有正面的助益。

"媒體內容工程：MPEG-4/7相關技術之研發"計畫中，擬實作兩項應用系統—“互動資訊媒體服務系統(Interactive Information Media Service System)”與”家庭媒體中心(Home Media Center)”。此兩項應用系統係整合了下列十項技術模組：

1. MPEG-4/7 互動媒體伺服器(MPEG-4/7 interactive server)
2. MPEG-4互動媒體播放器(MPEG-4 interactive player)
3. MPEG-4互動媒體場景編輯器(MPEG-4 interactive scene editor)
4. 影音特徵值粹取工具集(Audio/video feature extraction tools)
5. 內容分析與摘要模組(Content analysis and summarization module)
6. 資訊串流模組(Content streaming module)
7. 三維MPEG-4場景動畫建構工作台(3D workbench for MPEG-4 scene and animation construction)
8. 三維物件搜尋器(Query by example in 3D)
9. 傳送MPEG-4資料的MPEG-2傳輸流模組(MPEG-2 transport stream for MPEG-4 data)
10. MPEG-4/7保證服務品質的媒體內容代理伺服器(MPEG-4/7 Qos-supported proxy server)

十項技術模組將由四個分項計畫完成，而其功能區塊如下圖所示：


Content Creation
分項計劃二
Analysis
Content Composition
Content Delivery
Content Presentation
分項計劃二：MPEG-4/7 媒體內容分析與視訊串流技術
Consumer
A/V Content
MPEG-7 影音特徵值萃取工具集
MPEG-7 內容分析與綱要模組
分項計劃一 MPEG-4/7 資訊媒體內容互動架構之研究
MPEG-4 互動媒體場景編輯器
MPEG-4/7 互動媒體伺服器
資訊串流模組 (分項計劃二)
MPEG-4 互動媒體播放器
資訊串流模組 (分項計劃二)
3D Modal Content
三維物件搜尋器
三維 MPEG-4 動畫場景建構工作台
分項計劃三 MPEG-4/7 三維虛擬實境內容處理技術
MPEG-4/7 保證服務品質的媒體內容代理伺服器
多媒體傳輸整合骨架介面
MPEG-2 傳輸資料流
RTP/RTCP 模組
分項計劃四 MPEG-4/7 保證服務品質之媒體內容傳輸技術
本機儲存設備
Internet /Intranet
Content Flow
Metadata Flow
Value Chain

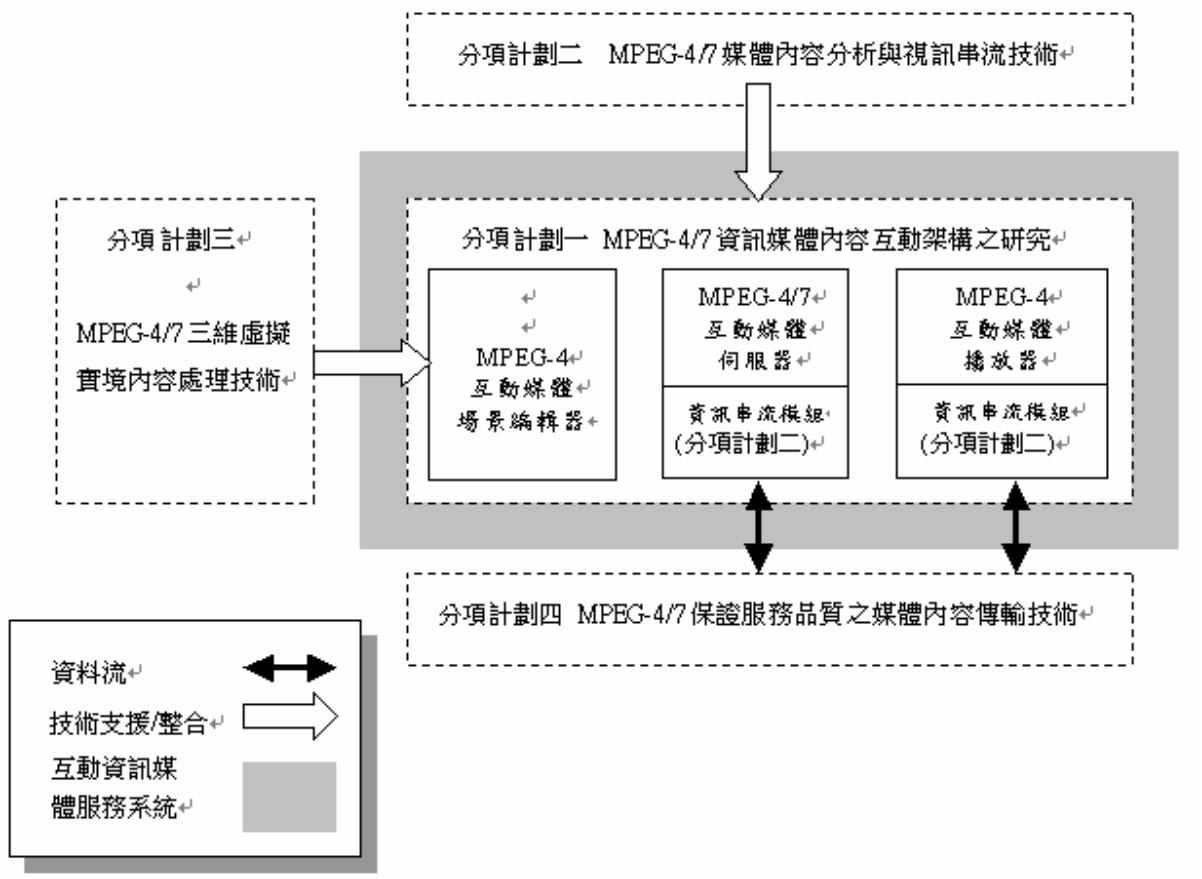
三、研究方法及成果

（研究方法）

為避免因所涵括的領域過於廣泛而使整個計畫失去焦點，本計畫將以開發兩套應用系統—“互動資訊媒體服務系統(Interactive Information Media Service System)”與”家庭媒體中心(Home Media Center)”為目標，亦即所發展出的上述 MPEG-4/7 技術模組，將用以開發此兩套系統。此兩項應用系統將由總計畫作整體規劃，以整合四個分項計畫所開發出的技術成果。實作部份，“互動資訊媒體服務系統(Interactive Information Media Service System)”將由分項計畫一負責系統之整體互動架構，並輔以分項計畫二所發展之媒體內容分析工具以及分項計畫三之場景建構平台，並由分項計畫二及四提供媒體內容傳輸所需之串流技術及網路服務；而”家庭媒體中心(Home Media Center)”將以分項計畫二及分項計畫三為主，開發 MPEG-7 相關之媒體內容分析工具以及三維物件搜尋器，俾以提供分項計畫一在建構 MPEG-4 動畫場景與分項計畫四在發展媒體傳輸技術時之應用。

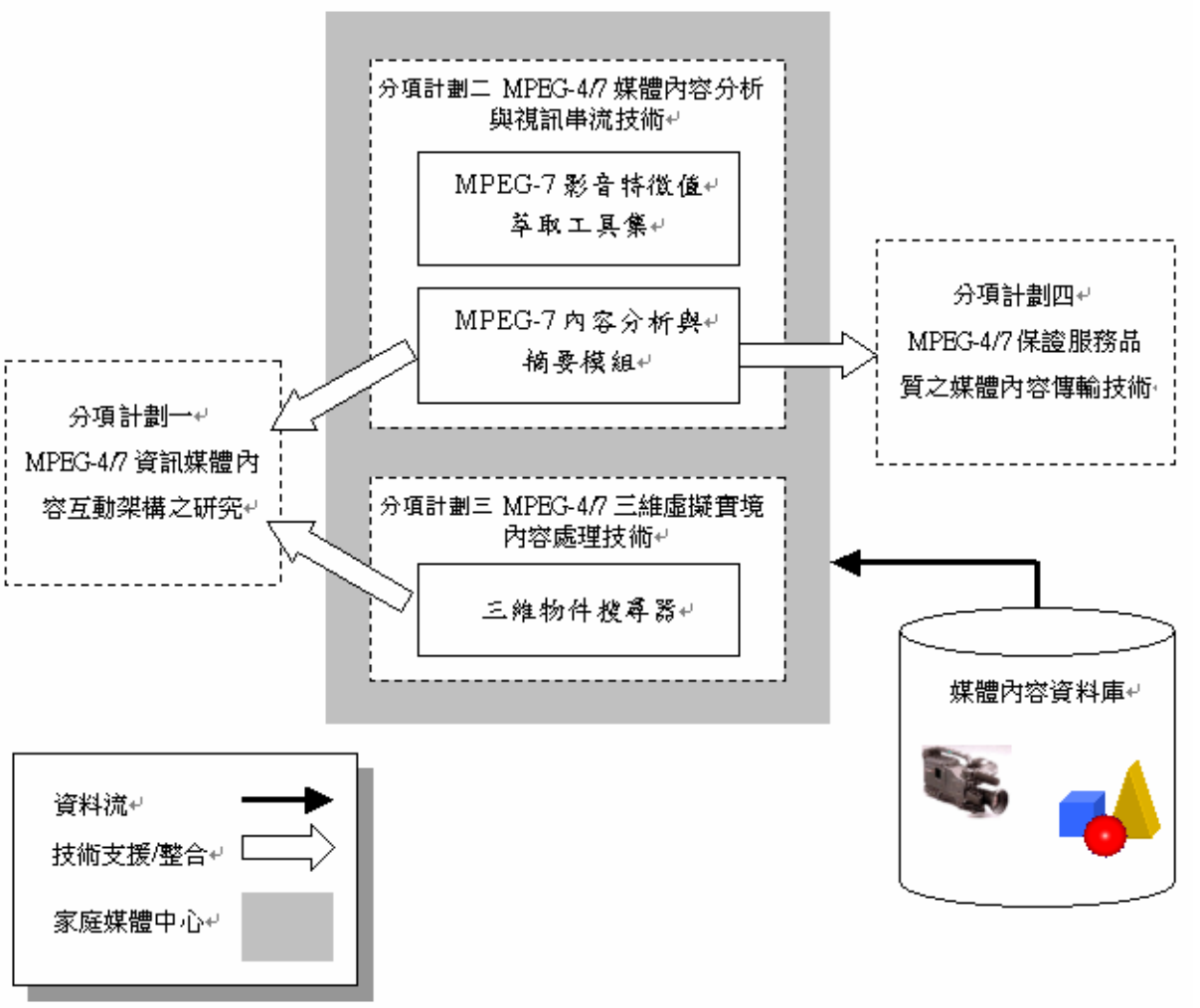
在"互動資訊媒體服務系統(Interactive Information Media Service System)"方面，由於電子產業與電視工業的蓬勃發展，使得數位電視(Digital TV)即將取代類比電視，成為新一代資訊家電(Information Appliance)的要角。我們可預見在不久的將來，電視不再只是單純地播放節目的機器，而進一步能提供多方面的資訊服務；更重要的是，它將顛覆被動地收看系統業者所提供頻道的傳統收視習慣。在互動媒體的架構下，使用者可以根據其需求，選擇最適合的節目觀看；或是查詢所需的資訊(例如某上市公司的即時股價)。節目(內容)提供者，也可以透過與使用者的互動過程，建立用戶收視習慣的資料庫，以期提供個人化(Personalization)收視的服務；或是設計各種互動式的節目，例如互動式電視遊戲、互動式電視教學等等。

本系統將以 MPEG-4 作為互動機制的解決方案，提供完整地互動平台，包括了分項計畫一中的【MPEG-4 互動媒體場景編輯器】、【MPEG-4/7 互動媒體伺服器】以及【MPEG-4 互動媒體播放器】等模組；其中，【MPEG-4 互動媒體場景編輯器】將與分項計畫三之【三維 MPEG-4 場景動畫建構工作台】整合；伺服器與播放端之間的網路傳輸服務則由分項計畫四負責提供。另一方面，由分項計畫二所開發的【MPEG-7 媒體內容分析技術】，則將整合至伺服器端，以提供使用者必要的資訊及更便利的服務。下圖為互動資訊媒體服務系統與各分項計畫之間的關係圖。進一步的研究方法，請參考分項計畫一的說明。



而在"家庭媒體中心(Home Media Center)"中，我們將提供一個多媒體內容分析管理系統，使用者可以很方便地管理經由數位攝影機、數位相機等設備所錄製的影音資料。透過對媒體內容的分析，可將媒體內容加以分類、摘要，並將萃取分析所得之 metadata 以 MPEG-7 所定義之敘述詞以及描述結構儲存起來。藉由這些 metadata 的協助，使用者可以很方便地尋找特定的影像片段，或是瀏覽影片的摘要畫面。針對影音媒體內容的分析與擷取，家庭媒體中心整合了【影音特徵值萃取工具集】、【內容分析與摘要模組】等技術模組，可支援本計畫另一系統"互動資訊媒體服務"，提供播放節目的預覽與摘要功能；其它的應用還包括家庭影片剪輯軟體、個人數位像簿等等。此外，此系統亦可提供分項計畫四中【保證服務品質的媒體內容代理伺服器】所需之相關技術。關於影音媒體內容分析技術的細節，請參考分項計畫二。

除了影音媒體之外，家庭媒體中心還可對三維模型與場景進行搜尋與分類，此功能將針對從事電腦動畫及虛擬實境人員的需求，建立三維資料庫以供查詢。由於 MPEG-7 標準對於三維物件的描述定義並不多，因此，在分項計畫三中的【三維物件搜尋器】模組，將針對三維物件設計所需之描述詞以及描述結構，並配合該分項計畫另一模組【三維 MPEG-4 動畫場景建構工作台】之協助，發展友善的人機互動界面以供查詢。詳細的研究方法，請參考分項計畫三的說明。下圖為家庭媒體中心系統與各分項計畫之間的關係圖。



本計畫在進行的過程中，將參照 MPEG-4/7 兩項國際標準中有關媒體內容呈現及描述方式的定義。接下來，我們將分別就 MPEG-4 以及 MPEG-7 標準中，與本計畫相關的研究步驟進行說明。

【與 MPEG-4 相關之研究方法】

本實驗室在完成前期產學合作計畫「MPEG-4 複合媒體及網路虛擬實境之研發(I)(II)(III)」之後，對於如何將兼具"物件導向(object-oriented)"以及"複合媒體(synthetic and natural hybrid media)"特色之 MPEG-4 標準，運用在媒體內容呈現、可調整性編碼以及傳遞技術等方面上，已經累積了相當的經驗，並有了初步的研發成果。本期產學合作計畫中，將著重在 MPEG-4 互動機制的研究上。這裡所謂的互動，不單只是改變場景中的物件外觀，或是控制媒體的播放行為(暫停、繼續)；更進一步地，我們將研究如何發展伺服器端可程式化(Programmatic)模組，並加強媒體播放端 BIFS(BInary Format for Scene)方面更完整的能力。此外，鑑於以往在建構動畫以及三維場景時，缺乏適當的人機界面，我們將在分項計畫三中，進行【三維 MPEG-4 場景動畫建構工作台】之研究。

其次在目前最受重視之 Video Streaming 技術方面，本實驗室在前期產學計畫中雖已有了初步的成果，但在相關技術日漸掌握與市場契機日漸明確的雙重鼓舞下，我們再次

仔細評估 MPEG-4 標準的技術研發方向及應用價值後，選定 Video streaming 為本次產學計畫的另一研發重點。一方面可承續前期計畫有關 MPEG-4 之研究成果，另一方面也可扮演與 MPEG-7 技術融合的橋樑角色。

【與 MPEG-7 相關之研究方法】

本計畫的另一個研究重點，將針對 MPEG-7 標準所制訂一系列對於媒體內容之分析與描述進行探討。有鑑於 MPEG-7 標準所制訂的描述詞(Descriptor)以及描述結構(Description Scheme)種類相當繁多，為了避免研究的領域過於廣泛，而使整個計畫失去焦點，本計畫將針對影像(Video)資料的分析與摘要進行研究，並將此技術整合在"家庭媒體中心"系統中，提供使用者能快速地瀏覽影片，並藉由影片的摘要資訊來得知影片內容，例如採用主要圖像 (Key Frame) 選取，或由連續視訊建立全景影像(Video Mosaic)等方式。而在"互動資訊媒體服務"系統中，使用者也可藉由伺服器所提供之節目預覽功能，選取喜愛的節目，而與伺服器端進行互動。在進行分析的過程中，我們將透過 MPEG-7 所定義的描述詞以及描述結構，替媒體內容建立其描述實體(Description)。

媒體內容的分析與摘要是個困難而充滿挑戰的議題。在分項計畫二中，我們將採取漸進的方式，先針對 MPEG-7 標準中所定義低階層影音特徵值，推算出其敘述詞實體值，這一部分將由【影音特徵值萃取工具集】負責完成，再由【內容分析與摘要模組】技術分項，從結構性(structure ascept)、知覺性(concept ascept)等方面，對媒體內容進行分析以及摘要，以期建立一符合人類知覺(human perception)、結構化的連續影像物件模型。詳細的研究方法，請參考分項計畫二。

另一方面，除了傳統的影音媒體內容之外，三維模型、三維場景等虛擬合成(Synthetic)的媒體，也是整個媒體內容產業中相當重要的一種素材，然而，MPEG-7 在描述三維模型方面著墨甚少，面對越來越多與豐富的三維模型資料庫，建立適當的管理系統是必要的。分項計畫三的【三維物件搜尋器】將針對三維模型的搜尋、分類進行研究，並將此技術整合至"家庭媒體中心"系統中，提供使用者在資料庫中搜尋特定的三維模型。詳細的研究方法，請參考分項計畫三的說明。

由於 MPEG-7 對媒體內容定義了詳盡的描述結構，豐富的 metadata 資訊也衍生出相當多的應用，其中一項便是媒體代理伺服器的管理。媒體代理伺服器的主要目的，在於縮短初始網路延遲(Network Initial Delay)、減少網路延遲差異(Network Jitter)，並降低伺服器的負擔(Load)。透過伴隨媒體內容本身的 metadata 之描述，媒體代理伺服器可以得知該影片的檔案大小、播放時所需的頻寬、傳輸速率等等資訊，利用這些資訊，可以對頻寬作有效的分配與使用，以期提供保證服務品質(QoS)的傳輸服務。詳細的研究方法，請參考分項計畫四的說明。

（第一年研究成果）

1. <u>互動資訊媒體服務系統(Interactive Information Media Service System)</u>

➢ MPEG-4場景編輯器

圖 1 為分項計畫一所開發完成的 MPEG-4 場景編輯系統，本系統延續前期產學合作計畫其中分項計畫一"MPEG-4 場景描述及瀏覽編撰工具" (NSC-89-2622-E-002-013)的結果，突破並增加以下的技術項目：

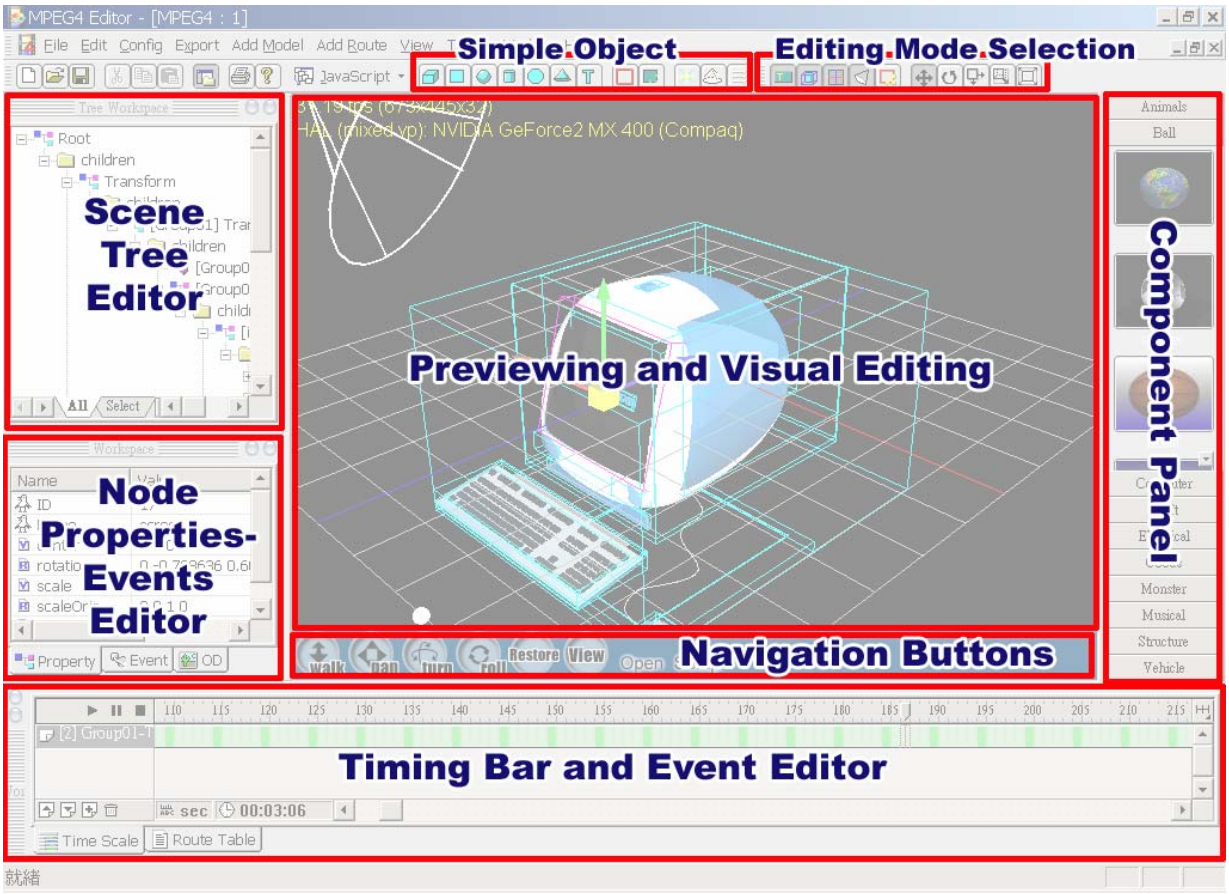


圖 1 MPEG-4 場景編輯器

1. 全新的視覺化編輯方式：新的編輯方式大為改進之前難以使用的缺點，使得使用者編輯場景大為容易。
2. BIFS-Audio的支援：BIFS-Audio讓MPEG-4場景在聲音方面的支援更為齊全
3. BIFS-Animation的支援：BIFS-Animation機制讓MPEG-4場景更富有豐富性
4. JavaScript支援：Script機制提供給MPEG-4場景更多互動的能力

透過上述之技術，目前 MPEG-4 場景編輯器已經能夠編輯出具備高度聲光效果與互動性的場景。圖 2 即為利用此編輯器製作而成的 MPEG-4 互動場景。

圖 2 MPEG-4 互動場景

➢ **MPEG-4媒體內容伺服器雛形**

MPEG-4 媒體內容伺服器利用串流技術傳遞影音資料流，並根據影音內容的複雜度以及頻寬的限制，把資料切割大小一致的網路封包。單位時間內的資料量盡可能維持穩定，在用戶端即時還原影音資料並加以呈現。

利用影音串流服務的機制，用戶端不再需要預先租借光碟或冗長的下載過程，可以即時從網路欣賞到高品質的影音內容，影音內容即時由串流伺服器端切割成連續的封包，透過網路傳遞到用戶端之後，由用戶端程式重組這些封包，還原成高品質的影音內容。關於串流技術之詳細說明，請參閱分項計畫二【MPEG-4/7 媒體內容分析技術與資訊串流模組】。

2.

”家庭媒體中心(Home Media Center)”：

➢ **MPEG-7多媒體搜尋系統**

圖 3 為分項計畫二所開發完成的 MPEG-7 多媒體搜尋系統。此系統結合了【影音特徵值萃取工具集】以及【內容分析與摘要模組】之初步成果，並已完成以 MPEG-7 參考軟體為基礎之多媒體搜索系統架構。

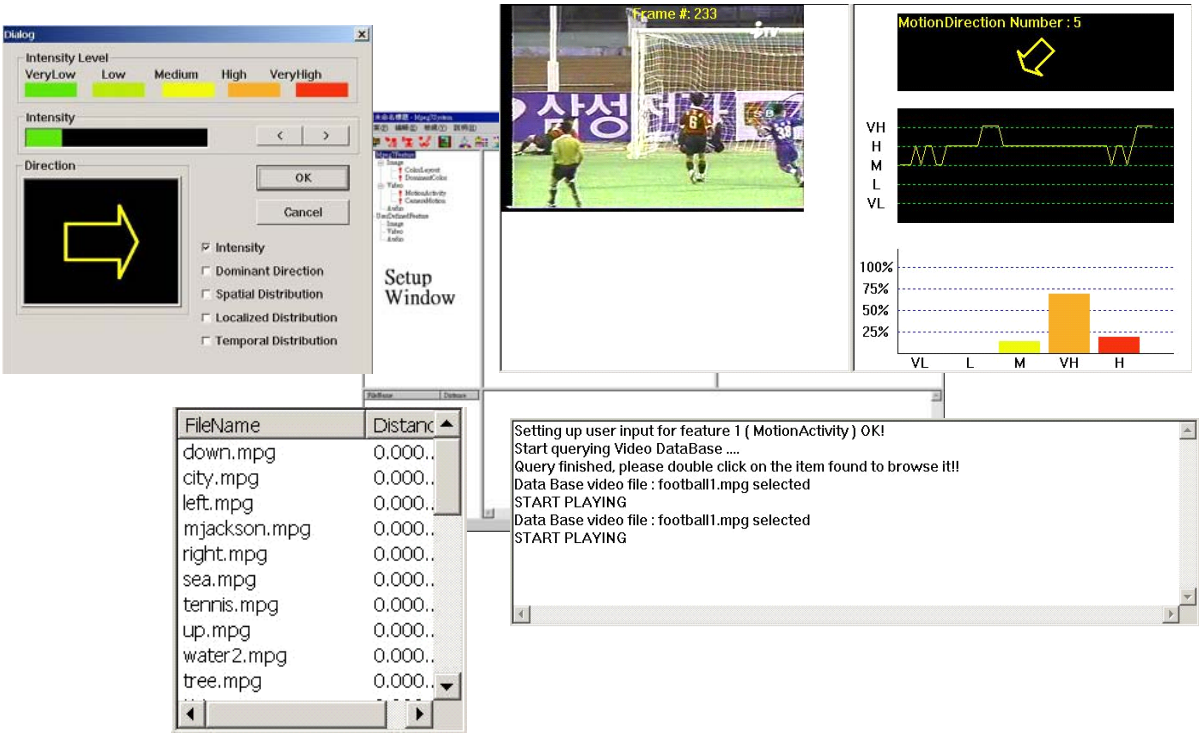


圖 3 MPEG-7 多媒體搜尋系統

目前完成的 MPEG-7 述詞(descriptor)包括了 color layout, dominant color, motion activity, contour shape, region shape 等，並以一個初步的搜尋系統，包含魚類搜尋及商標比對來展示其結果。

另外，在摘要模組之初步成果中,已實作連續視訊建立全景影像(Video Mosaic)及主要影像(Key Frame)之擷取。

➢ **三維物件搜尋器**

本計畫中，我們採用 MPEG-7 中，Multiple View 的觀念，使用正十二面體包住三維模型，運用正十二面體的二十個頂點，作為視點，每一視點上取一張剪影，再使用 MPEG-7 中提供的 Region Shape 的方法，來作為二維剪影的比對方式。

我們的三維模型比對方式是以範例作為搜尋比對關鍵(Search by an Example)。搜尋結果依相似程度由高至低排列。圖 4 為搜尋系統的執行畫面，圖 5 為此系統的搜尋結果。目前資料庫有 445 個三維模型，在 CPU 為 Pentium III 800 MHz 的機器上，搜尋時間約為 11 秒鐘。

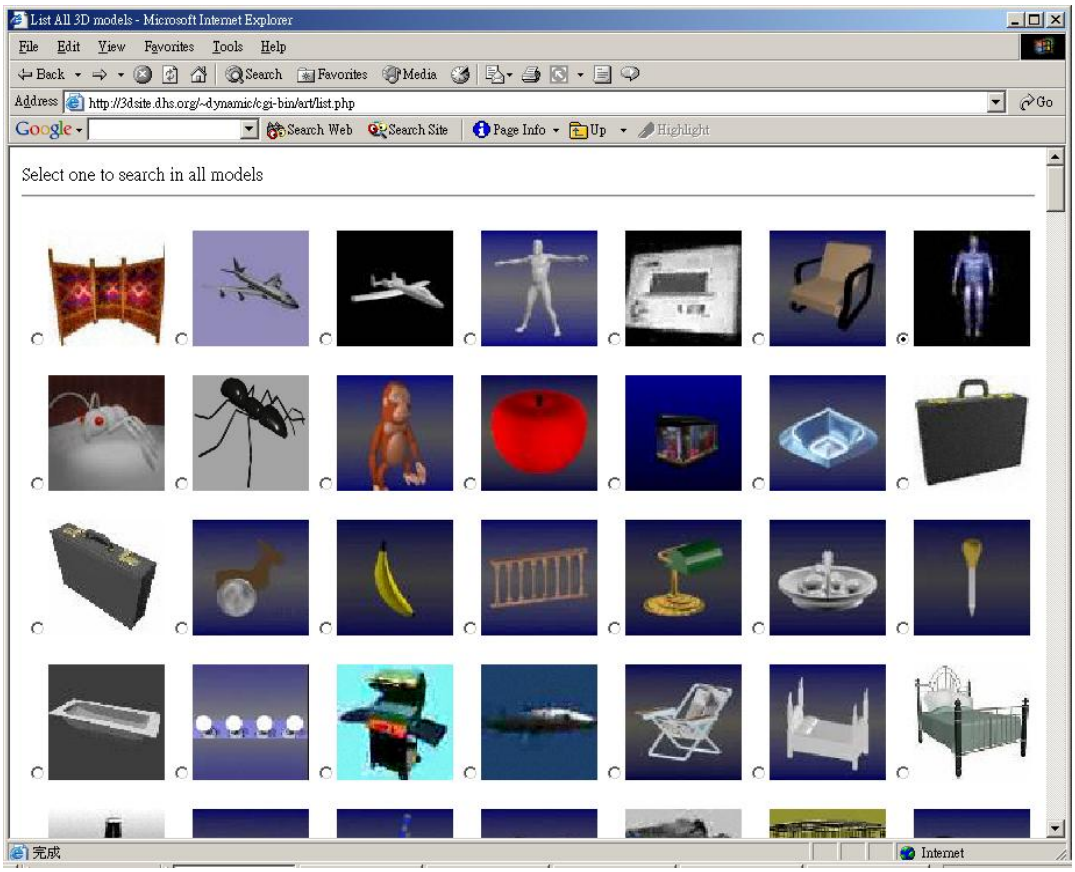


圖 4 三維物件搜尋器之執行畫面

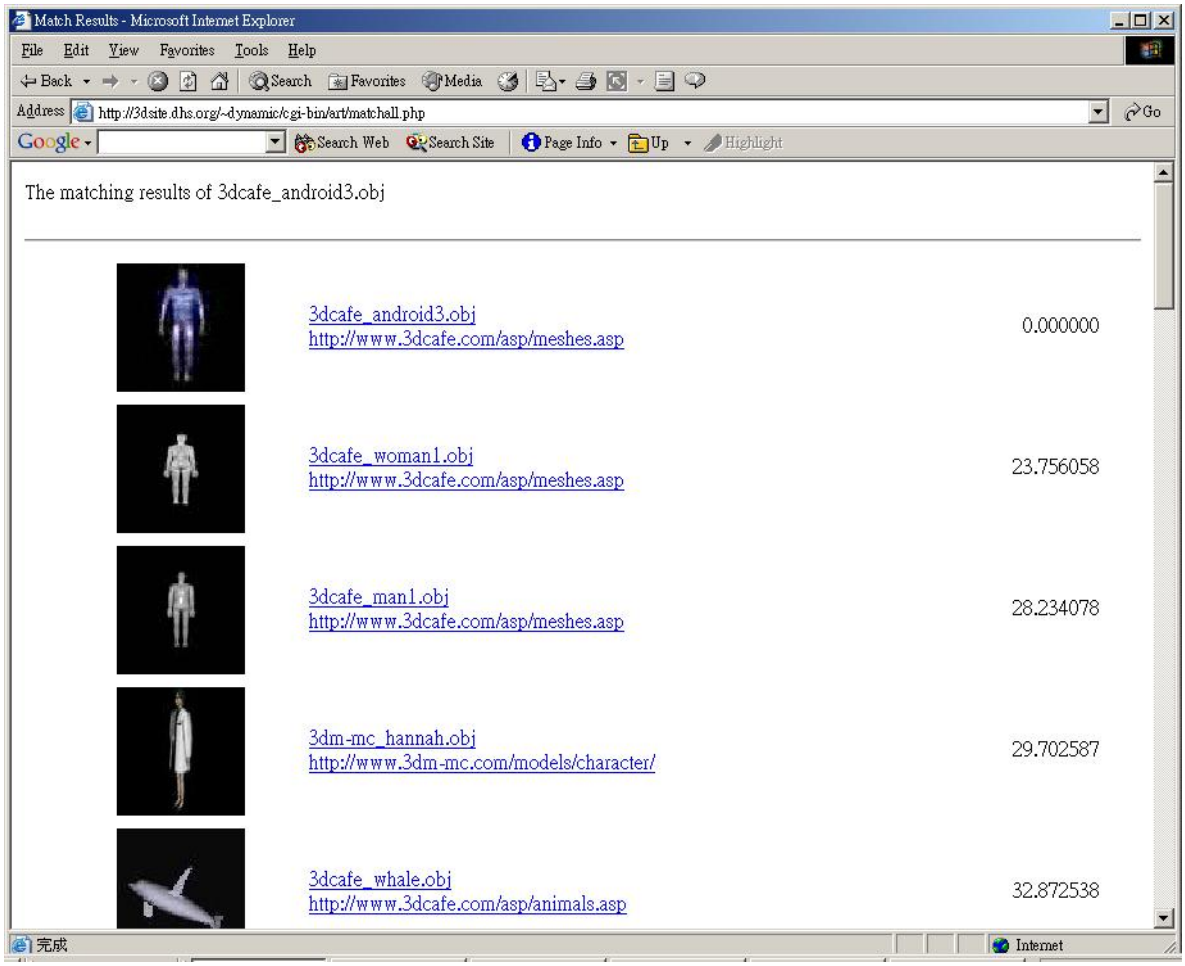


圖 5 三維物件搜尋器之搜尋結果

3.

”Rich-media Platform for Training”：

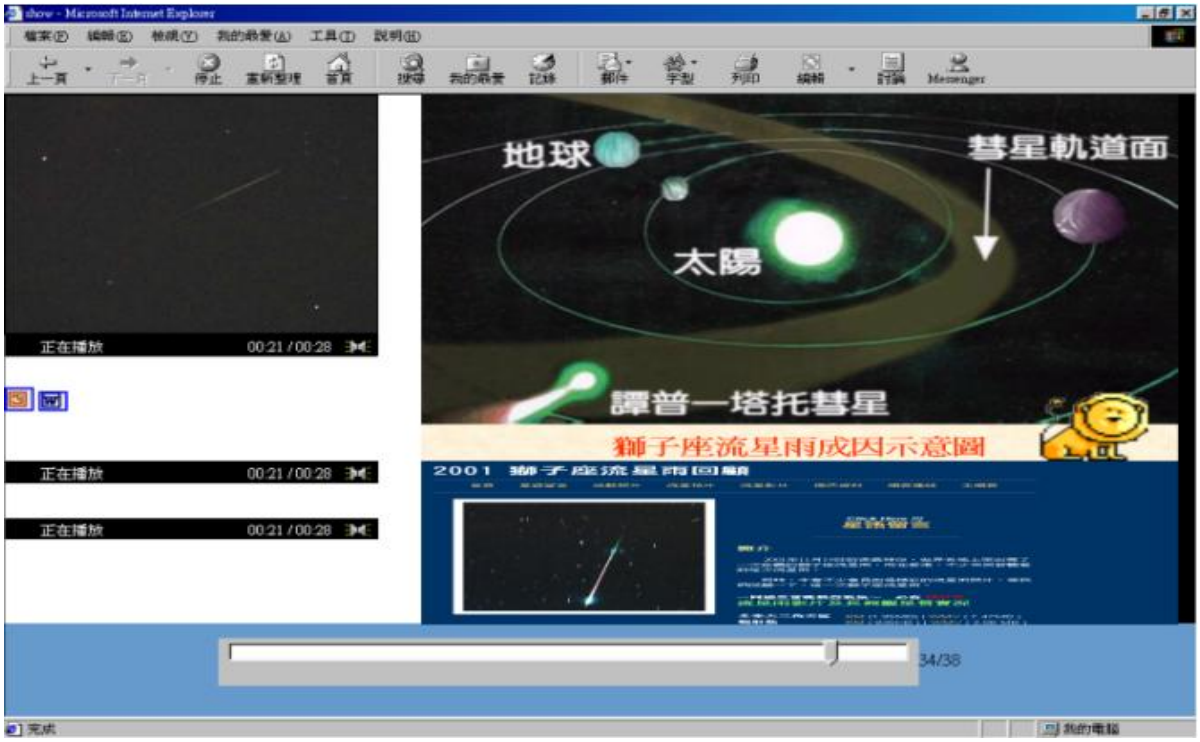


圖 6 Rich-media Presentation 的外觀

我們所提出的系統，是架構在以網頁為基礎(Web-Based)的平台上。如圖 6 為介紹獅子座流星雨的簡報，其中涵蓋各類型的多媒體，包括影片、聲音、投影片、參考性文件…等七個媒體內容。此外，一個完整頁面可任意切割為多個視窗框架(Frame)。以下圖為例，左上角正播放著星空拍攝的影片；其下方的小圖示為參考性文件的連結；接下來是擔任簡報講說的聲音檔，負責講解流星雨的形成過程；以及優美的背景音樂。而右上角為說明流星雨成因的投影片；而右下角為回顧 2001 年獅子座流星雨的網頁。最下面為簡報的主時間軸(Timeline)，用來控制物件之間的同步。

（第二年研究成果）

1. 互動資訊媒體服務系統(Interactive Information Media Service System)相關成果

- **MPEG-4 場景編輯器之改良**

本系統承續了第一年的研究成果，為 MPEG-4 場景編輯器增加了對於 JavaScript 及 mp4 檔案格式的支援.

新完成的場景編輯器提供了方便的使用者介面，可以讓使用者簡單的點選想要處理的事件，系統隨即便會處理好其他的細節，包括新增 Script 物件節點、填好基本的函式宣告、連結事件表等等，並可進入程式撰寫模式。

另外，編輯好的場景，可經由 MP4 檔案格式輸出精靈(MP4 File Format Exporting Wizard)的協助，輸出成.mp4 檔案格式；此外，影音媒體檔案也可使用 MP4 Builder 將其封裝成.mp4 格式，以便日後使用。

- ***互動式*MPEG-4媒體播放器與伺服器**

經由編輯器編輯好的互動場景，輔以場景中所定義的各種不同影音媒體物件，可存放在MPEG-4伺服器，再透過本分項計畫所開發支援JavaScript 的MPEG-4播放器，便可體驗互動式MPEG-4媒體的強大功能與聲光效果。

本計畫於第二年所開發的MPEG-4伺服器，係參考 MPEG-4 Standard Part 8 -- 4onIP 中所提出的資料封裝格式以及網路協定標準(RTSP/RTP)，因此未來能與同樣實作此標準的用戶端相通。此外，除了提供事先儲存的影音媒體串流之外，本伺服器也能透過視訊擷取裝置，傳送即時的影音串流。

- **MPEG-4伺服器端與媒體播放器端的通訊協定**

基於資料特性與目的的不同，伺服器端與媒體播放器端之間的通訊協定可分成三類：影音媒體串流、Multi User World控制訊息以及非串流資料等。圖7 為伺服器端與媒體播放器端的通訊協定示意圖。其中傳輸第一類與第三類資料所需之通訊協定，本計畫係採用工業界普遍使用的RTSP/RTP以及HTTP/TCP標準，並加以實作。此外，為顯示本系統之實用性，本計畫亦整合完成了一套【衛星電視訊號擷取與播放模組】。

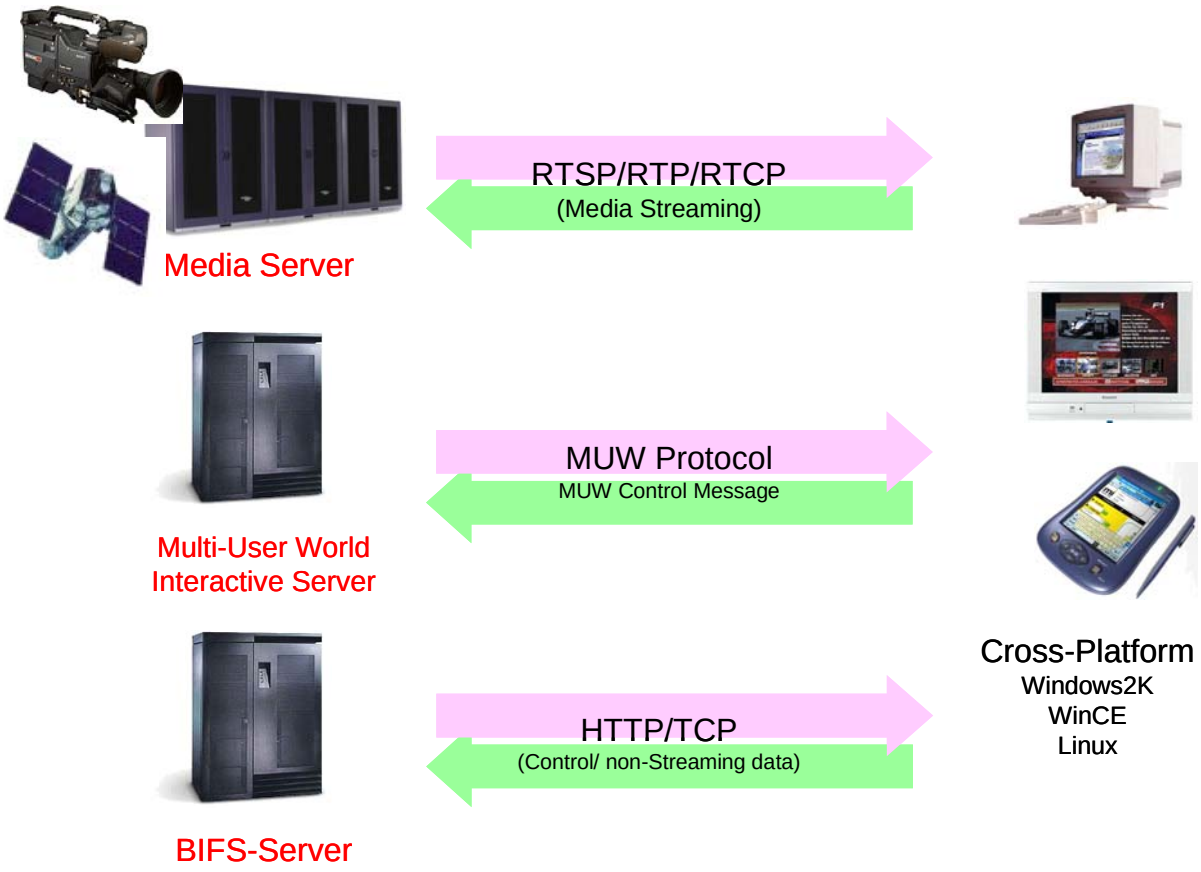


圖7 伺服器端與媒體播放器端之間的通訊協定

此外，本計畫今年所新增之Multi User World部分則是參考MPEG-4 Standard中有關MUW的設計，並已實作出一個Multi User World的雛形系統。（以上成果之詳細內容請參考分項計畫一之成果報告。）圖8 即為本計畫所開發出來的MPEG-4 Multi User World系統操作畫面。

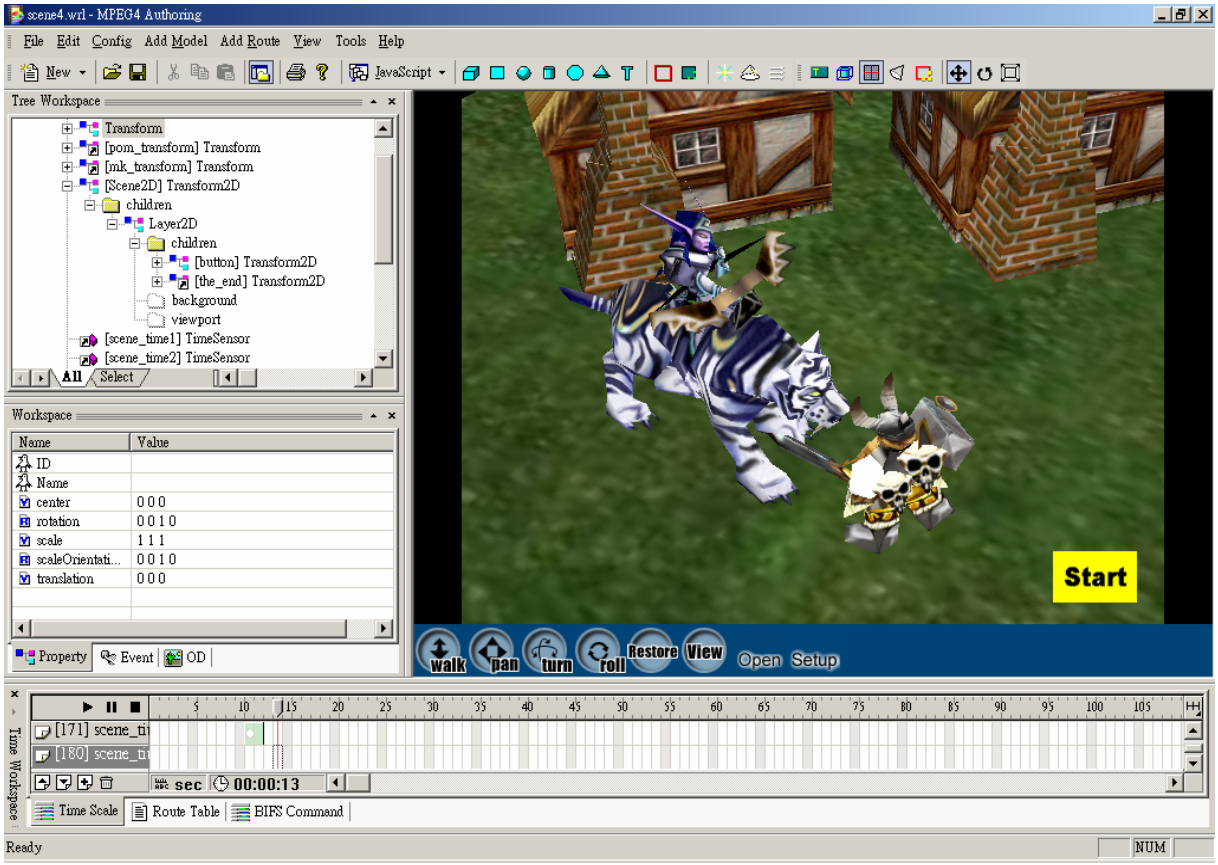


圖8 MPEG-4 Multi-User World 系統之操作畫面

➢ 視訊代理伺服器

代理伺服器(Proxy Server)的基本功能即是作為作為來源伺服器與客戶端的中繼站，將客戶端要求的多媒體資料部分暫存在離客戶端較近的代理伺服器，用來減少網路資料的流量、縮短初始網路延遲、網路延遲差異，並藉由頻寬管理機制的允入控制，對頻寬進行有效的分配及使用。

圖 9 為新一代代理伺服器的架構圖。假設網路多媒體應用程式的客戶端有指定代理伺服器，則所有對原伺服器所發出的要求，將會先傳送給代理伺服器，再查詢代理伺服器所暫存的相關資料，對原客戶端的要求做適當調正，然後才真正對原伺服器發出要求，並回應客戶端。對於任何一個要求，代理伺服器都會分析系統現況，如果可用網路頻寬、系統資源均滿足要求則允入，否則將拒絕要求。之後，代理伺服器將先到片頭暫存表檢查客戶端所要求的資料是否有暫存在片頭緩衝區內，若已存在，則到媒體描述表內，取出此媒體內容的媒體資訊及要求的服務品質，然後建立一條適當的連線，直接將片頭資料由代理伺服器傳向客戶端，並向原伺服器要求後續的媒體資料，再將由原伺服器所接收到的資料儲存在平滑暫存區以確保資料供給的穩定性。反之，若無法在片頭暫存表找到客戶端所要求的資料，則代理程式直接向原伺服器要求客戶端所需要的資料，並將片頭部分儲存在片頭緩衝區，更新片頭暫存表及媒體描述表，且在接受資料的同時，將所接受到的資料傳送給客戶端。至於資料的暫存及取代則由暫存決策器依特定的演算法來決定。本計畫今年已經完成的項目有媒體描述表(MD Table)與片頭暫存(Prefix Table)這兩個主要的模組。(詳細內容請參考分項計畫四之成果報告。)

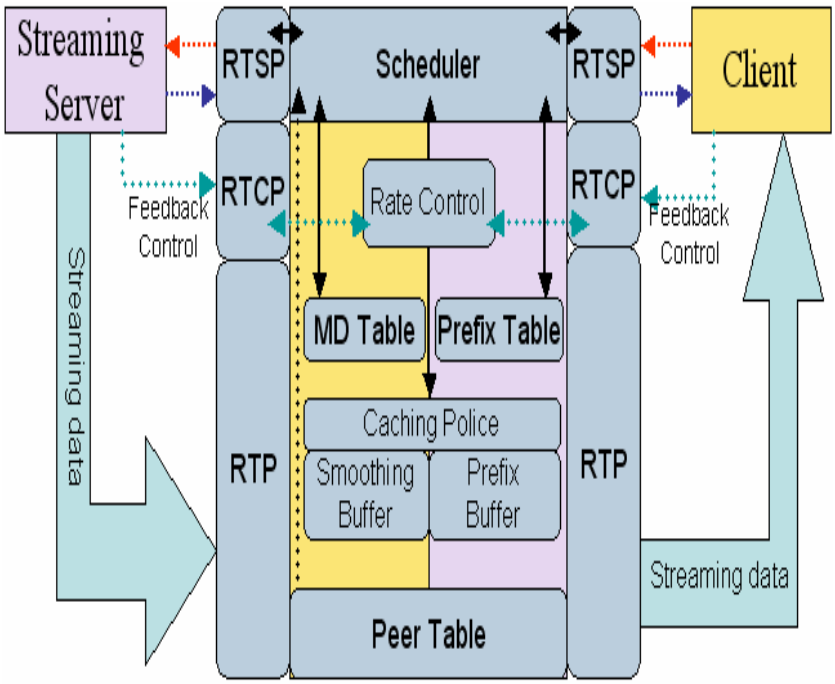


圖 9 代理伺服器的架構圖

➢ **可調整串流服務**

針對 MPEG-4 的可調整特性(Scalability),我們發展出了具備可調整功能之編解碼器，不但透過編解碼參數的調整，讓各種壓縮技巧有更大的彈性，並能接受使用者直接互動式的線上參數調整及與 QoS 配套之可調整性壓縮模式。此編碼解碼器提供了五種可調節性，此五種可調節能力分別為: 畫面呈現品質的可調整性，空間解析度上的可調整性，時間解析度上的可調整性，壓縮位元率的可調整性以及計算能力需求的可調整性。圖 10 為可調整性影像壓縮編碼及解碼器的影像品質調整架構示意圖（詳細內容請參考分項計畫二之成果報告）

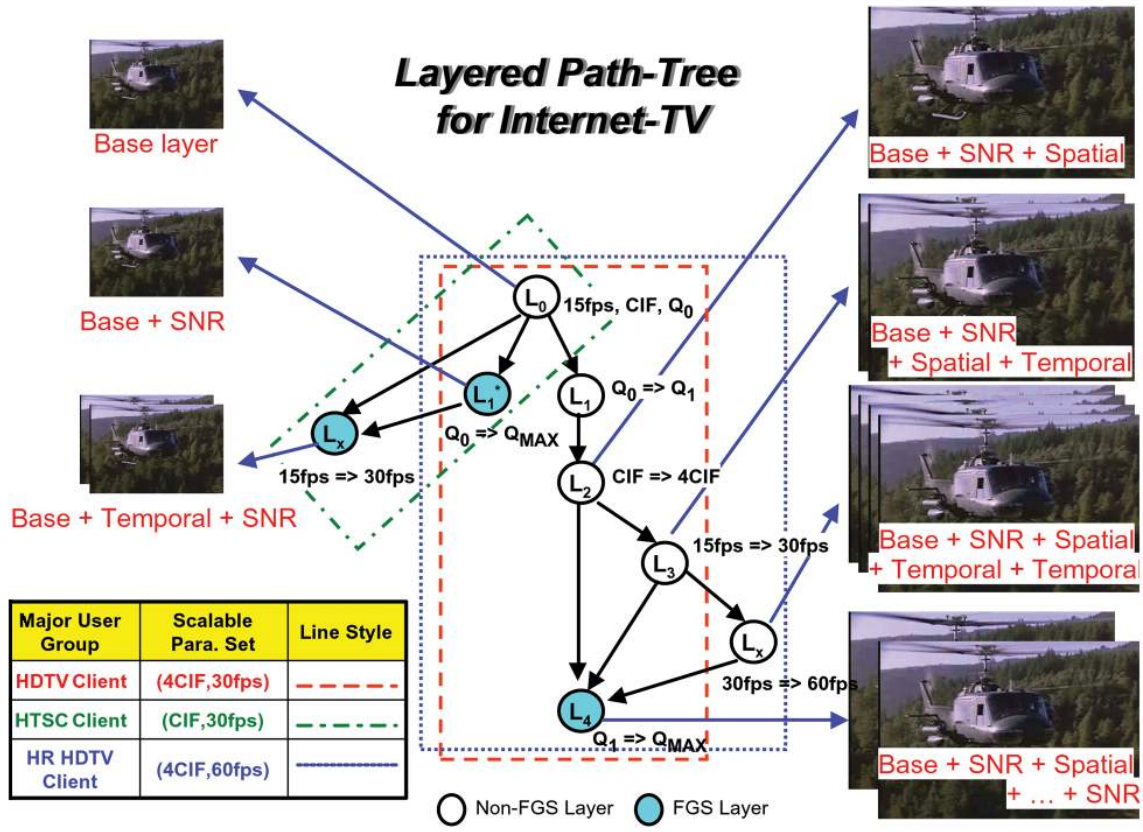


圖 10 可調整性影像壓縮編碼/解碼器系統示意圖

2.

家庭媒體中心(Home Media Center)相關成果

➢ 具備影像語意功能之數位照片管理系統

數位照片為目前家庭內最常產出之數位內容型態。本計畫今年完成的數位照片管理系統，為家庭媒體中心內針對數位照片管理所設計的模組，除了具備以完整之【MPEG-7 影音特徵值粹取工具集】為基礎之影像索引與查詢功能，另外還增加了如人類臉部偵測與辨識，視覺化通訊錄，風景/人類照片分類，智慧型縮圖影像顯示等具備語意概念（semantic meaning）的進階數位照片管理功能。圖 11 為我們所完成的具語意功能之數位照片管理系統之使用者介面。

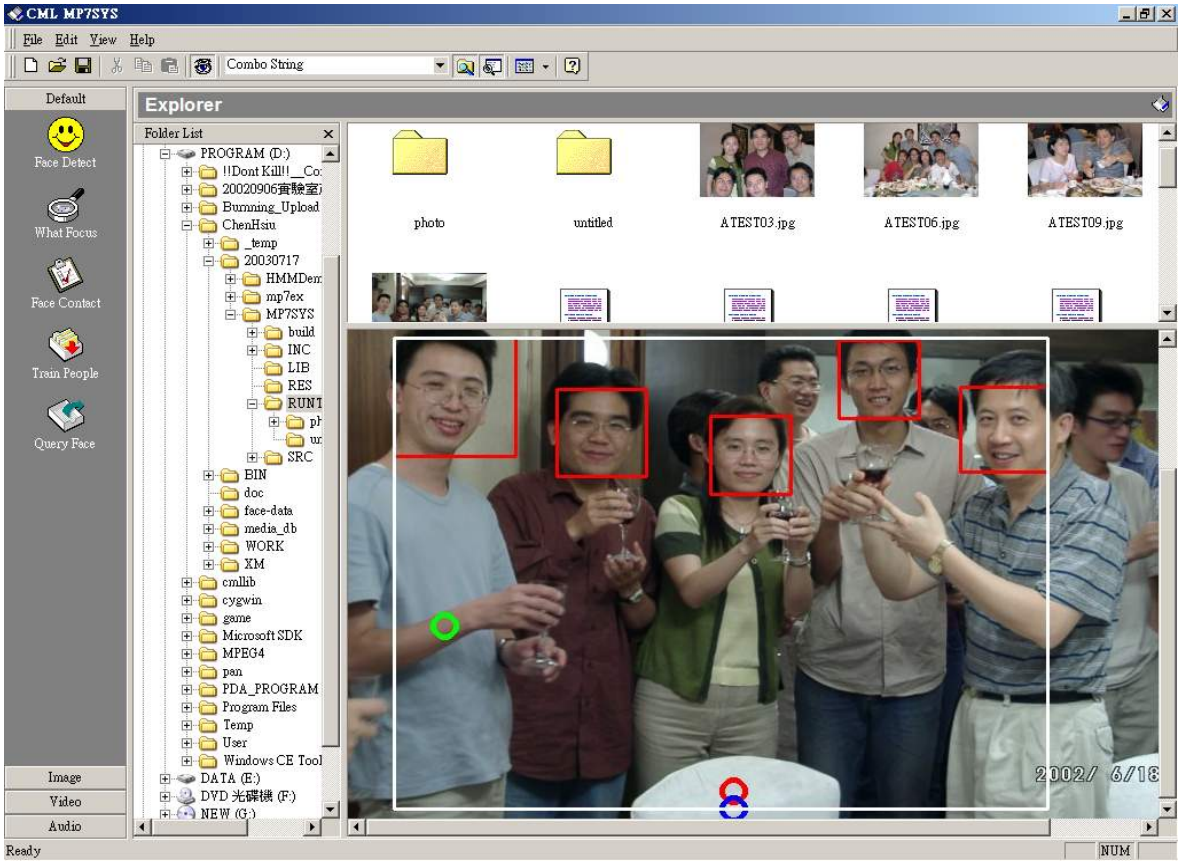


圖 11 數位照片管理系統之使用者介面

- 電視新聞分析與摘要模組

本年度的另一項成果為針對新聞影音資料之有效摘要及瀏覽之工具。此工具係利用由上而下的方式，以對於“重要”語意的了解來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，這個技術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, 分別給予重要性標示。藉由對攝影師/播報員/後製人員/新聞現場人物…等四類人物的反應及行為追蹤, 我們歸納出某些事件的發生具有非常強烈的重要性意義。而經由這個以重要性標示為基礎的架構，相當程度地反映出人類對於新聞內容重要程度的看法。

- 使用者注意點分析與偵測系統

另外, 我們還提出了一套可偵測影像及視訊資料中使用者的注意點的系統。藉由此系統偵測出的注意點，影像/視訊編碼器在壓縮時可以更符合使用者的需求。此系統藉由影像畫面中的亮度、色彩、人類膚色以及物體運動方向的分佈情形，預測出人眼於此影像中的注意點。(以上成果之詳細內容請參考分項計畫二之成果報告)

➢ 三維物件搜尋器之改良

本研究成果為第一年成果之【三維物件搜尋器】的功能增進版本，不論在搜尋方式、速度以及正確率等方面皆有相當幅度之進展。

在搜尋方式方面，二維的繪圖搜尋介面增加了更多的繪圖工具，使搜尋更為簡便；以關鍵字搜尋的技術亦被同時引進於本系統中，使用者只需輸入標的物件的名稱或相關詞彙，便能達到快速搜尋的目的。不論由繪製二維圖形或輸入關鍵字詞，其搜尋結果皆能以三維模型相互比對的方式做更進一步的準確搜尋。圖 12 即為由使用者繪製出之二維圖形搜尋出之結果。

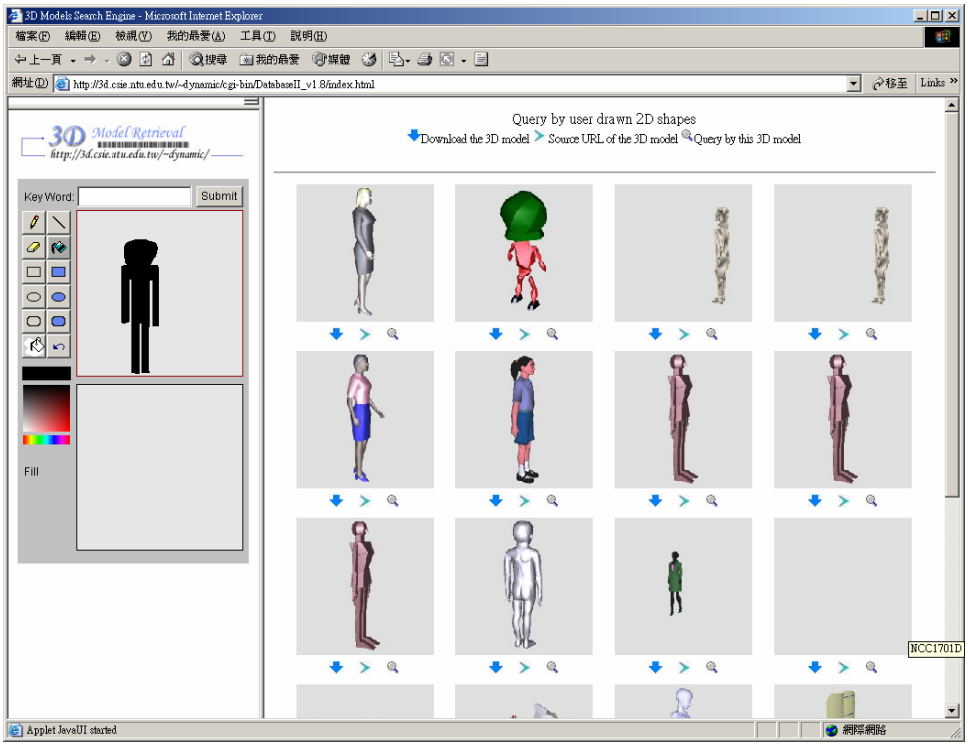


圖12 在 3D 物件搜尋器中根據使用者繪製之二維圖形進行搜尋之結果

在以三維物件相互比對的部分，由於比對過程分為六個階段、並於各個過程逐次作淘汰的動作，另外用以描述三維物件的述詞皆以少量的資料量表示。因此，由全世界網路上不重複的一萬多個三維模型中找出相似模型的時間由原來的一分鐘大幅減少為兩秒鐘，而由二維圖形搜尋相似三維物件的過程也降為0.1 秒

另外，本年度在描述三維物件所使用的述詞有較去年有所調整，即Fourier descriptor, Circularity 以及Eccentricity 的加入。此三者與原有的 Zernike moment 結合後，就搜尋的正確率有顯著的正面效益。本系統與MPEG-7 的Multiple view descriptor、Shape descriptor 以及由美國 Princeton 大學以3D spherical harmonics 所建立之搜尋系統的比較。整體而言，本系統的搜尋正確率較此三者分別高出22%、95%、與 43%，是為目前已發表的搜尋方式當中效果最佳者。[17]

- 三維 **MPEG-4** 動畫場景建構工作台

以第一年的的三維模型雕塑系統為基礎，第二年在三維MPEG-4 動畫場景建構工作台方面，三維模型雕塑系統多階層的空間分割已發展完成。除了提供原有的 Extended Marching Cube method (EMC)來保留雕塑物品的特徵邊緣之外，還引進了多重解析度的概念。利用雕刻面的高斯曲率(Gaussian curvature)來作為多接層的空間分割的條件。當雕刻工具的曲率值越大時，就以越精細的解析度來表示。此概念使得本系統可以做出比單一解析度系統更加精細的模型，這樣的改進使得此系統得以更接近實用階段。

另外，還套用了更為精確的力回饋力量計算模型，稱為正切平面力模型(Tangent-plane force model)。這套系統令使用者能以更直覺而方便的方式來建立動畫場景中的物件。使用者可以利用系統中各種不同的形狀的雕刻刀，如球形，立方體等工具，來進行三維模型的雕刻，並在使用時感受到更加精確的碰撞力量。（以上成果之詳細內容請參考分項計畫三之成果報告）

（第三年研究成果）

1. 互動資訊媒體服務系統(Interactive Information Media Service System)相關成果

➢ MPEG-4 多使用者應用開發架構

隨著寬頻與無線網路基礎建設的高度發展，電腦運算速度、顯示設備與儲存媒介等硬體技術的一日千里，人們透過各類普及運算設備(Pervasive Devices)取得多媒體資訊，並與其他使用者之間互傳訊息、共享資訊甚至進行虛擬會議等應用已經成為趨勢。而以往電腦輔助群組協同工作(Computer Supported Cooperative Work, CSCW)已廣泛地被運用在遠距教學、軍事模擬以及多人連線遊戲等等不同的領域。然而這些系統多半採用侷限於特定應用的封閉標準(例如美國陸軍的模擬系統 SIMNET)；或是應用現有技術以符合特定需求(例如以 Web 技術發展而成的聊天室)，前者缺乏擴充性與互通性，後者則因為遷就現有架構，並不適用於新興的媒體格式或網路技術。本研究提出的 MPEG-4 多使用者應用開發架構，可以簡化開發各類不同的多使用者互動系統的過程。

本系統利用 MPEG-4 標準所定義的工具並參考其所提出的應用程式引擎(MPEG-J)以及多人世界(Multi-User World)標準，設計出一個 MPEG-4 應用開發架構，以縮短多使用者互動的應用程式開發時程。這個架構包括了四項核心模組，分別是：(1)MPEG-4 表現引擎，(2)媒體存取層，(3)MPEG-4 應用程式引擎以及(4)多人應用程式伺服器。MPEG-4 表現引擎是一個跨平台的互動媒體呈現架構，配合用戶以及伺服端的媒體存取層，能滿足多點傳輸媒體資料流的需求。而 MPEG-4 應用程式引擎提供了一組應用程式介面，外部程式可藉此來存取並操控 MPEG-4 表現引擎。多人應用程式伺服器則提供了可程式化的伺服端架構，配合與用戶端之間的遠端函式呼叫實現多人連線下的同步機制。

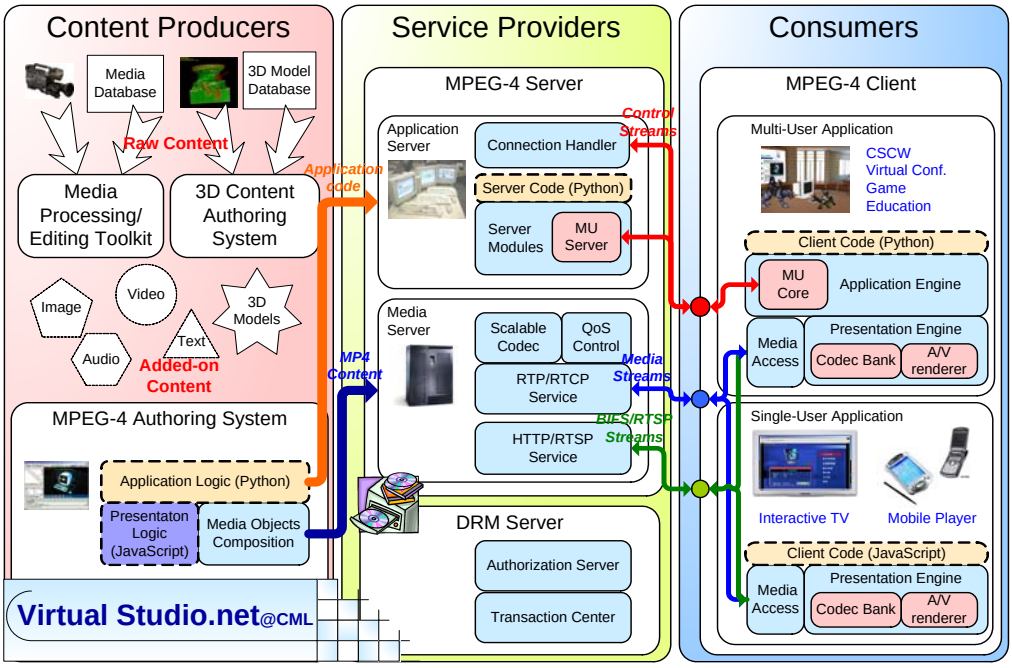
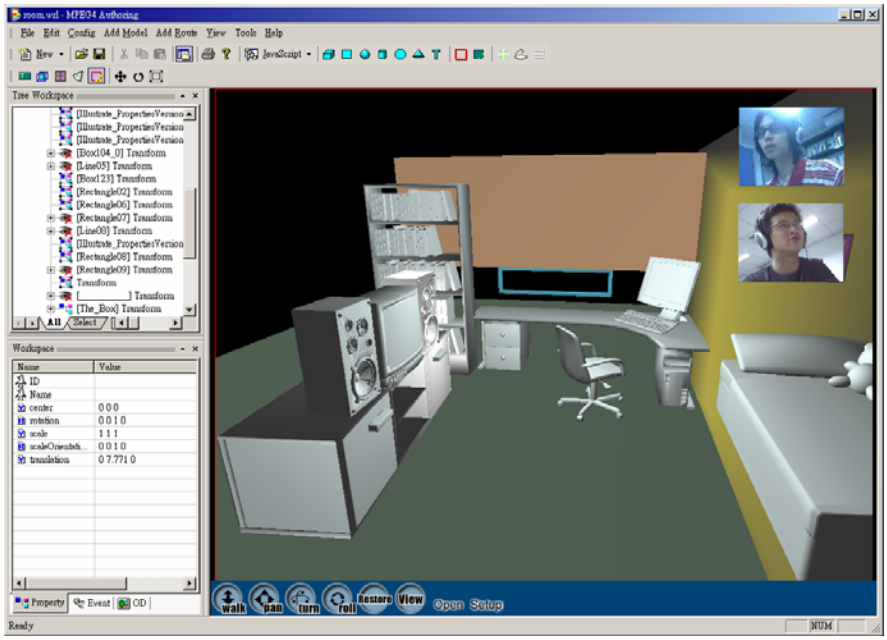


圖 13　MPEG-4 多使用者應用開發架構

本系統完成的這套架構，使得程式開發者能快速開發出不同類型的多媒體群體合作系統。為了驗證本架構之適切性，我們開發了「線上室內設計」、「具有三維共享物件之虛擬會議」以及「多人連線遊戲」等多人連線系統。由開發者的經驗可知，本架構確實降低了這些多媒體群體合作系統的開發難度與成本。

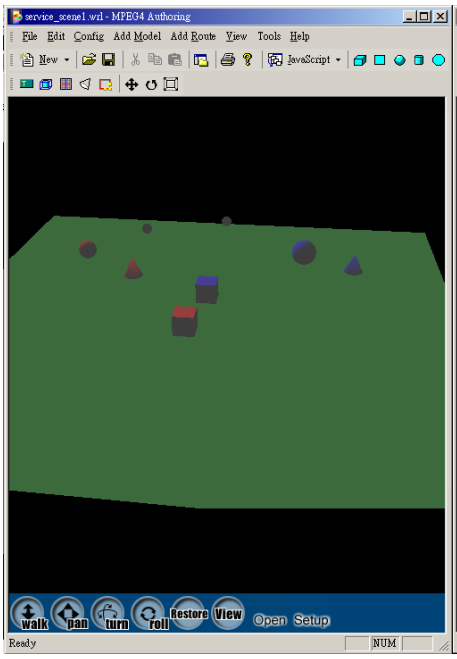


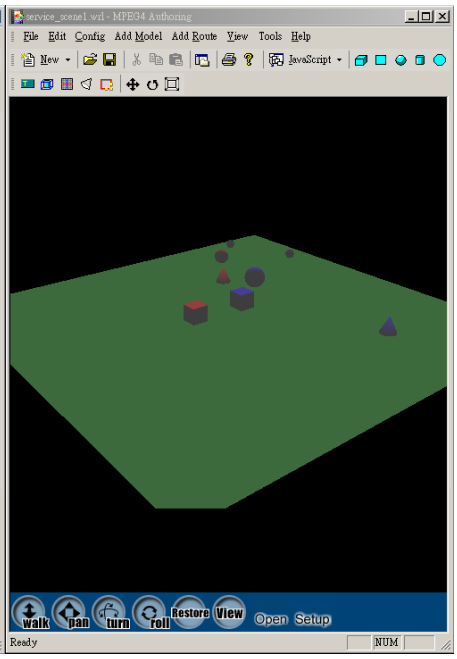
線上室內設計

透過視覺化的 3D 虛擬場景，加入設計會議的使用者可以任意的移動房間中的擺設，或替房間增添家具和擺飾。其他使用者所看到的房間狀態會即時同步改變，達成最直接的意見交流。右方則有視訊會議視窗讓使用者面對面的溝通。

虛擬會議室

使用者可以建立或是加入不同的虛擬會議室，每個會議室都可根據不同主題有著不同的功能，例如虛擬白板、簡報系統、共享的 3D 物件等等，使用者之間可以透過視訊設備進行交談，也可以製作 Avatar 來代表虛擬世界中的人物。





多人連線遊戲

The Cube，兩人即時戰略遊戲。使用者可以使用滑鼠或鍵盤，移動自己的 Cube 或攻擊對手的 Cube。此遊戲展現出 MPEG-4 多人應用的即時性與同步性。

➢ 廣告素材切割系統

對於廣告製作公司來說，倘若能夠再次的利用之前已經拍攝好的廣告來對客戶說明新的廣告提案，是非常有利而節省成本的一件事情。然而，如何找出合適的影片，需要相當好的記憶力之外，其實也是相當不容易的一件事。即使是有經驗的工作者，也可能必須花上許多時間。如果能建立適當的資料庫，讓電腦來幫忙尋找這些片段，相信一定能夠大幅地節省時間及人力。將廣告公司片庫中相連的廣告素材影片切割為不同則的廣告，是在建立資料庫之前的一項必要研究。

拍攝電視廣告的公司都會把拍攝完成的廣告儲存在數位錄影帶(digital beta cam)之中，但是這些錄影帶通常並沒有以良好的格式來規劃，而是一捲捲的收在片庫裡，等要用的時候才找出來。在此，我們藉由分析影片中某些特定的順序，來切割出一則一則的廣告。

經過觀察，影片中每一則廣告裡會出現的鏡頭大致如下圖所示: 首先是會只顯示色條的鏡頭(color bar shot)，緊接著會有一個鏡頭顯示這則廣告的標題，長度，版本之類的資訊(字幕鏡頭，caption shot)，再來會是倒數，接著是一個大約一秒鐘的全黑的鏡頭(black shot)，之後是這一則廣告的影片，最後會再接一個黑的鏡頭。 本研究即利用這個規則將大致符合的廣告切割出來。圖 14 即為電視廣告素材的格式範例之一。

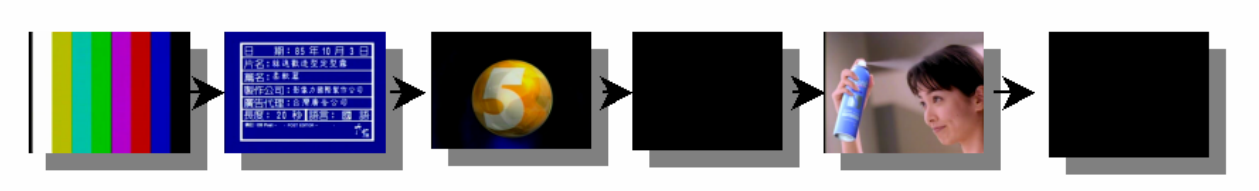


圖 14 電視廣告素材的格式範例

- Linux 上的 MPEG-4 互動場景播放器

Linux作業系統是近年來在各個領域相當熱門的主題，它對軟體開發提供非常高的自由度，開發工具的健全及作業系統的高效能與穩定度，使它不論在即時系統應用、多媒體開發平台、嵌入式平台至手持式行動裝置等都成為一個深具潛力的明日之星。因此，結合Linux的平台優勢，本實驗開發之 Linux MPEG-4 互動式場景播放器在可預見的將來，可提供廣泛且強力的多媒體內容呈現，亦為邁入 Multimedia Everywhere的第一步。

本實驗室所開發之 Linux MPEG-4 互動式場景播放器具有以下特點：

1． 支援 ISO 媒體檔案格式 (MP4)

可以播放符合 MPEG-4 標準的多媒體檔案、串流內容服務，意即符合 MPEG-4 BIFS 標準的多媒體場景。本實驗室已開發有 MPEG-4 Scene Editor 致力於產生 MPEG-4 的互動式場景，其產生之檔案即為 MP4 格式。透過此 MPEG-4 播放器，將可在 Linux 相容平台上播放豐富互動的多媒體內容。

2．實作 MPEG-4 BIFS 及 Scene Graph

BIFS 是 MPEG-4 用來描述場景中影音物件如何組織構成整個場景，它階層式地將物件編碼成二進位的共通格式並支援百餘種不同的物件。本實驗室開發之 Linux MPEG-4 播放器採用 BIFS 為核心場景描述語言，並依照 MPEG-4 標準實作各種不同的 BIFS 節點。

BIFS 節點可包含不同的物件，包括聲音(MP3 或 WAVE 檔)、影片、文字、3D 物件等。

3． 可 Render 2D/3D 合成的景場

近來多媒體內容已不再滿足於單純 2D 的畫面，3D 的設計提供更多可能的感觀效果。而 MPEG-4 本身就支援自然與合成的多媒體物件，因此本實驗室開發之 Linux MPEG-4 播放器可以在 Linux 平台下播放 2D/3D 混合的場景，讓 Linux 平台上的媒體呈現更豐富。

4． 支援 JavaScript 程式

Script 機制提供 MPEG-4 場景更多更強的互動能力。本實驗室所開發的 MPEG-4 JavaScript 引擎(見圖 15)可以製作具有互動性的 MPEG-4 多媒體場景。在此 Linux MPEG-4 播放器中我們亦加上 JavaScript 模組，讓我們可以在 Linux 上處理 script 對場景的動態改變。

此 Linux MPEG-4 播放器不僅能呈現完整的 MPEG-4 互動場景，並能根據使用者點選場景中的物件所引發的事件，正確地執行相對應的 JavaScript 程式碼。進而可以設計各式不同的多媒體互動場景，像是數位電視購物、2D/3D 遊戲、遠距教學等等...

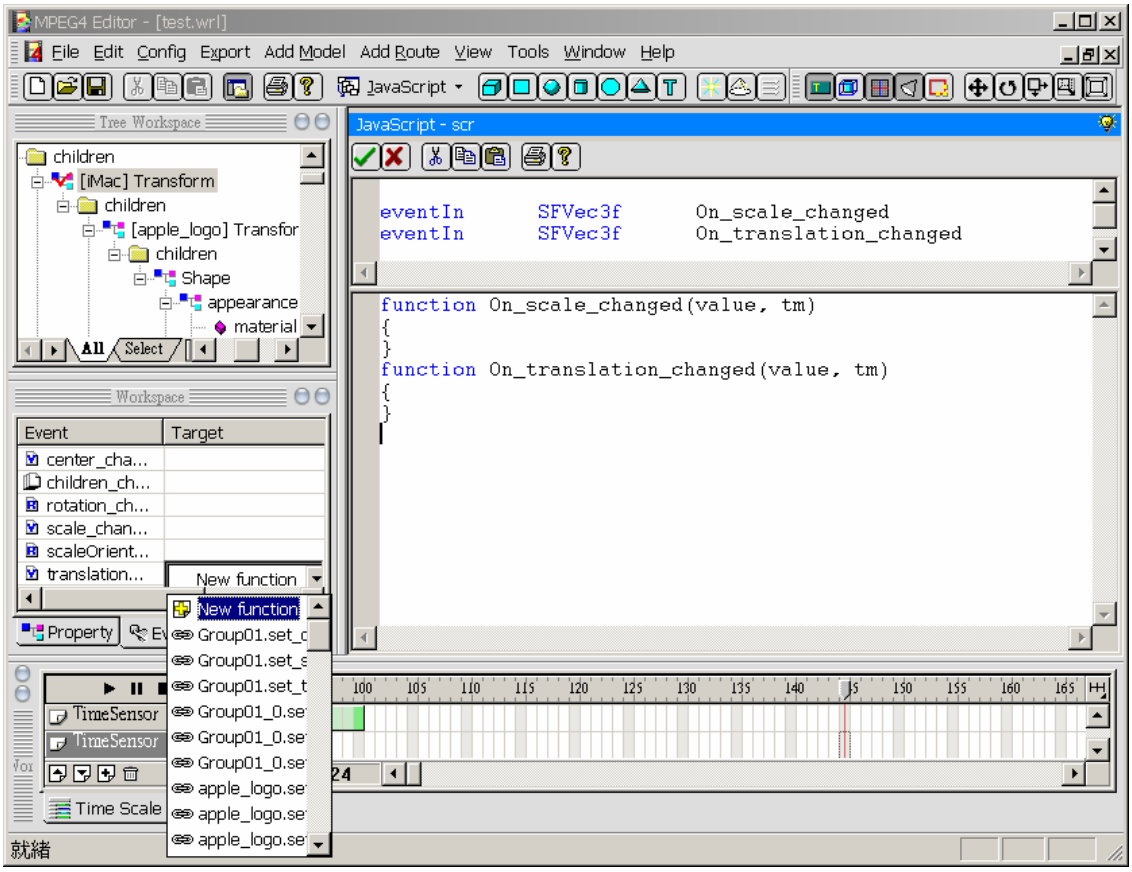


圖 15

在Linux MPEG-4互動式場景播放器的實作上，我們採用開放並跨平台之軟體函式庫(SDL、OpenGL)，因此，此MPEG-4播放器提供了開發MPEG-4播放器之軟體範本，吾人可以用低成本的方式開發出在其它平台上之MPEG-4播放器。

- 智慧型頻寬調節傳輸模組的改良

在本計畫的前兩年中，我們已經完成了【智慧型頻寬調節傳輸模組】的初步架構，本年度計畫針對其上的功能作更進一步的改良與加強。我們在速率調節（Rate shaping）模組中加強了SDF（Selective Frame Discard）功能。

RTP 提供了傳送具有即時性資料的功能，但是它並沒有想像中的去做資源預約(resource reservation)或保證服務品質(QoS) 的功能，它的運作方式就是在要送出的資料前加上 RTP 的封包標頭 (packet header)，如圖九，利用這些標頭它可以提供許多有用的資訊，例如： payload type identification、sequence numbering、timestamping 和 delivery monitoring，再加上 RTCP 的回饋 (feedback) 功能，我們就可以更清楚的了解網路狀況，以達到即時性傳送的要求。

在經由資源有限的網路傳輸影片時，最原始的方法就是不理會網路資源的限制，傳送每一個訊框(Frame)。但在網路資源不足時，封包遺失是在所難免的。然而每個訊框對於使用者觀看的影片服務品質的重要性不同，若是為了傳送重要性較低的訊框而造成重要的訊框被丟棄，不啻為一種網路資源的浪費。此外，使用者端的播放程式也可能因為訊框到達的時間太晚，來不及在該訊框的播放時間播出，而將之丟棄。這也造成網路資源的浪費。

為了改善上述問題，選擇性丟棄訊框(Selective Frame Discard)的概念被提出，簡稱為SFD。其基本想法是由串流伺服器考量網路資源、使用者端緩衝區大小的限制、影片內容特性及使用者對服務品質的要求，選擇性地丟棄對於服務品質較不重要的訊框。

2. 家庭媒體中心(Home Media Center)相關成果

- 電視新聞節目中的廣告偵測

時至今日，數位錄影的裝置已經變的方便又便宜，這件事情促進了數位電視上和一般電視上的時間平移錄影(time-shifted recording)這些應用的出現。對於觀眾來說，時間平移錄影最常被用在除去不必要的資訊上面，例如廣告。精確的廣告切割可以節省錄影的空間，同時也會讓人在觀看錄影的時候更加的舒適。無疑的，廣告偵測是在能完全的把廣告切掉之前很重要的第一步。

相反的，對於廣告商來說，監督電視台是否有按照合約上記載的時間來播送廣告是很重要的一件事情。另一方面，在對電視新聞作內容分析的時候，廣告通常得利用手動的方式來切除，再對剩下的新聞影片來做分析。所以偵測電視新聞當中的廣告，不管是偵測出來之後要把廣告留下來還是把廣告去掉，是一項非常有意義的研究。

由於廣告拍攝手法和內容的多樣性，讓自動偵測電視節目中的廣告一直是一個困難的問題。本論文在研究了有關於廣告偵測文獻以後，加上一些製作廣告的規則以及對於廣告的觀察，提出了一個由上而下的方法來偵測電視節目當中的廣告。

因為廣告通常是影片中觀眾比較不感興趣的部分，所以導演會利用剪接等影片編輯的手法，或是在廣告中特別強調某些顏色，來讓廣告影片變得有趣，並且達到吸引觀眾和強調產品的效果。本研究利用找出影片中這些現象的出現頻率，當作是廣告出現的依據。除此之外，為了使偵測到的廣告的片段更加的精確，本系統在廣告偵測系統之中接著加入了影片場景切換的偵測（video scene boundary detection）試著來自動的找出新聞和廣告之間準確的切換點。實驗中並顯示，本系統可以良好的把新聞中的廣告區域偵測出來。倘若使用者想要再次對電腦自動選出來的廣告區域加以調整，此系統還能夠列出其他可能的切換點來讓使用者選擇。這使得實做出來的系統變成一個非常便利的廣告切割輔助工具。圖 16　即為此新聞節目廣告偵測系統之介面。

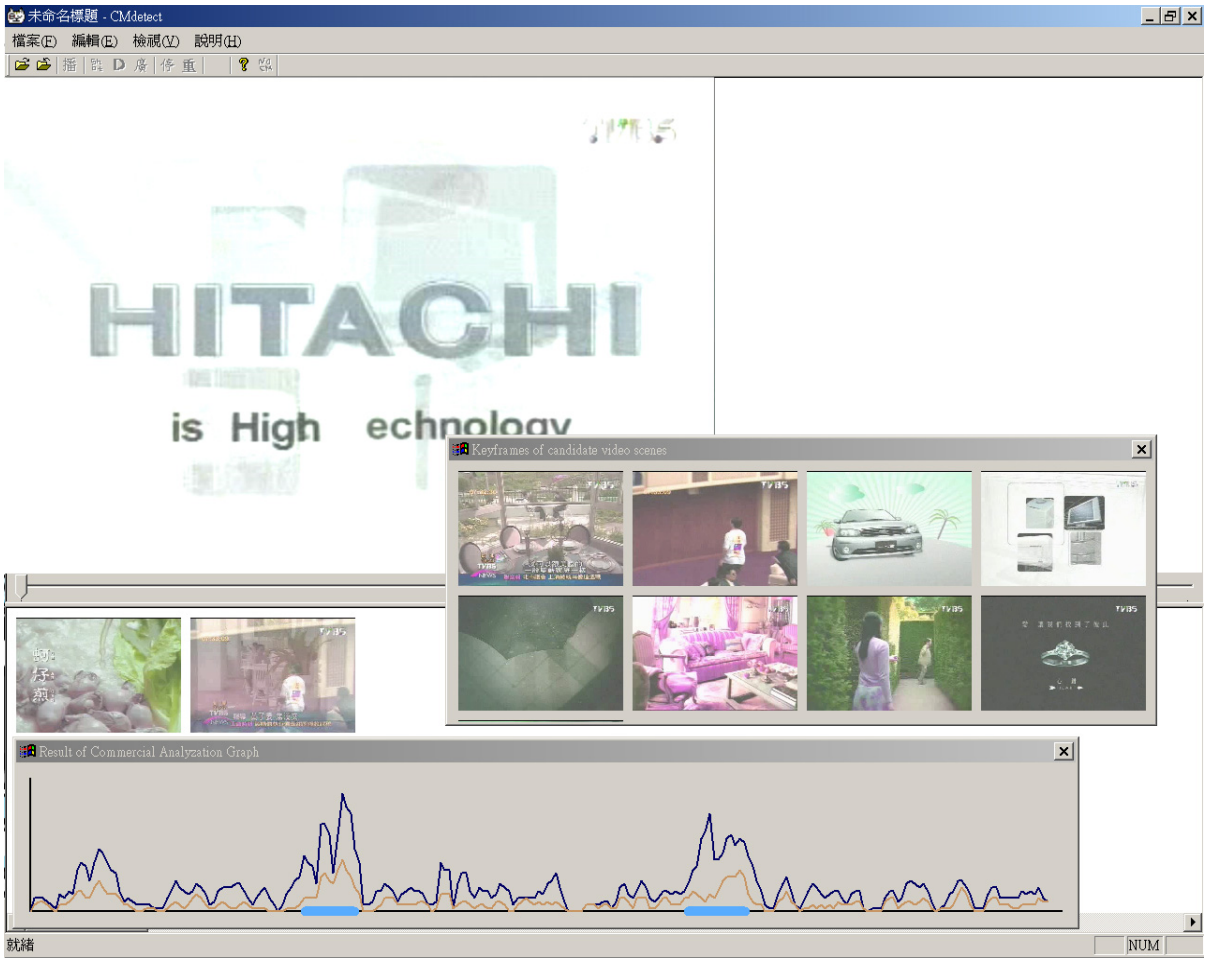


圖 16　新聞節目的廣告偵測介面

我們分析的是從電視上錄下來的新聞節目，在研究廣告偵測及廣告拍攝相關文獻之後，利用廣告時段影片用來吸引觀眾的手法，利用廣告的鏡頭變換(shot change)次數通常會比新聞節目多的一個特性，先大略的標出廣告的可能存在區域之後，更進一步利用廣告影片均特別強調顏色的特性，和廣告和新聞的不同點，計算出可能區域裡頭不同場景間的凝聚性(coherence)來找出更精確的廣告區域。最後可以依據不同的應用模式，來決定只留下廣告或是刪去廣告。

- 視訊興趣區自動決定系統

近年來，隨著多媒體產生技術的快速進步，加上網際網路對於資訊傳播的便利性，造成了多媒體文件 在數量上呈現急遽的增加。面對如此大量的多媒體資訊，越來越多的使用者希望能僅獲得該些文件簡要的內容精華，而非原始的完整文件。而另一方面，雖然在影像處理、人工智慧及電腦視覺等相關領域已有許多重要的研究成果，然而，其對於如何能自動分析出多媒體文件內容之語義仍無法解決。因此，如何能依照人類知覺上的感知，來發展出一套符合人類知覺需求的興趣區 (Region-of-Interest, ROI)決定系統，也就是可找出多媒體文件中對於觀察者較具意義或重要性的部份，將是數位化時代多媒體管理及研究當前一個重要的課題。

傳統的興趣區分析主要著重於兩種多媒體型式：影像（image）與視訊（video）。然而，在視訊方面的研究成果卻遠落後於影像的相關研究。這種情形肇因於忽略了許多視訊獨有的特性。第一、視訊可視為由一連串的訊框（frame）沿時間軸所構成，而越相近的訊框彼此間具有的相關性越高，而非互相獨立存在的影像；第二、在視訊的產生過程中，拍攝者常會使用運鏡（camera motion）的技巧來強調視訊中的重點或藉以引導觀眾的注意力。因此在分析視訊興趣區時，我們將考慮「應用媒體美學」（applied media aesthetics），也就是利用視訊拍攝時慣用而遵循的基本準則，來增加興趣區分析時的準確度與代表性。

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。在這個研究中，視訊的注意特徵值（attentive features）及應用媒體美學的知識都被同時考慮且利用。本研究將成為達到更高階具意義性視訊分析的一個重要基礎。

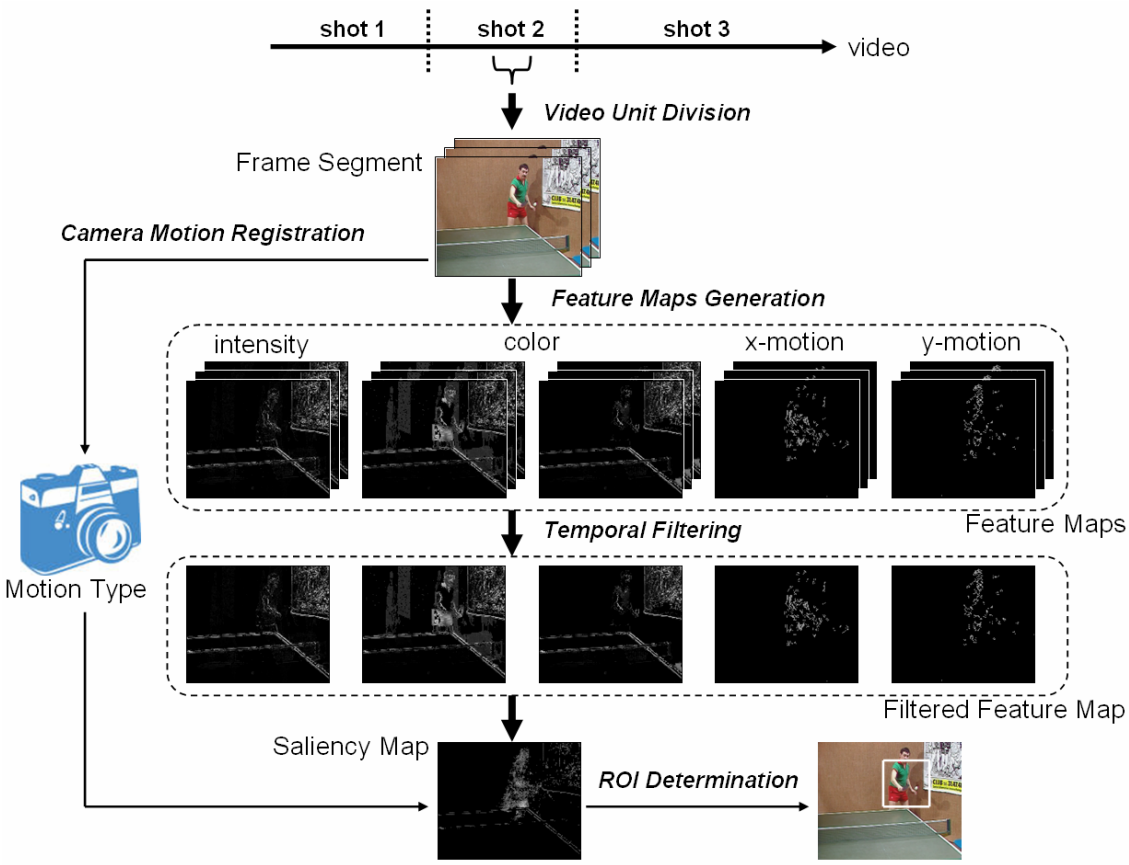


圖 17　決定視訊興趣區之系統流程

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 17　即為本系統之流程圖示。

➢ 音樂導向視訊摘要系統

根據經驗，人們對於沒有劇情、沒有旁白，以及沒有配樂的影片最缺乏耐心，而這些正是一般家庭影片所具備的特色。因此，未經過剪輯的家庭影片通常都無趣且乏善可陳。另一項在 MIT 的研究報告也指出，使用者如果對於同樣的影片搭配不同的配樂，較好的配樂會讓人們覺得影片的內容較佳。此外，我們認為，如果能將影片與配樂上進行良好的搭配（讓影片配合音樂的節奏作剪輯），可以收到顯著的加成效果。

因此，為了解決前述家庭影片所遇到的問題，我們設計了一套全自動化的視訊摘要系統，允許使用者選擇自己喜愛的音樂，讓系統根據音樂的節奏，以及影片本身的重要性，配合音樂長度自動地摘要剪輯出一段搭配音樂的影片（類似於商業電影的預告片）。本系統可配合音樂的節奏，加入適當的轉場特效，剪輯出具有專業質感的影片。也由於整個過程是全自動的，可以讓使用者免除複雜的視訊剪輯軟體學習過程。圖　為本音樂導向視訊摘要系統的使用者介面。

圖 18 音樂導向摘要系統的使用者介面

本音樂導向摘要系統的流程可細分為三個部分，茲列舉如下：

(1) 媒體分析（Media Analysis）：這部分包含針對輸入的家庭影片以及所選擇的配樂進行分析，利用媒體內容分析領域所發展出來的一些方法及工具，分析出影片中的重要片段，以及配樂裡的音樂節奏。
(2) 媒體同步結合（Media Synchronization）：利用前一部份所得到的分析資訊，我們提出兩種不同的演算法，配合參數的設定，將選出影片重要片段配合音樂的節奏加以剪輯，並加上適當的轉場特效。
(3) 媒體產出（Media Production）：第二部分中所產生的僅是存在電腦記憶體中的結構化資訊，在媒體產出的這部分，我們可以針對不同的需求，使用編輯腳本語言（Editing Scripting Language）直接輸出成影片—全自動化的產出方式。也可以輸出成目前市面上視訊編輯軟體的專案格式，再由這些編輯軟體進行更細部的編輯動作—此為半自動化的產出方式。

在最後完成的開發成果中，我們使用了Microsoft DirectShow Editing Service作為全自動化的編輯腳本語言；而使用訊連科技CyberLink PowerDirector視訊編輯軟體的專案格式作為輸出，並使用PowerDirector作為半自動化的細部編輯軟體。本系統以及其相關的演算法已被合作廠商訊連科技採納成為其核心技術，並將相關技術規劃為視訊剪輯商業產品MagicDirector，以及應用於現有產品PowerDirector中的新技術MagicCut等各項實際商業應用。

四、結論與討論

在各分項技術及兩項系統的開發過程中，參與本計畫的研究人員，已充分掌握到 MPEG-4/7 所採用的關鍵技術。由於 MPEG-4/7 所涵蓋的領域相當廣泛，包含系統、資料壓縮、三維繪圖、網路通訊、串流技術、訊號處理、圖形辨識等等，而所使用的技術亦相當具有深度。因此，除了可協助業界在未來 MPEG-4/7 相關應用產品的研發上，快速取得領先的地位外，對於整體資訊軟體業的技術提昇，亦有相當大的助益。除此之外，由於 MPEG-4/7 中包括了一些目前尚未成熟的技術，如 MPEG-4 所提出的多人世界 (MUW, Multi User World)、MPEG-7 在內容分析上所提出的知覺模型(Conceptual Model) 等等，仍是極為熱門的研究題目。所以，本計畫除了對業界有所幫助外，在學術上的研究，預期也將會有不少的突破及論文發表。

五、參考文獻

[1] Meng-Jyi Shieh, Chien-Feng Huang, Cha-Dong Duh, Wei-Teh Wang, Wen-Chin Chen, “Implementation of an efficient MPEG-4 2D/3D Mixed Renderer for Visual Editing and Interactive Applications,” International Conference on IEEE MPEG-4 3rd Workshop & Exhibition, San Jose, CA, U.S.A., June, 2002.

[2] 林劭穎 ：”一個以 MPEG-7 參考軟體為基礎的多媒體搜索系統架構（An MPEG-7 Content-Based Analysis System Architecture）”, 碩士論文。

[3] Ding-Yun Chen and Ming Ouhyoung, “A 3D Object Retrieval System Based on Multi-Resolution Reeb Graph”, Computer Graphics Workshop, Tainan, Taiwan, June 2002.

[4] 梁菁秀, “Design and Implementation of Prefetching Algorithm for Rich-Media Presentation across the Network”, 碩士論文。

[5] 陳盈慧, “Design and Implementation of Synchronization Mechanism for Rich-Media Presentation across the Network”, 碩士論文。

[6] 台大資工所謝孟吉，博士論文，”MPEG-4系統的設計與實作”，民國92年6月。

[7] 台大資工所禹智鏹，碩士論文，”MPEG-4平順傳輸串流之設計與實作”，民國92年6月。

[8] 台大資工所楊鈞傑，碩士論文，”MPEG-4多人世界系統架構之研究”，民國92年6月。

[9] 台大資工所林聖斌，碩士論文，”衛星電視訊號擷取之模組”，民國 92 年 6 月。

[10] 台大資工所何嘉強，博士論文，”以使用者為本視訊串流技術之研究”，民國 92 年 6 月。

[11] 台大資工所郭晉豪，博士論文，”以媒體內涵為處理基礎之相關研究及其在電視新聞節目影像上的應用與實作”，民國 92 年 6 月。

[12] 台大資工所陳賢碩，碩士論文，”MPEG-4 數位影像壓縮系統之軟體實作”，民國 92 年 6 月。

[13] 台大資工所王碩文，碩士論文，”H.264/AVC 數位影像解壓縮系統之探討與實作”，民國 92 年 6 月。

[14] 台大資工所曹智富，碩士論文，”自動於數位視訊中偵測字幕”，民國 92 年 6 月。

[15] 台大資工所方子維，碩士論文，”針對新聞影音資料之有效摘要及瀏覽工具”，民國 92 年 6 月。

[16] 台大資工所郭振彬，碩士論文，” 基於視覺和聽覺分析的新聞視訊分割”，民國 92 年 6 月。

[17] Ding-Yun Chen, Xiao-Pei Tian, Yu-TeShen and Ming Ouhyoung, “On Visual Similarity Based 3D Model Retrieval”,Eurographics 2003, Granada, Spain, Sep 2003

[18] Wan-Chun Ma, Fu-Che Wu, Ming Ouhyoung, “Skeleton Extraction of 3D Objectswith Radial Basis Functions”,Shape Modeling International 2003,Seoul, Korea, May 2003.

[19] Pin-Chou Liu, Fu-Che Wu, Wan-Chun Ma, Rung-Huei Liang, Ming Ouhyoung,

“Automatic Animation Skeleton Construction Using Repulsive Force Field” Pacific Graphics 2003, Canmore, Alberta, Canada,Oct 2003.

[20] Fu-Che Wu, Wan-Chun Ma, Ping-Chou Liou, Rung-Huei Laing, Ming Ouhyoung, “Skeleton Extraction of 3D Objects with Visible Repulsive Force”, currently submitted.

[21] Ding-Yun Chen, “On Visual Similarity Based 3D Model Retrieval”, Ph.D. thesis, Dept. of Computer Science and Information Engineering.

[22] Xiao-Pei Tian, “A 3D Model Retrieval System Based on MPEG-7 3DSSD”,Master thesis, Dept. of Computer Science and Information Engineering.

[23] Yan-Hong Lu, “A Feature Preserving Multiresolution Volumetric Sculpting System”, Master thesis, Dept. of Computer Science and Information Engineering

[24] 其他參考文獻請參考總計畫書(NSC 92-2622-E-002-002)。

誌謝

本計畫之第一期、第二期及第三期得以順利進行，首先要感謝行政院國家科學委員會、訊連科技股份有限公司、太極影音科技股份有限公司以及台灣大學資訊工程系所的大力支持。此外，所有參與計畫的助理，研究生，以及合作伙伴的工程人員，才是本計畫目前成果的幕後功臣。

參與本計畫之人員名單如下：國立台灣大學資訊工程學研究所吳家麟教授，陳文進教授，歐陽明教授，黃肇雄教授，項潔教授，周承復助理教授。國立台灣大學資訊工程學研究所博士班研究生：黃奕勤、何嘉強、童怡新、林奕成、莊玉如、林佳緯、葉正聖、郭晉豪、謝孟吉、黃俊翔、劉柏伸、吳賦哲、何建璋、馬萬鈞、林育慈、朱威達、陳昭宇。國立台灣大學資訊工程學研究所碩士班研究生：郭振彬、王頌文、歐俊岳、曹智富、陳賢碩、方子維、黃振修、沈至豪、陳萱瑋、鄭文皇、葉人豪、潘廷建、梁致豪、楊鈞傑、林聖斌、禹智鏘、林彥光、張燕君、楊書旻、殷圖駿、楊國鑫、徐明煒、蔡鈞傑、莊光庭、蘇家樟、田曉珮、杜佩璇、呂彥宏、涂正廷、劉秉周、羅婉琪、卓聖堯、林弘德、沈育德、張培根。研究助理：陳雅琳，賴怡嘉。

國立臺灣大學資訊工程學研究所 博士論文

指導教授：陳文進 博士

以 MPEG-4 為核心之多媒體群體合作應用架構

A Multimedia Collaboration Application Framework Based on MPEG-4 Standard

研究生：黃奕勤 撰

學號：F86526064

中華民國九十三年六月

中文摘要

隨著寬頻與無線網路基礎建設的高度發展，電腦運算速度、顯示設備與儲存媒介等硬體技術的一日千里，人們透過各類普及運算設備(Pervasive Devices)取得多媒體資訊，並與其他使用者之間互傳訊息、共享資訊甚至進行虛擬會議等應用已經成為趨勢。而以往電腦輔助群組協同工作(Computer Supported Cooperative Work, CSCW)已廣泛地被運用在遠距教學、軍事模擬以及多人連線遊戲等等不同的領域。然而這些系統多半採用侷限於特定應用的封閉標準(例如美國陸軍的模擬系統 SIMNET)；或是應用現有技術以符合特定需求(例如以 Web 技術發展而成的聊天室)，前者缺乏擴充性與互通性，後者則因為遷就現有架構，並不適用於新興的媒體格式或網路技術。本論文將研究開發一個具有擴充性的多媒體群體合作應用架構。透過這個架構，可以簡化開發各類不同的多媒體群體合作系統(Multimedia Collaboration System)的過程。

要成功地發展出一套多媒體群體合作應用架構，有幾項關鍵技術必須加以考量。首先是互動媒體的呈現技術，由於媒體壓縮標準的蓬勃發展，輔以個人化服務的盛行，使得具備高度互動性的複合媒體需求日益增加。其次，網路連線遊戲、視訊會議以及遠距教學等應用的普及，多使用者環境下的同步技術變得相當重要；最後，對多媒體系統開發者而言，可程式化的應用程式介面將加速開發時程並減少開發成本。

基於以上幾點觀察，本論文將利用 MPEG-4 標準所定義的工具並參考其所提出的應用程式引擎(MPEG-J)以及多人世界(Multi-User World)架構，設計出一個多媒體群體合作應用架構。本論文設計並實作了四項核心模組，分別是：(1)MPEG-4 表現引擎，(2)媒體存取層，(3)MPEG-4 應用程式引擎以及(4)多人應用程式伺服器。MPEG-4 表現引擎是一個跨平台的互動媒體呈現架構，配合用戶以及伺服端的媒體存取層，能滿足多點傳輸媒體資料流的需求。而 MPEG-4 應用程式引擎提供了一組應用程式介面，外部程式可藉此來存取並操控 MPEG-4 表現引擎。多人應用程式伺服器則提供了可程式化的伺服端架構，配合與用戶端之間的遠端函式呼叫實現多人連線下的同步機制。

本論文所提出的這套架構，使得程式開發者能快速開發出不同類型的多媒體群體合作系統。為了驗證本架構之適切性，我們開發了「線上室內設計」、「具有三維共享物件之視訊會議」以及「多人連線遊戲」等多媒體群體合作系統。由開發者的經驗可知，本架構確實降低了這些多媒體群體合作系統的開發難度與成本。

國立臺灣大學資訊工程學研究所 碩士論文

指導教授：陳文進 博士

MPEG-4 應用程式開發系統的設計及實作

Design and Implementation of an MPEG-4 Application Engine

研究生：楊國鑫　撰

學號：R91922072

中華民國九十三年六月

中文摘要

MPEG-4 系統提供了一套工具，以開發具有高互動性的媒體內容，透過在場景中內嵌 JavaScript 的方式，可以做出許多種場景和使用者間的互動。

透過使用現有的 MPEG-4 系統的經驗，我發現到目前存在系統的一些限制，使得要開發更進階的使用者互動和網路應用時，有許多的困難。爲了克服這些限制，本論文開發了一個 MPEG-4 整合網路應用的系統，包括了使用者端和伺服器端的設計和實作，並藉著這套系統開發出一些 MPEG-4 的網路應用，以證實這套系統的可行性。

國立臺灣大學資訊工程學研究所 碩士論文

指導教授：陳文進 博士

Linux 上的 MPEG-4 互動式多媒體播放器實作

An Implementation of MPEG-4 Interactive Media Player on Linux

研究生：楊書旻　撰

學號：R91922101

中華民國九十三年六月

中文摘要

MPEG-4 是由 Moving Picture Expert Group 所訂定的一套 ISO/IEC 國際標準。MPEG-4 的目標是爲了整合各種自然與合成的媒體。藉由它所題供的場景描述機制，我們能以物件導向的方式建構一個同時包含各種異質媒體的場景。因爲 MPEG-4 的複雜度很高，要實作一個 MPEG-4 的系統並不容易，以致於 MPEG-4 不常被選作多媒體系統的解決方案。

本論文中提出一套 MPEG-4 互動式多媒體播放器的實作方式。該播放器是在 Linux 作業系統平台上實作。也因爲採用了許多跨平台的程式庫，使得播放器可以容易地移植到許多不同的平台上。播放器也能夠同時呈現各種不同的媒體類型，包括二維及三維的幾何物體、圖片、影片以及聲音。使用者也可以透過週邊裝置與媒體內容產生互動，動態地改變媒體呈現的內容。

國立臺灣大學資訊工程學研究所碩士論文
指導教授：吳家麟博士

應用內容知覺機制於音樂導向的視訊摘要系統
A Musical-driven Video Summarization System Using Content-aware Mechanisms

研究生：黃振修　撰
學號：R91922001
中華民國九十三年六月

中文摘要

在本篇論文中，我們基於一些內容知覺的機制，完成一個應用於家庭視訊影片，音樂導向的視訊摘要系統。在論文中我們探討了許多音訊/視訊的特徵，以用來輔助分析輸入的音訊和視訊資料，並使用這些分析後的描述資料，將輸入的音訊和視訊資料結合成我們的音樂影片。音訊與視訊的結合是將視訊的節奏搭配音訊的節奏而完成。

對於音訊與視訊的結合，我們也提出了四種不同的群組方式，給予使用者在影音結合的過程中擁有較多的彈性。由於使用者對於音樂有著較直接的認同感，本系統所產生的音樂影片顯示出較高的專業性及較佳的娛樂效果。根據我們所做的客觀測試，所有的測試者都對於我們的系統感到驚奇並留下深刻的印象。大多數的測試者都很高興能有這樣的應用程式來幫助他們自動地編輯他們所創作的影片。

國立臺灣大學資訊工程學研究所碩士論文
指導教授：吳家麟博士

利用使用者注意模型決定視訊之興趣區
User Attention Model in Region-of-Interest Determination on Videos

研究生：鄭文皇　撰
學號：R91922002
中華民國九十三年六月

中文摘要

隨著多媒體文件在數量上的急遽增加，人們對於如何簡明地表現該些文件的精華變得更加熱切。其中一個重要的技術即為興趣區 (region-of-interest, ROI) 決定。傳統的興趣區分析主要著重於兩種多媒體文件型式：影像 (image) 與視訊 (video)。然而，對於視訊方面的研究成果卻遠落後於影像的相關研究。這種情形肇因於沒有適當地考量影像及視訊兩者間在本質上的差異，同時更忽略了視訊獨有的部份特性。

面對如此一個具挑戰性的研究課題，我們提出了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。在這個研究中，視訊的注意特徵值 (attention features) 及應用媒體美學 (applied media aesthetics) 的知識都被同時考慮且利用。我們將視覺注意特徵值區分為三個基本種類：亮度（intensity）、顏色 (color) 及運動 (motion)。參考美學的原則，這些特徵值以一個新提出之稱為訊框切片 (Frame-segment) 的視訊分析單位為基礎，同時依據攝影機運鏡 (camera motion) 的種類而加以整合。在實驗中，對於數種不同的視訊資料進行了興趣區分析及使用者相關研究並證明了所提架構的有效性。我們視本研究為達成更高階具意義性視訊分析的一個重要基礎。

國立臺灣大學資訊工程學研究所碩士論文
指導教授：吳家麟博士

電視新聞節目中的廣告偵測
TV Commercial Detection in News Videos

研究生：葉人豪　撰
學號：R91922035
中華民國九十三年六月

中文摘要

由於廣告拍攝手法和內容的多樣性，讓自動偵測電視節目中的廣告一直是一個困難的問題。本論文在研究了有關於廣告偵測文獻以後，加上一些製作廣告的規則以及對於廣告的觀察，提出了一個由上而下的方法來偵測電視節目當中的廣告。

因為廣告通常是影片中觀眾比較不感興趣的部分，所以導演會利用剪接等手法，或是在廣告中特別強調某些顏色，來讓廣告影片變得有趣，並且達到吸引觀眾和強調產品的效果。本論文利用找出影片中這些現象的出現時間，當作是廣告出現的依據。除此之外，為了使偵測到的廣告的片段更加的精確，本論文在廣告偵測系統之中接著加入了影片場景切換的偵測(video scene boundary detection) 試著來自動的找出新聞和廣告之間準確的切換點。實驗中並顯示，本系統可以良好的把新聞中的廣告區域偵測出來。倘若使用者對電腦自動選出來的廣告區域不滿意，此系統還能夠列出其他可能的切換點來讓使用者選擇。這使得本系統變成一個便利的切廣告輔助工具。

國立臺灣大學資訊工程學研究所碩士論文

指導教授：歐陽明博士、陳炳宇博士

針對形變建立可調整精細度的三維動畫模型

Constructing Scalable 3D Animated Model by Deformation Sensitive Simplification

研究生：卓聖堯　撰

學號：R91922029

中華民國九十三年六月

中文摘要

迄今，為了表現出比較重要的特徵或是比較細緻的結構，這個需求帶來了愈來愈多的高解析度的三維動畫模型。然而有些時候這樣高解析度的模型是不需要的，或者是不希望有這麼高的解析度。例如: 當我們只是在預覽一個 3D 模型動畫要來決定是否要下載時，我們就不需要去預覽這樣高解析度的動畫模型 ; 再舉另一個例子: 在一些互動的系統中，為了要有更好的系統執行效率，也不會去使用到這麼高解析度的動畫模型。

雖然現在有很多有名的演算法可以去化簡 3D 模型，但是都受限於靜態的模型上。一般來說，這些演算法是在一個固定姿勢的模型上估計化簡後所構成的誤差來做為化簡模型的順序來簡化這個模型，如此的話，在某些比較平常的姿勢上，可以得到一個很好的簡化模型，但是在一些特別的姿勢時，就可能破壞掉這個模型的重要特徵。

在此論文中我們提出了一個會考量到3D模型的所有動作的全自動化簡演算法，並且保留這些模型在不同的姿勢時所有的幾何上的特徵。我們也將用一些統計方法來與其他化簡方式做一個比較。

國立臺灣大學資訊工程學研究所碩士論文

指導教授：周承復博士

改善無線串流服務：
訊框丟棄與封包遺失原因之區分演算法

Improving Streaming Performance over Wired and Wireless networks : A Similarity Based Frame Discard Algorithm and A Trend Based Loss Differentiation Algorithm

研究生：徐明煒　撰

學號：R91922064

中華民國九十三年七月

中文摘要

近年來串流服務已成為網路的主要應用之一，加上無線網路快速普及，在無線網路上提供串流服務已成為一項重要的議題。然而由於無線傳輸媒介的特性，如較低的頻寬和較高的位元錯誤率，使得在無線網路上提供串流服務比在有線網路上，更具有挑戰性。

當我們經由一資源有限的網路傳輸串流影片時，有時候因為頻寬的不足，導致部分的資訊遺失是不可避免的。為了提高網路頻寬的使用率，我們必須先傳送對影片品質有較大影響的訊框(frame)。因此我們提出一以相似度為基礎的訊框丟棄演算法(Similarity-based Frame Discard)，依相似度決定各影片訊框的重要性。

目前串流連線普遍採用 TCP Friendly Rate Control (TFRC)為其壅塞控制機制以達到和傳統 TCP 連線公平地競爭網路資源的目的。如同 TCP，TFRC 在遇到封包遺失時，會認為網路已經進入壅塞狀態並調降傳送速度，以免網路的負荷量持續增加。但在無線網路上，由於較高的位元錯誤率，封包遺失可能是由無線傳輸媒介的不穩定所造成。因此，盲目地遇到封包遺失就調降傳送速度將會造成頻寬使用率的低落。為了解決此一問題，我們提出了以趨勢和遺失密度為基礎的遺失原因區分演算法，利用封包傳送時間的改變趨勢和封包遺失的密集程度，判斷封包遺失的原因是否為網路壅塞，抑或為無線媒介所造成。

分項計畫一：MPEG-4/7資訊媒體內容互動架構之研究 (III)

The Interactive Framework for MPEG-4/7 Instantaneous Information Media Content (III)

成果報告

計畫編號：NSC91-2622-E-002 –002

全程計畫：民國 90 年 08 月 01 日至民國 93 年 07 月 31 日

本年度計畫：民國 92 年 08 月 01 日至民國 93 年 07 月 31 日

計畫主持人　：陳文進　台灣大學資訊工程系教授
共同主持人　：吳家麟　台灣大學資訊工程系教授
歐陽明　台灣大學資訊工程系教授
黃肇雄　台灣大學資訊工程系教授
周承復　台灣大學資訊工程系助理教授
陳炳宇　台灣大學資訊管理系助理教授

一、摘要

本分項計畫為「MPEG-4/7 資訊媒體內容互動架構之研究」，用以支援總計畫【媒體內容工程: MPEG-4/7 相關技術之研發】所提出的“互動資訊媒體系統”。主要的研發目的是發展 MPEG-4 互動式媒體內容從伺服器端直至媒體播放端整套的播放機制及從伺服器端到媒體播放端的 Client/Server 溝通機制。強調能與使用者間互動的『資訊媒體』將是未來產業界的重點。如何讓媒體動態地提供給使用者想知道的最新資訊，是這一波媒體內容(Content)熱潮最關鍵的技術之一，而這些核心技術，都在 MPEG-4 與 MPEG-7 這兩套標準的規範之中。

在 MPEG-4 技術方面，本計畫強調『可程式化』的技術。由於目前世面上的 MPEG-4 播放架構非常沒有彈性，一項新的應用就需要重新撰寫一個新的 MPEG-4 播放器與伺服器。換言之，播放器與伺服器往往為了一項新的應用就必須重新改寫。本計畫預計設計一個『媒體播放端與伺服器端的 Java 解決方案』，試圖建構一個『可程式化』的媒體播放架構，利用伺服器端以及媒體播放端可抽換的 Java 程式碼，根據媒體目的的不同即時性的產生動態資訊媒體，去解決現階段媒體播放架構不夠一般化的問題，如此，我們便可以很容易和其他分項計畫的需求整合。例如，MPEG-7 的整合便可以很容易的在伺服器端利用『內嵌式的 Java 程式碼』來完成。

由於 MPEG-4 目前最直接的應用與最大的商機便是『互動電視』的應用。本計畫所發展的互動機制也正是『互動電視』最需要的核心技術。因此，將本計畫所發展的伺服器端技術應用在互動電視上便是一個很好的驗證。但是，若是要成功的應用在互動電視上，本實驗室前期產學合作計畫所發展的 MPEG-4 媒體播放器就必須做適度的調整，並增強其聲光效果，這方面也將跟分項計畫三『MPEG-4/7 三維虛擬實境內容處理技術』相互整合。

This sub-project is made for “The Interactive Framework for MPEG-4/7 Instantaneous Information Media Content,” and supporting “The Interactive Information Media Service System” of the main project “Content Engineering: Research on MPEG-4/7 Multimedia Technologies”. Its principal purpose is to develop the technology of the interactivity framework between the server-side and the MPEG-4 media player-side. In the future, the industry will put their focus on the information media that is able to interact with users. How to dynamically provide users with interesting information and content will be the key technologies, which are defined by MPEG-4 and MPEG-7 standard. We try to solve some major difficulties, including the communication mechanism between the server and the player, the architecture of creating real-time interactive media content in server side, and the framework of the client side accepts the input from users and responses to users by using the client side interactivity or deliveries the event back to the server for getting the latest information media content.

MPEG-4 is a international standard of multimedia application, which can meet all kind of requirement of different application and different hardware. However, the current framework of the MPEG-4 player is inflexible, we have to rewrite an extra MPEG-4 server for one new application. Moreover, the server and the client may have to be totally rewritten to meet a new requirement. This project expects to design “the Java solution of player’s and server’s end” and try to create a programmatic framework for whole of MPEG-4 playing process. We use replaceable Java codes at the server-side and client-side to solve the problem of lacking vague generalization in MPEG-4 playing framework at present, so we can integrate with the needs of other sub-projects easily. For example, the integration of MPEG-7 can use “embedded Java component” at server-side to achieve easily.

Because the most direct application and greatest merchandise chance of MPEG-4 is apply to interactive TV; besides, interactive function developed by this project is also the core of the most needed technology. Therefore, it’s very good to put to the proof to apply the technology developed by this project at server’s end to interactive TV. But if it is really applied to interactive TV, our MPEG-4 medium player developed by the previous experiment needs to be adjusted appropriately and enhanced its abilities, such as in 2D domain, the capability for multimedia application, and execution of Java embedded codes. This part will also be integrated with the sub-project “Content Manipulation for MPEG4/7 3D Virtual Environment”.

二、計畫緣由與目的

隨著個人電腦計算能力的一日千里，儲存媒介容量亦成倍數成長，再加上網際網路的普及、無線與寬頻網路技術的推波助瀾，各種類型的多媒體應用將如雨後春筍般地進入市場。從基本的單一媒體播放，到複雜度較高的複合媒體整合系統，像是導覽系統、網路 3D 連線遊戲、視訊會議等等，現在已經變的非常的普及。然而，隨著多媒體元素的多樣化以及複雜度的提昇，多媒體系統的開發變得越來越複雜，所必須面對的媒體格式也越來越多，以至於整個的開發成本變得越來越高。然而，這種為了特定目的而開發的多媒體系統不但難以擴充，而且與其他系統交換媒體資料更顯得窒礙難行。同樣地，媒體內容卻難以在各種不同的平台之間流通，無形中使得越來越多的成本浪費在無意義的工作上。

為此，MPEG 組織定義了 MPEG-4 標準，這個國際標準定義了一個新的多媒體框架(Framework)，試圖利用這個框架去整合各式各樣豐富的媒體，讓這些媒體可以互相合作去完成各種的多媒體的應用。MPEG-4 打破傳統以畫面為主的影片呈現方式，取而代之的，是以物件的方式呈現一個多媒體畫面的組成。這些物件的形式包括了圖片、影片、聲音、文字、3D 模型、2D 向量圖形等等。MPEG-4 並定義了一個方法去描述這些物件在時間上、空間上的安排，稱為 MPEG-4 場景描述。場景描述不僅可以描述各種不同異質媒體之間在畫面上以及時間上的呈現，更可以描述這些物件彼此之間如何的互相互動(事件驅動機制)，以及撰寫物件間彼此的邏輯關係(利用 JavaScript)，而這些特性，正好符合目前市場的發展趨勢。

另一方面「互動電視」這個名詞最近幾年經常出現在各報章雜誌。這是一個後 PC 時代的新技術，在網際網路興起、PC 市場已飽和的情形下，電視與網際網路的結合被視為下一波的網際網路應用的主流。最主要的著眼點在於電視與一般家庭生活緊密結合在一起，可將網際網路使用族群由原有的個人電腦使用者擴及整個的消費大眾，也就是說希望把 10-65 歲的電視觀眾帶入網際網路的世界。

當要將網際網路應用推廣至一般電視觀眾時，要考量幾個限制因素。一、此類應用的模式應不離使用者觀看電視的操作習慣，而所謂觀看電視時的操作習慣表示使用者手中所握的將不會有鍵盤、滑鼠，而是至多只有電視遙控器來控制所有顯示；二、必須在電視功能外加入符合網際網路的應用，如在電視上可收發電子郵件、進行線上聊天、或網頁瀏覽等。三、電視與網際網路結合後必須有加乘價值的應用，所以並非只是一俱備上網功能的電視，而應是除了兩者原有的功能外，可另提供結合後的另類應用。如電子節目導引（EPG）、即時互動的電視節目等。

雖然互動電視在許多年前就已經出現在市場上，但經過多年的競爭，除了一些歐洲國家之外，目前並沒有成功的案例出現，它不普及的原因是大家無法互相信服彼此的規格，各大廠商互不相讓的結果，最後只造成互動電視一直處於喊口號的階段。

但這一切都將在 MPEG-4 這個標準的出現而有所改觀，因為 MPEG-4 並非由廠商

主導，而且是個世界公認的標準，它的互動機制又是目前最齊全的，因此，預計很快就成為新一代互動電視的標準，目前許多研究 MPEG-4 的廠商或是學術單位，大多都有將 MPEG-4 應用在互動電視的計畫，無論從研究或是商業上的角度來看，互動電視都實在是一個不可缺席的研究領域。

因此，本計畫將會把 MPEG-4 為主要的技術核心應用在互動電視上，並支援總計畫中”互動資訊媒體服務系統”(Interactive Information Media Service System)之研發。MPEG-4 原本是個龐大而且複雜的標準，讓許多組織望而卻步，但對本實驗室而言，這卻是一個很好的機會，因為本實驗室已經在前三年的產學合作計畫【MPEG-4 複合媒體及網路虛擬實境之研發】(NSC89-2622-E-002-013) 中累積了足夠的背景知識以及成果，這是一個很好的機會讓本實驗室驗證 MPEG-4 互動機制之可行性。但是，一個 MPEG-4 Interactive TV 與一個 MPEG-4 Player 之間的架構仍然有許多的差異性，本分項計畫也將試圖突破這當中的技術難題。

本分項計畫為三年計畫，其逐一的目標在:

1. 加強前期產學合作案在 MPEG-4 方面的成果，使其能夠完全展現 MPEG-4 系統核心的主要元素，包括了聲音特效(BIFS-Audio)，動畫機制(BIFS-Animation)，以及可程式化互動機制(JavaScript 引擎)。
2. 建構高彈性的 MPEG-4 媒體串流伺服器，並整合多人互動機制，完成整套完整的 MPEG-4 解決方案。
3. 依當時電視工業現況將 MPEG-4 成果應用在互動電視的領域上，並且整合其他分項計畫。

三、研究方法

【Linux MPEG-4 互動媒體播放器】

1. 前言與研究目的

Linux Operating System 是近年來在各個領域相當熱門的主題，它對軟體開發提供非常高的自由度，開發工具的健全及作業系統的高效能與穩定度，使它不論在即時系統應用、多媒體開發平台、嵌入式平台至手持式行動裝置等都成為一個深具潛力的明日之星。因此，結合 Linux 的平台優勢，本計畫所開發之 Linux MPEG-4 互動式場景播放器在可預見的將來，可提供廣泛且強力的多媒體內容呈現，亦為邁入 Multimedia Everywhere 的第一步。

2. **研究方法**

在 Linux MPEG-4 互動式場景播放器的實作上，我們採用開放並跨平台之軟體函式庫(SDL、OpenGL、FFMpeg)，因此，這個實作提供了開發 MPEG-4 播放器之軟體範本，吾人可以用低成本的方式開發出在其它平台上之 MPEG-4 播放器。

Linux MPEG-4 互動式場景播放器過程採用模組化的設計，整個播放系統包括以下模組：

- 媒體存取模組(Media Access module)
- 場景執行模組(Scene Execution module)
- 場景合成模組(Compositor module)
- 媒體解碼器模組(Media Decoder module)

圖 1為系統示意圖，模組內容詳述於後。

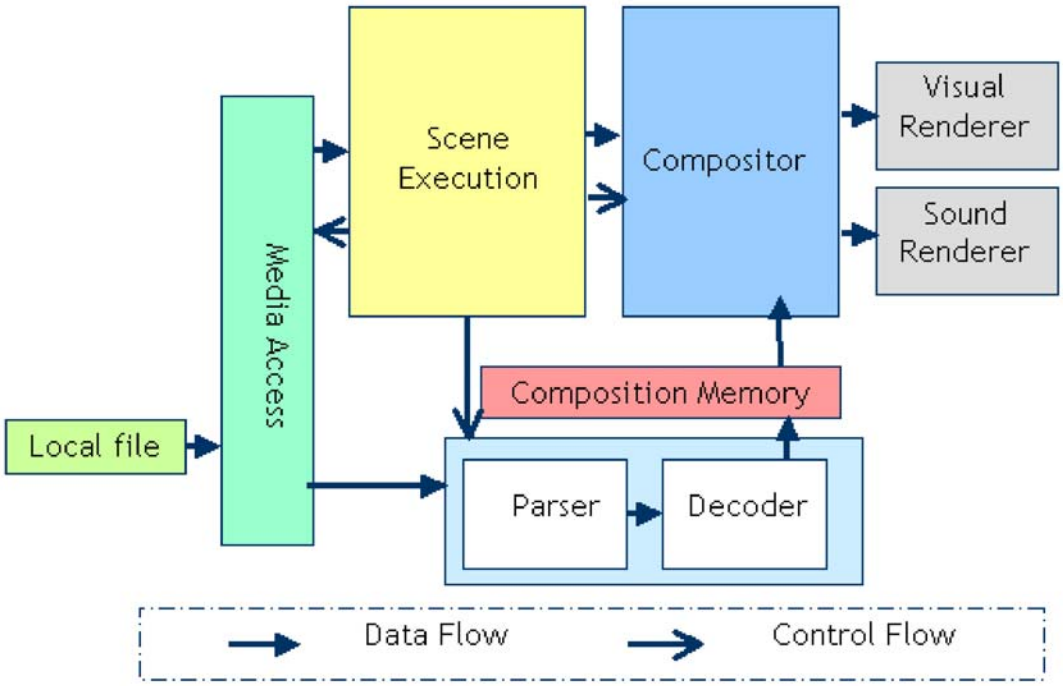


圖 1: MPEG-4 系統架構圖

甲、媒體存取模組(Media Access module)

媒體存取模組負責讀取場景描述或是媒體串流，再將分析過後的場景描述交給場景執行模組來執行。而當場景合成模組需要媒體資料的時候，其模組也會送出控制訊息給媒體存取模組，媒體存取模組會將所需的部份取讀並傳送給場景合成模組來解碼。

此模組讀取的對象有兩種，一種為場景描述(Scene Descriptions)，一種為原生的媒體檔，例如影片檔等。

乙、場景執行模組(Scene Execution module)

場景執行模組會依場景建出一個場景圖(Scene Graph)，它會負責初始化場景圖並且處理訊息路由(Routing)連結。當場景初始化後，場景執行模組就會依固定週期動態地更新場景，而會造成場景變化的機制，主要有二：

(1) 事件路由(Event Routing)
(2) Scripting

丙、場景合成模組(Compositor module)

場景合成模組負責合成並呈現聲音及視覺的物件，它利用建立 Rendering Resource 的方式加速處理複雜的 MPEG-4 多媒體場景，這是 Linux MPEG-4 多媒體撥放器中核心地位的一個模組。

一般來說，一個 MPEG-4 的場景通常包括有 2D 幾何物件、3D 幾何物件、影片、聲音等等。下面描述場景合成模組如何處理這些類別：

2D/3D 幾何物件：3D mesh 物件由 vertices 座標、法向量及材質坐標構成，我們利用 OpenGL 對 Vertex Array 的支援來達到加速的效果，使得幾何物件的 rendering 能有效率地完成。

2D 圖片：在 Linux MPEG-4 播放器的實作中，圖片被視為幾何物件的材質。也就是說我們用一個簡單的平面幾何代表圖片跟場景的連結。實作上，在載入圖片檔的同時我們也建立了其 Mipmap，雖然建立 mipmap 需要額外時間，但是在遠近的視覺效果上，較不會出材質突然改變的突兀情形，有較好的視覺效果。

影片：影片在場景中同樣是以材質來處理，通常影片的 frame rate 約在 24fps 到 30fps 之間，因此場景合成模組更新材質的速度有很高的要求，在實作中，若更新速度不及 frame rate 的話，將會捨棄其 frame。

我們採用 FFMPEG 的 libavcodec 函式庫來解碼影片及聲音物件。libavcodec 函式庫是具有跨平台特性的函式庫，在大部份平台都能編譯，並且擁有良好的解碼效率。

由於解碼影片需要大量計算，為了應付 render 的即時需求，我們引用了暫存緩衝區機制來減少播放時的延遲，由一個 Packet 執行緒來填放 Packet Buffer，然後由一個 Video 執行緒來即時解碼 Buffer 中的 Packet 並將解碼後的畫面放至一個 queue 中，此 queue 中的畫面即可用來即時繪製到材質上。

聲音：相同於影片，聲音之 packet 由 Packet 執行緒讀出並放至 Buffer 中，由另一個解碼執行緒來解碼出聲音的資料，並放到音效卡的 Buffer 中撥放。若有二個以上的聲道，我們採用 SDL_mixer 函式庫來進行混音的動作。

丁、媒體解碼器模組(Media Decoder module)

這裡包含了各種媒體的解碼器，包括 MPEG 1/2/4、Mp3、JPEG、WMV、BMP、PNG…等等的解碼器。原則上我們採用的都是 Open-Source 的函式庫。

戊、同步(Synchronization)

MPEG-4 場景中可包含各種類型的媒體物件，然而，要確保這些異質物件(MovieTexture、TimeSensor、AudioClip 等等)能夠在不同裝置上同步地被 render 是一件不容易的事。我們利用 SDL 所提供的時間函式來實作一個計算同步的架構，確保在每個 frame 更新的時候，每個 BIFS 節點能正確地被 render。圖 2為 MovieTextureNode 的 Timing Model，若 movieTexture 的 isActive 為 true，則場景合成模組就會依時間資訊(見圖 3)從 Buffer 中拿出一張該點的 frame 來播放，而達到同步的效果。

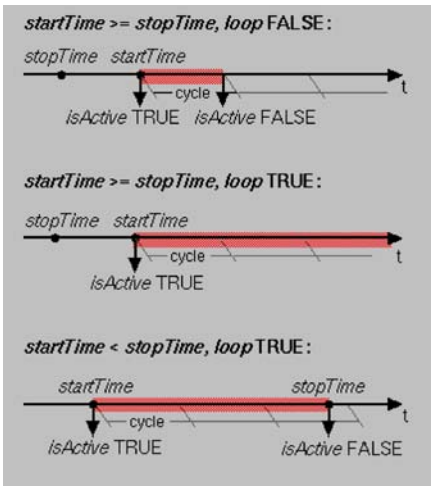


圖 2: MovieTexture Timing Model

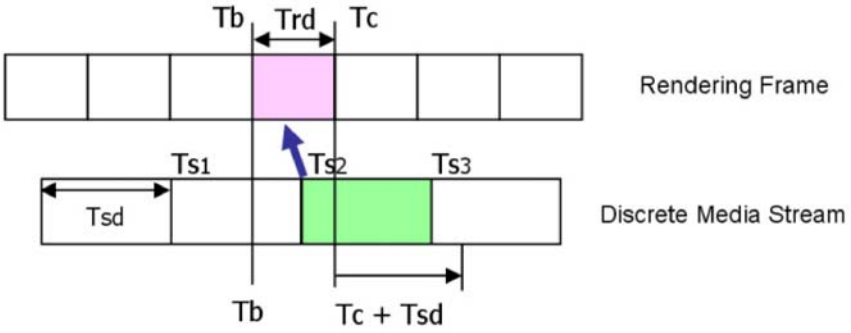


圖 3: MovieTexture 之同步

3. 實驗及成果

本研究成功地開發並實作 Linux 環境下之 MPEG-4 互動式多媒體播放器，能提供新一代 MPEG-4 多媒體應用於 Linux 上，將 MPEG-4 標準進一步推廣到非 Windows 之平台。將來能成為客戶端進行基本單一媒體撥放、複雜度較高之複合媒體整合系統，像是導覽系統、網路 3D 遊戲、視訊會議等應用之平台。

此 MPEG-4 互動式多媒體播放器具有以下特點：

1. 支援 ISO 媒體檔案格式 (MP4)

可以播放符合 MPEG-4 標準的多媒體檔案、串流內容服務，意即符合 MPEG-4 BIFS 標準的多媒體場景。本實驗室已開發有 MPEG-4 Scene Editor 致力於產生 MPEG-4 的互動式場景，其產生之檔案即為 MP4 格式。透過此 MPEG-4 播放器，將可在 Linux 相容平台上播放豐富互動的多媒體內容。

2. 實作 MPEG-4 BIFS 及 Scene Graph

BIFS 是 MPEG-4 用來描述場景中影音物件如何組織構成整個場景，它階層式地將物件編碼成二進位的共通格式並支援百餘種不同的物件。本實驗室開發之 Linux

MPEG-4 播放器採用 BIFS 為核心場景描述語言，並依照 MPEG-4 標準實作各種不同的 BIFS 節點。

3. 可 Render 2D/3D 合成場景

近來多媒體內容已不再滿足於單純 2D 的畫面，3D 的設計提供更多可能的感觀效果。而 MPEG-4 本身就支援自然與合成的多媒體物件，因此本實驗室開發之 Linux MPEG-4 播放器可以在 Linux 平台下播放 2D/3D 混合的場景，讓 Linux 平台上的媒體呈現更豐富。

4. 支援 JavaScript 程式

Script 機制提供 MPEG-4 場景更多更強的互動能力。本實驗室所開發的 MPEG-4 JavaScript 引擎可以製作具有互動性的 MPEG-4 多媒體場景。在此 Linux MPEG-4 播放器中我們亦加上 JavaScript 模組，讓我們可以在 Linux 上處理 script 對場景的動態改變。

此 Linux MPEG-4 播放器不僅能呈現完整的 MPEG-4 互動場景，並能根據使用者點選場景中的物件所引發的事件，正確地執行相對應的 JavaScript 程式碼。進而可以設計各式不同的多媒體互動場景，像是數位電視購物、2D/3D 遊戲、遠距教學等等…

下圖是一個在 Linux 平台上播放的互動電視應用範例。這個多媒體場景整合了 2D/3D 物件、多媒體影片、及互動式選單等。可以提供使用者觀賞電視節目像是新聞、氣象報導、電視購物等功能，展現 MPEG-4 多媒體播放器其應用功能。

圖 4: MPEG-4 互動電視應用範例

【PocketPC MPEG-4 互動媒體播放器】

1. 前言與研究目的

隨著硬體技術的進步以及寬頻網路的發展，MPEG-4 逐漸成為最重要的國際標準。MPEG-4 和一般以 frame 為基礎的多媒體標準不同，它採用物件導向的方式，並把靜態圖片、二維三維圖形、動畫、影片、音效、以及虛擬實境等多媒體技術整合成一個整體的架構。MPEG-4 提出了場景(Scene)的概念，以及一個描述場景中各個多媒體元件間時間空間關係的機制。此機制在 MPEG-4 標準中稱為「Scene Description」，正式名稱是「BInary Format for Scenes」(簡稱 BIFS)。透過 BIFS，場景作者可以定義媒體物件之間的互動行為和邏輯關係。由於 MPEG-4 具有「以物件為基礎」以及「可程式化編輯」的這兩項特性，將為我們帶來嶄新的、各式各樣的多媒體應用。

MPEG-4 的重要特性之一為其跨平台的特性。近年來嵌入式裝置已成為業界與學界熱門的研究課題，可以預見的是，可攜式裝置必將成為新興多媒體應用的消費平台。然而，嵌入式裝置常常受限於其較低的計算能力以及有限的硬體支援，因此，在嵌入式裝置上成功地開發 MPEG-4 系統，不論在設計或實作方面都還有許多技術議題尚待克服，包括了：1)較低的計算能力、2)較少的硬體加速支援、3)受限的記憶體空間、4)缺乏內建的多媒體函式庫。針對前三項議題而言，我們需要一些最佳化設計以降低系統的複雜度。如果忽視這些系統資源的限制，則將造成過長的回應時間以及沉重的耗電量。對於最後一項議題，嵌入式裝置上通常沒有像 Windows DirectX 這類高階多媒體應用程式介面(API)，以方便程式設計師使用硬體(如 3D 加速晶片、音效卡等)所提供的多媒體功能。再者，用來播放多媒體串流的函式庫，例如微軟的 DirectShow，對嵌入式裝置而言又太過複雜與龐大。因此，我們需要為嵌入式裝置上的影音串流的解碼與同步設計一套解決方案。

綜合以上討論，本研究將試圖在嵌入式裝置上，開發一個互動式的 MPEG-4 媒體播放器，除了具備 MPEG-4 物件導向、互動性與可程式化等功能之外，並針對嵌入式裝置上硬體資源受限與缺乏完整多媒體支援等議題，研發適當的演算法以及系統最佳化設計。

2. 研究方法

甲、系統架構

MPEG-4 處理流程

圖 x 說明了呈現一份 MPEG-4 媒體的處理流程。首先，存放在本地檔案或遠端伺服器的 MPEG-4 媒體中的場景資料先被擷取出來，場景資料接著被解碼，然後組成場景描述(Scene Description)。接著 Scene Graph 被建造出來。根據 Scene Graph，媒體資料可以從不同的來源被擷取出來。這些媒體串流被解碼到 Media Buffer，準備被 Compositor 使用。最後，Compositor 負責呈現整個 MPEG-4 媒體的工作。

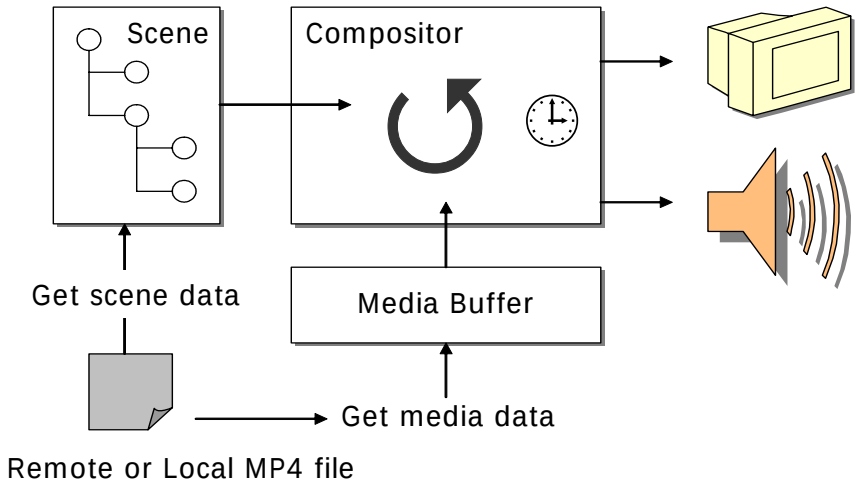


圖 5: MPEG-4 媒體處理流程

要成功實作出一個 MPEG-4 場景播放器，必須解決一些必要的課題：首先，Scene Graph 及其相關資源必須有良好的管理；再者，系統必須支援各式各樣的媒體格式，而這些媒體資料可能來自不同的來源，因此在串流的方式上，系統不但要能夠存取本地的儲存裝置，還要能存取遠端伺服器的儲存裝置。最後，所有相關媒體必須要能同步的呈現。由以上的討論，我們的系統設計如圖二所示。此系統由好幾個模組組成：場景圖管理者、媒體解碼架構、合成緩衝器、聲音引擎、2D/3D 圖形引擎。

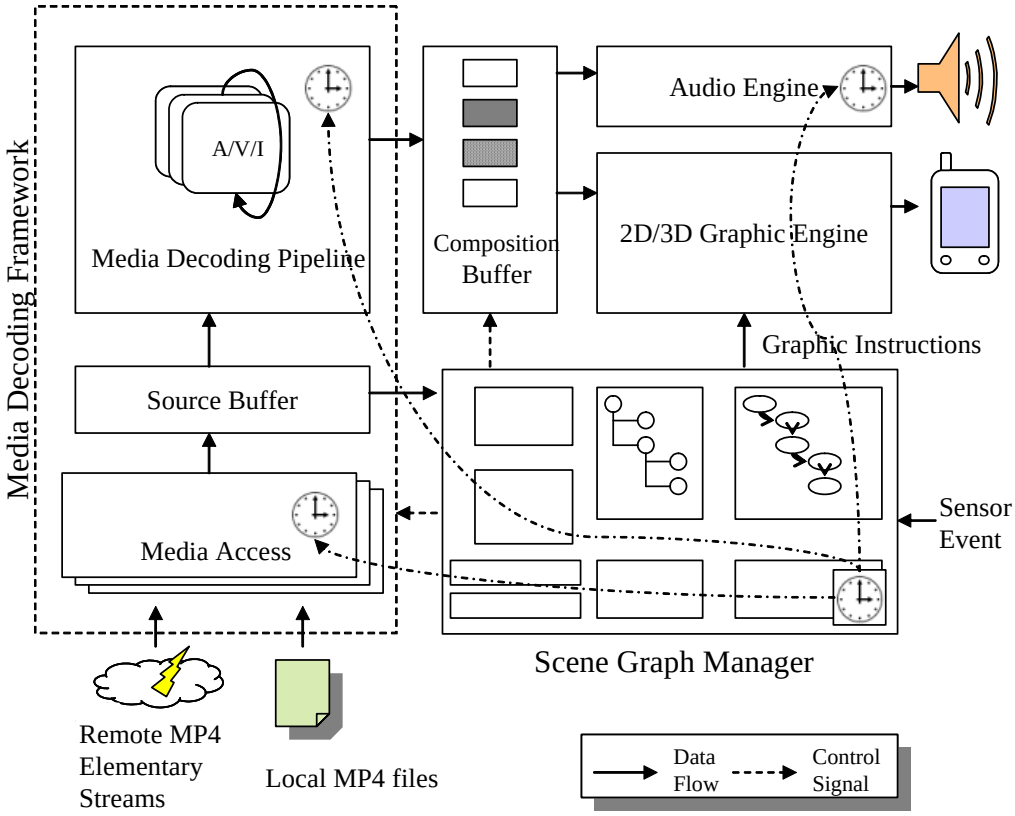


圖 6: 系統架構圖

各個模組的功能如下：

- Scene Graph Manager：系統的核心，管理場景圖的一些操作。
- Media Decoding Framework：負責擷取各式各樣的多媒體資料，並且加以解碼。
- Composition Memory：用來儲存已解碼的資料。

在 semi-pull model 下，場景圖管理者使各種媒體組成物件的呈現方式能夠同步化。為了合成整個場景，場景圖管理者也必須要根據被場景描述所定義媒體物件的 2D/3D 空間關係的資訊，來產生圖形化的指令。根據這些指令，圖形引擎就能夠有效的描繪出 2D/3D 圖形的基本圖元，就像影像一樣。同時，聲音引擎也能夠有效的呈現出已解碼的聲音串流。

媒體解碼架構

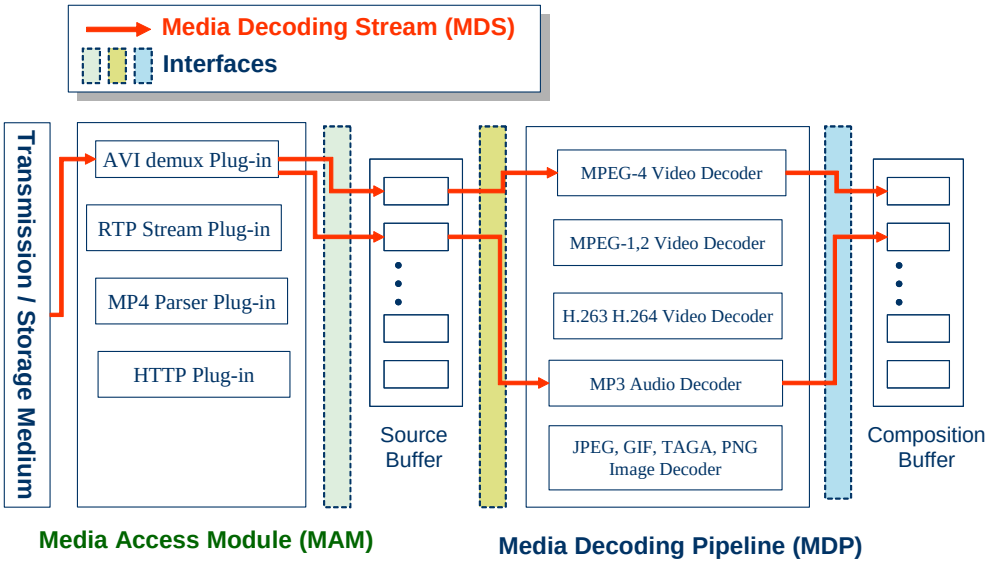


圖 7: 媒體解碼架構

MPEG-4 支援來自不同來源，由不同檔案格式或網路協定傳送過來的各式各樣的多媒體格式。為了滿足這些需求，於是我們提出一個用來對廣泛媒體串流解碼的一般化架構。在這個架構中，一個媒體解碼串流(Media Decoding Stream ,MDS)被建立，此被用來連接媒體存取模組(Media Access Module ，MAM)和媒體解碼管線(Media Decoding Pipeline ，MDP)，如圖 X 所示。三個一般化的輸入/輸出介面和來源緩衝器，就像連接的橋樑，使得 MAM 和 MDP 可以獨立的運作。此三個介面分別為 MAM 的輸出介面和 MDP 的輸入/輸出介面。不論何時，如果一個實作此介面的新元件被加入此系統中，則此系統皆能夠處理新的媒體資料。此架構提供一個合理的、可延伸的機制來控制各種被支援的多媒體串流。

場景圖管理者

場景儲存圖形管理(SCGM，Scene Cache Graph Management)：ISO/IEC 14496-1

具體規定大量種類的節點及資料型態。為了用一個圖形資料結構(即場景圖 ，Scene Graph)來管理這些節點，於是採用一個一般化的形式來呈現每一個節點。為了加速場景圖的繪製，每一個節點皆會歸屬於一個相對應的高速儲存裝置繪製資源(cache render resource)，事先計算好的資訊便置於其內。而這些高速儲存裝置繪製資源會組成另一個圖，稱之場景儲存圖(Scene Cache Graph)。只有當此圖中某些欄位的值被改變時，一個節點才需要被重新計算其相關的資訊，並且更新其相對應的儲存資源。用此方式，可省下大量的計算機資源。

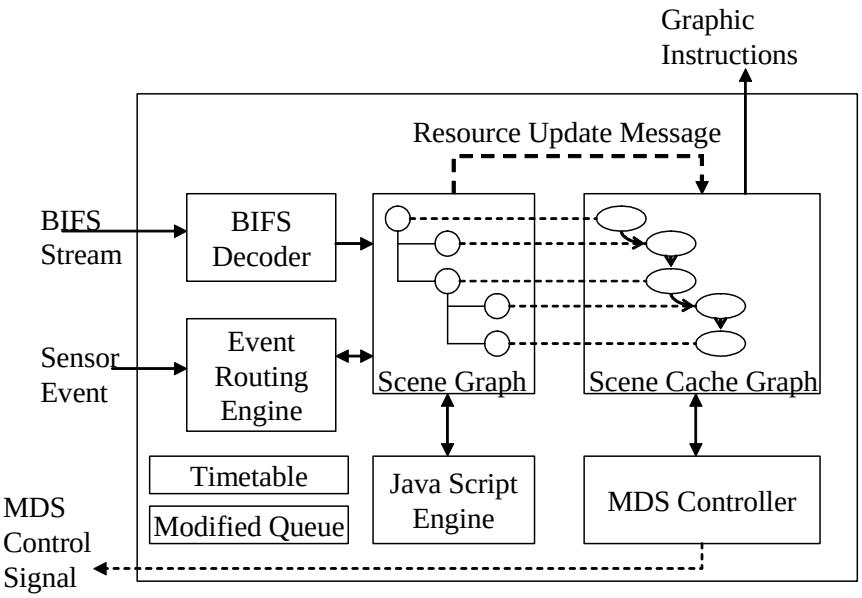


圖 8: 場景圖管理架構

在實驗結果部分中，我們將探討場景儲存圖形管理(SCGM ，Scene Cache Graph Management)的效能。

媒體解碼串流（Media Decoding Stream ,MDS）控制器：媒體解碼串流控制器是用來控制媒體解碼架構，其從解碼的媒體中，建構出需要的媒體解碼串流到合成緩衝器。藉由連接適當的媒體存取模組（MAM）和媒體解碼管線（MDP），媒體解碼串流（MDS）被建造出來。媒體解碼串流控制器也會個別驅動在媒體存取模組的擷取執行緒和媒體解碼管線中的解碼執行緒。

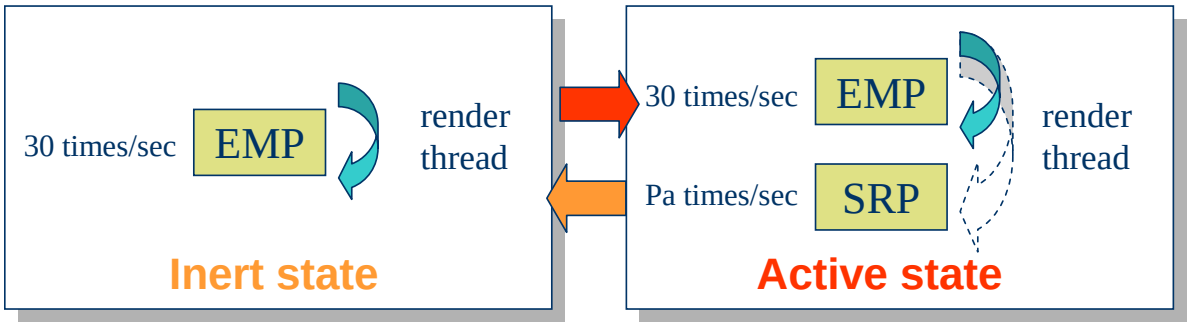
乙、實作與最佳化：

在我們的系統中，有兩個繪製執行緒。一個是負責繪製視覺的畫面；另一個則是負責呈現聲音的串流。視覺繪製執行緒負責驅動場景圖管理者和圖形引擎。它花很大的部分在系統有限的計算機資源上，因此我們提出一個最佳化的機制來減少此執行緒的工作量。

Visual render thread	Procedure Type	Overhead	Execution Freq. (times / s)	Exec. Freq.
	EMP	Low	30	Fixed
	SRP	High	0~30	Adaptive

表 1: 視覺繪製執行緒分類表

視覺繪製執行緒的工作被分成兩個程序，事件監督程序(EMP ，Event Monitoring Procedure)及場景繪製程序(SRP ，Scene Render Procedure)，如表 1所示。EMP 會偵測、發送、處理即時的媒體，如事件。SRP 會追蹤場景圖，並驅動圖形引擎來繪製 2D/3D 的圖形。EMP 的負擔是相當少的，而且執行的頻率決定系統事件的反應時間，因此 EMP 的執行頻率是固定的。但是 SRP 則有很大的負擔，而且其執行頻率決定場景的畫面更新速度，因此我們想要儘可能的降低 SRP 的執行頻率，並且同時要能夠維持畫面的品質。為解決這個問題，我們提出一個合適的畫面速率（AFR ,Adaptive Frame Rate）機制。



在此機制中，一個執行中的場景不是處於積極狀態（active state），就是處於遲緩狀態（inert state）。當至少有一個MovieTexturec或TimeSensor正在執行時，系統即處於積極狀態，否則系統即處於遲緩狀態。當系統處於遲緩狀態時，繪製執行緒採用緩慢繪製的策略，即只有週期性的執行EMP。換句話說，系統只有在某些事件發生，或者場景某處被修改時，繪製執行緒才會執行SRP。當場景處於積極狀態時，為了不停地繪製MovieTexture，繪製執行緒會在一個特定的週期$\mathbf{P_a}$，執行SRP，除此之外，繪製執行緒仍會執行EMP。因此，我們提出一個演算法來估計出合適的週期$\mathbf{P_a}$：

Definition: $\mathbf{F} = \{f_1, f_2 \ldots f_n \mid f_i$: the frame rate of MovieTexture i$\}$

Input: $\mathbf{F}$ Output: $\mathbf{P_a}$

```
If "Fa not init" or "F changed"
  Fa = Min( Median(F), MAX-FPS );

If (Average-CPU-Utilization > threshold)
  Fa *= FPS-DECREASE-FACTOR;
Else
  Fa += FPS-INCREASE-FACTOR;
Pa = 1/Fa;
```

丙、 實驗及成果

Scene 1

- Experimental Scene: Scene 1
- Adaptive Frame Rate (AFR): Off
- Scene Frame Rate: fixed at 12 fps

Scene Cache Graph Management (SCGM)	CPU load
Without SCGM	98.65%
With SCGM	6.37%

Scene 2

- Experimental Scene: Scene 1 & Scene 2
- Source video format: MPEG-4 simple profile & QCIF & 15fps
- Measurement Parameter: lifetime of battery (100% → 20%)
- AFR: On SCGM: On

Experimental Scene	Gains compared with “AFR:Off”
Scene 1	118%
Scene 2	27%

圖 9: 實驗結果

我們在 Toshiba e740 Pocket PC 上開發本系統，該平台採用 Intel XScale PXA 250 CPU，並且內建 802.11b 無線模組。圖呈現了兩個 MPEG-4 場景。以下說明 SCGM 在場景一的效能測試結果。當重繪圖框速率固定在 12 fps 的情況下，不使用 SCGM 則 CPU 負載為 98.65%。啟動 SCGM 則改善到 6.37%。換句話說，節省了約 15 倍的計算資源。

場景二則包含了三個物件，每一個物件上都各自結合了電影的表面材質。場景中央的橢圓形物件設定了透明度 0.3，並隨機的在場景中移動。播放的影片皆為 QCIF，且其圖框播放率為 15 fps。三段影片中其中兩段影片來源是由遠端是伺服器及時串流傳輸，第三部影片來源則是存放在裝置上的檔案。使用 AFR 機制時，場景的最高圖框播放率約達 14 fps。

我們計算電池容量從 100%變化到 20%的壽命期間來測量 AFR 的效能。實驗結果顯示在播放第一個場景時電池的壽命期間增長了 118%。在相同的情況下播放第二個場景則只改善了 27%。AFR 在播放第一個場景時效能較佳的原因是由於場景一大部分都處於插入狀態，而 AFR 在插入狀態下效能較好。

【MPEG-4 互動電視平台】

1. 前言與研究目的

為了快速提供較先進的電視服務，各網路服務廠商在初期快速且積極地投入發展具備特色的互動電視平台；其中中介軟體扮演著最重要的角色，它的主要目的是將應用程式與硬體與網路的架構區分開來，以減少開發應用程式時的複雜度、提高開發速度及增加應用程式的共通性。

中介軟體的功能應包括對資料的製作、展現、傳送及應用程式界面..等的規範。目前市場上傾向延用既有的網際網路技術如 TCP/IP、Web、HTML、JavaScript、DOM、CSS..等。讓電視網頁瀏覽與電視互動有著共同的使用者界面。最重要的是讓有 WEB 經驗的技術人員可輕易地進入互動電視的市場、以加速互動電視應用的普及度。

內嵌在 MPEG-4 場景中的 JavaScript 原本是專為撰寫網頁設計的語言，而非一般用途(general purpose)的程式語言，因為它在模組化及物件導向方面功能的不足，使得要使用 JavaScript 不適合開發較大規模的程式。此外 JavaScript 只有比較少的一般用途的標準函式庫可用，有許多功能都不容易找到現成的函式庫來使用。因為函式庫較少，JavaScript 似乎也少有直接支援使用 Socket 等方法來傳送資料的功能，所以要單純以 JavaScript 來開發一般的網路應用程式也很不方便。另外，因為原本的方式是將 JavaScript 程式嵌在場景中，更使得這些程式難以達到模組化的設計及重覆使用。

2. 研究方法

為了要開發一個以 MPEG-4 為基礎的互動電視平台，我們必須開發下列技術模組：

甲、應用程式引擎

MPEG-PY Interface：MPEG-PY System 的核心在其中的 MPEG-PY Interface，MPEG-PY Interface 的目的是提供一套 Python 的介面以控制 MPEG-4 System，MPEG-PY Interface 的設計，主要是參考 MPEG-4 Standard Part1:System 中的 MPEG-J。MPEG-J 提供一套以 Java 來控制 MPEG-4 System 的介面。MPGE-PY Interface 目前提供的介面主要分為兩個部份：Scene API 以及 Rendering API，Scene API 用來控制 MPEG-4 System 中場景的內容，Rendering API 則用來設定 MPEG-4 System 如何畫出場景。下面有一節會再詳細介紹 Scene API 的部份。

乙、伺服端可程式化環境

互動電視服務的一個重要的特色，是能讓使用者與伺服器之間的高度互動性，也就是說，互動電視媒體伺服器所提供的場景、多媒體串流或應用程式，能夠依使用者的要求動態地調整其內容。舉例說明，使用者把自己所喜歡的影片類型，傳送給隨選視訊系統之伺服器，伺服器便可依照此一資訊篩選出符合使用者需求的影片，並動態地產生一個 MPEG-4 影片選單的場景，當此場景下載至用戶端，用戶端的 MPEG-4 場景播放器便

能合成出這個視覺化的選單場景，以供使用者點選。另一個例子則是透過互動電視之服務進行網路遊戲，由於任何型態的網路遊戲都會透過伺服端與用戶端的連線來傳遞遊戲所需之各類資訊，例如更新使用者最新的狀態，倘若這些工作都交由用戶端 Script 來處理，將會使得系統過於複雜並加重用戶端的負擔。因此，在 MPEG-4 伺服端亟需一個可程式化的架構，以滿足這些動態改變的需求。

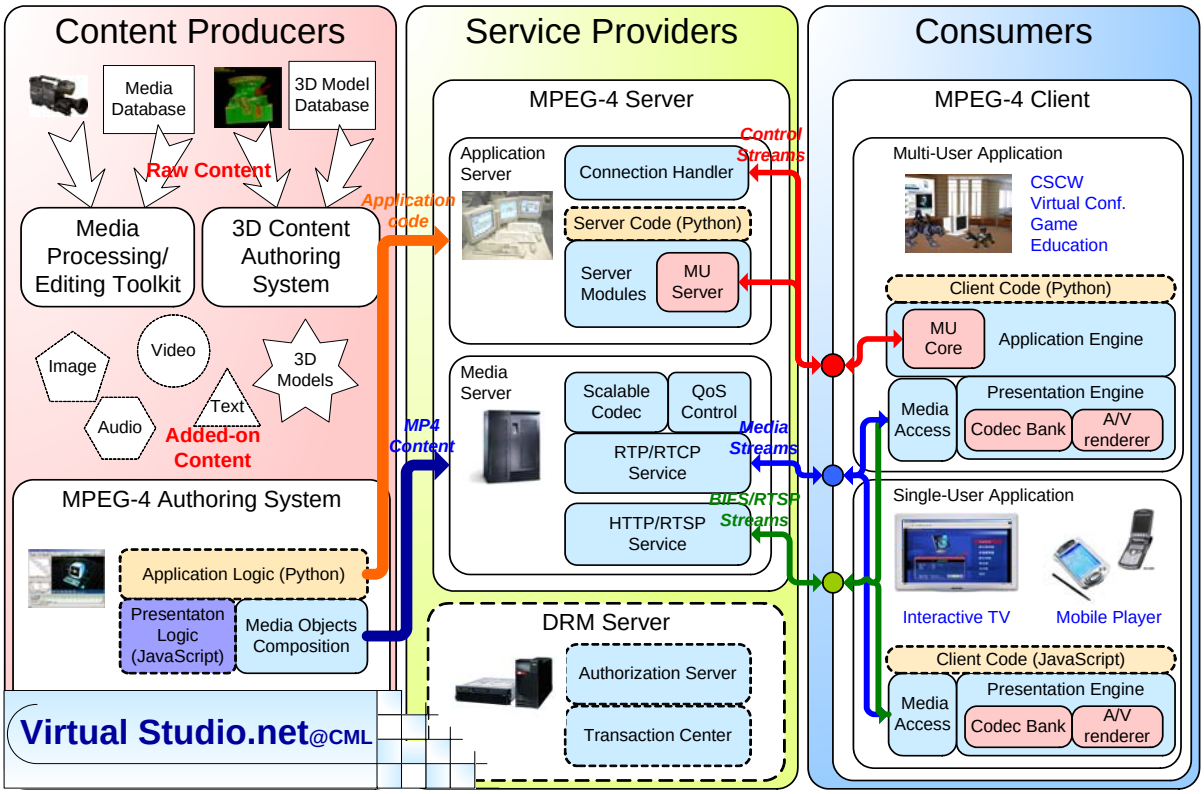


圖 10: MPEG-4 互動電視平台

丙、媒體伺服器

MPEG-4 影音串流伺服器支援了 RFC 一系列工業標準，包括 RTSP,SDP 以及 RTP。此外，在 MPEG-4 over IP 標準中定義了如何利用 RTP Packet 傳送 MPEG-4 Elementary Stream (ES)的方法。本系統將 MPEG-4 ES 所傳送的 Access Unit(AU)切割封裝在 RTP Payload 裡，再透過 RTP 模組傳送至接收端。RTP Packet 傳送至接收端之後，系統會將其還原解碼。

3. 實驗及成果

本實驗室承續過去在 MPEG-4 領域研發的成果，開發出 Virtual Studio.NET 互動電視平台，這是一個以人為中心的數位媒體互動平台，在這裡有 Media Archive 儲存了包括視訊與圖片的各式媒體、有 iTV 提供互動節目，還可以在 Meeting Room 中跟朋友進行虛擬會議或是在 Game Zone 裡娛樂一下。使用者可以利用各種不同的 Device 進入這個虛擬平台，包括了 PC、手機、PDA 甚至家中客廳的 TV。無論處在何時何地，Virtual

Studio.NET 都將帶來全新的體驗科技。

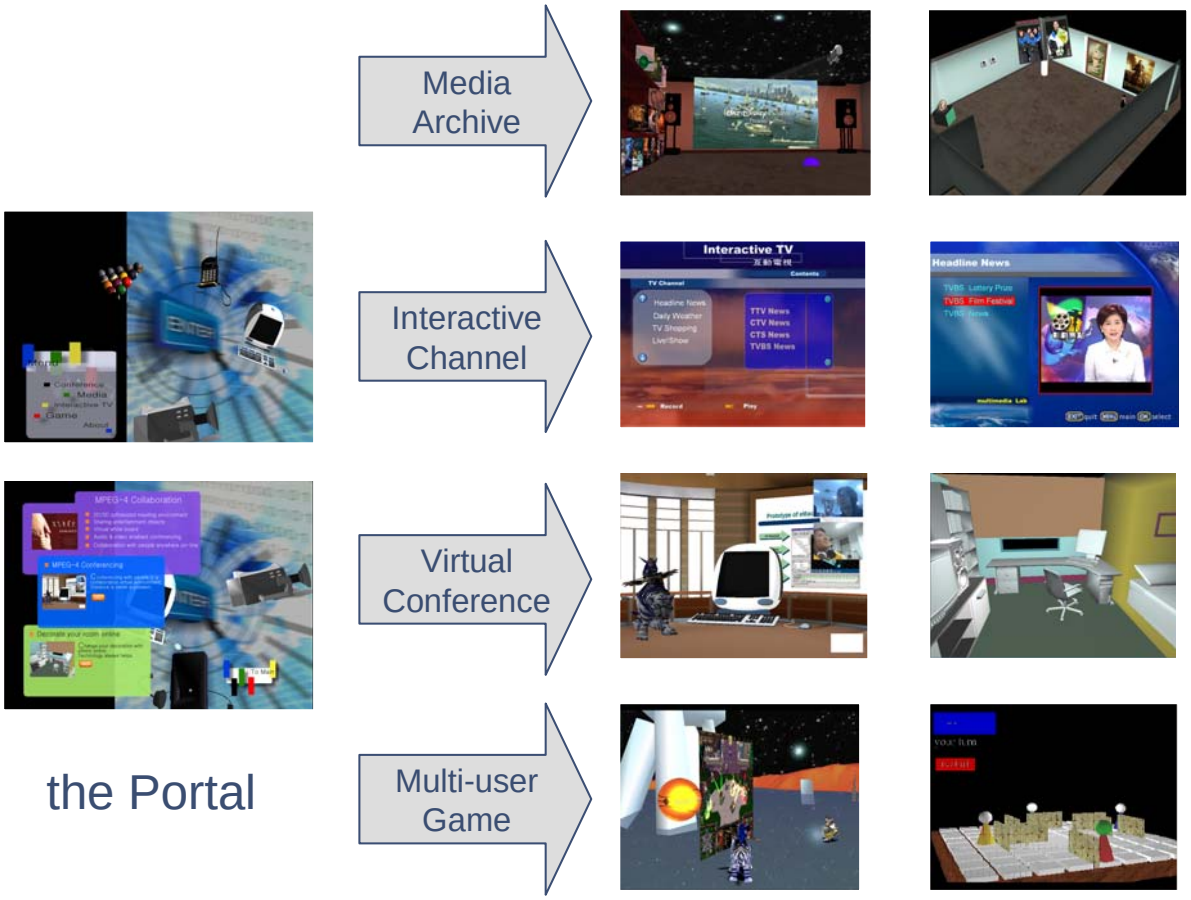


圖 11: Virtual Studio.NET 示意圖

Virtual Studio.NET 勾勒出一個以 MPEG-4 為核心的互動電視服務系統的平台，我們分別在 Windows、PocketPC、Linux 平台上實作了 MPEG-4 Presentation Engine，除了針對平台採用不同的繪圖與音效函式庫之外，核心模組均可重複使用，正符合 MPEG-4 的跨平台特性。各個系統模組的功能描述與實作成果如表 2：

系統模組	功能描述與實作成果
Presentation Engine	Windows, PocketPC, Linux
Application Engine	JavaScript Interpreter, MPEG-Python
Application Server	Python-based Server
Media Access Layer	AVI,ASF, MPEG-1,2, MP4 File, RTP Packet
Streaming Server	HTTP Server, RTSP/RTP Server

表 2: 系統模組的功能描述與實作成果

【第三年預期成果與具體成果之比較表】

預期工作項目	實際工作成果	說明
MPEG-4 互動電視的播放機制與架構	1.MPEG-4 多人世界應用伺服器 2.MPEG-4 多點傳輸串流伺服器 3.支援不同平台之 MPEG-4 播放器 4.伺服端與播放端之互動架構	符合。
與其他分項計畫整合，實作出完整功能的『互動資訊媒體服務』系統	互動資訊媒體服務系統 (Virtual Studio.NET)	符合。
融合業界最新技術，進行系統之錯誤更正與效能調整	完成。	符合。
撰寫技術手冊與使用手冊	完成。	符合。
落實研發成果至合作廠商	完成。	符合。

四、相關論文完成

[1] 以 MPEG-4 為核心之多媒體群體合作應用架構（“A Multimedia Collaboration Application Framework Based on MPEG-4 Standard” 黃奕勤，博士論文，2004）

[2] Linux 上的 MPEG-4 互動式多媒體播放器實作（“An Implementation of MPEG-4 Interactive Media Player on Linux” 楊書旻，碩士論文，2004）

[3] MPEG-4 應用程式開發系統的設計及實作（“Design and Implementation of an MPEG-4 Application Engine” 楊國鑫，碩士論文，2004）

五、國際期刊與國際會議論文

國際會議論文

[1] Yi-Chin Huang, Tu-Chin Yin, Kou-Shin Yang, Yan-Jun Chang, Meng-Jyi Shieh, Wen-Chin Chen, “Design and Implementation of an Efficient MPEG-4 Interactive Terminal on Embedded Devices,” TP3-4, (CD-ROM) Proc. of IEEE 2004 International Conference on Multimedia and Expo (ICME04), Taipei, Taiwan, June, 2004.

[2] Yi-Chin Huang, Meng-Jyi Shieh, Chien-Feng Huang, Ching-Che Kao, Shu-Min Yang, Wen-Chin Chen, “A Visual MPEG-4 Scene Editor,” PD1, (CD-ROM) Proc. of IEEE 2004 International Conference on Multimedia and Expo (ICME04), Taipei, Taiwan, June, 2004.

[3] Meng-Jyi Shieh, Chien-Feng Huang, Cha-Dong Duh, Wei-Teh Wang, Wen-Chin Chen, “Implementation of an efficient MPEG-4 2D/3D Mixed Renderer for Visual Editing and Interactive Applications,” International Conference on IEEE MPEG-4 3rd Workshop & Exhibition, San Jose, CA, U.S.A., June, 2002.

[4] Cha-Dong Duh, Jiann-Rong Wu, Chien-Fen Huang, Ching-Che Kao, Meng-Jyi Shieh, “Integrating a Real-time Interactive Virtual Character into MPEG-4,” International Conference on IEEE MPEG-4 3rd Workshop & Exhibition, San Jose, CA, U.S.A., June, 2002.

六、參考文獻

其餘參考文獻請參照分項計畫一計畫書參考文獻。

七、相關論文成果 (見後頁)

Design and Implementation of an Efficient MPEG-4 Interactive Terminal on Embedded Devices

Yi-Chin Huang, Tu-Chun Yin, Kou-Shin Yang, Yan-Jun Chang, Meng-Jyi Shieh, Wen-Chin Chen
Communication and Multimedia Lab.,
Department of Computer Science and Information Engineering,
National Taiwan University, Taipei, Taiwan

E-Mail: {yichin, birding, voyager, ckuan, feitian, wcchen}@cmlab.csie.ntu.edu.tw

ABSTRACT

We present an efficient MPEG-4-based interactive player for PDA-like embedded devices in this paper. Our system can receive media from various sources, then decompress and compose them in an object-oriented manner. Embedded devices such as PDA usually have limited computational resources, memory size, power budgets, and multimedia capabilities. To overcome these constraints, two novel mechanisms were introduced, namely *adaptive frame rate (AFR)* and *scene cache graph management (SCGM)*. Furthermore, a *media decoding framework* was proposed such that multimedia objects of different formats can be retrieved from different sources and then be decoded. This framework can be extended by add-on components. A *semi-pull model* was also designed for the synchronization of heterogeneous media objects. Finally, all the key modules were efficiently implemented and optimized. These modules include MPEG-4 video decoder, 2D/3D graphic engine, buffer management, script engine, and streaming service.

1. INTRODUCTION

With rapid hardware advance and the growth of broadband networks, MPEG-4 is becoming one of the most important international standards. Unlike the traditional frame-based multimedia standard, MPEG-4 adopts the object-oriented methodology, which integrates the existing multimedia technologies, such as still images, 2D/3D graphics, animations, video, audio, and virtual reality into one single architecture. MPEG-4 specifies a mechanism for describing the spatial and temporal relations among the different multimedia components of the application. This mechanism is called the "MPEG-4 Scene Description" in the standard, which is termed "BInary Format for Scenes" (BIFS). Furthermore, users can define the interactive behaviors and logical relationships among these media. Being object-based and programmatic, MPEG-4 will bring on a variety of new multimedia applications.

One of the essences of MPEG-4 is the interoperability of different platforms. With the popularity of embedded devices such as PDA, mobile devices will surely be the platforms for the emerging mobile multimedia applications. However, embedded devices are usually under the constraints of lower computing power and limited hardware supports. Therefore, there are many technical issues need to be tackled in designing and implementing an efficient MPEG-4 system on embedded devices. These issues are: 1) Lower computing power; 2) Less hardware acceleration; 3) Constrained memory size; and 4) Lack of built-in multimedia APIs. For the first three issues, some optimization efforts are needed to reduce the time and space complexity. Unacceptable responding time and heavy power consumption may occur if cares are not taken for these system resource constraints. For the last one issue, on embedded devices, there is usually no high-level multimedia APIs, e.g. DirectX, available to help programmers to access the advanced features of high-performance hardware such as 3D graphics acceleration chips and sound cards. Moreover, APIs for playing back multimedia streams, e.g. DirectShow, are too complex for embedded devices. Thus, we should design a framework for audio-visual streams decoding and synchronization.

This paper consists of six parts. Part 1 gives an introduction to MPEG-4 and the problems while designing MPEG-4 system on embedded devices. Part 2 describes the system architecture. Part 3 presents the synchronization mechanism for heterogeneous media. Part 4 presents some implementation and optimization details. Part 5 shows the experimental results of the system and the performance of the novel mechanisms we proposed. Finally, Part 6 gives future plans and concludes the paper.

2. SYSTEM ARCHITECTURE

2.1 The MPEG-4 Media Process Flow

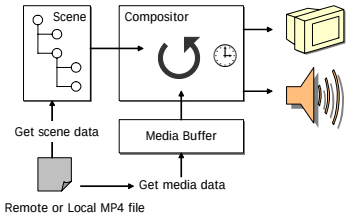


Figure 1. The processing flow of a MPEG-4 presentation

Figure 1 gives the general processing flow of an MPEG-4 media presentation. At first, the MPEG-4 media data are retrieved from local files or remote server. These data are decoded to constitute the scene description of the content. A scene graph is then built. Based on the scene graph, the media data are retrieved from different sources accordingly. These media streams are decoded into media buffer and are ready to be used by the compositor to present the whole MPEG-4 content.

There are some challenging issues for implementing a MPEG-4 scene player: First, a BIFS scene graph must be well-managed with related resources. Second, the system should support various media formats. Furthermore, media may come from different sources. The system must access not only local storages but also remote services in streaming fashion. Finally, the whole scene with all associated media must be properly presented in a synchronized manner. From the above discussion, our system is designed as shown in Figure 2. The system consists of several modules: Scene Graph Manager, Media Decoding Framework, Composition Buffer, Audio Engine, and 2D/3D Graphic Engine.

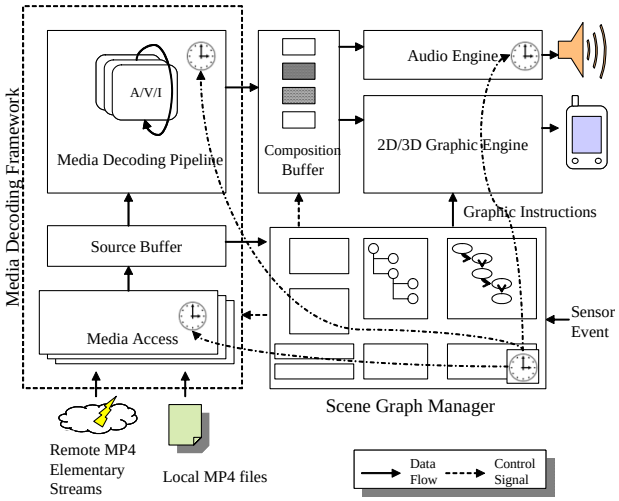


Figure 2. System Architecture

The function of each module is as follows. Scene Graph Manager, which is the kernel of the system, manages the operations of the scene graph. Media Decoding Framework is in charge of retrieving various media data and then decoding them. Composition Buffer stores the decoded data. Under the operation of semi-pull model, Scene Graph Manager synchronizes the presentation of heterogeneous media objects. In order to compose the whole scene, Scene Graph Manager also generates graphic instructions according to the information and 2D/3D spatial relationship of media objects defined by the scene description. According to these instructions, Graphic Engine then efficiently renders 2D/3D graphic primitives as well as textures. At the same time, Audio Engine renders decoded audio streams.

2.2 Media Decoding Framework

MPEG-4 supports media of various formats, from different sources and conveyed by different file formats or network protocols. To fulfill these requirements, a generic framework for decoding comprehensive media streams is proposed. In this framework, a **Media Decoding Stream** (MDS) is established by connecting the **Media Access Module** (MAM) and a **Media Decoding Pipeline** (MDP), as shown in Figure 2. Three generic input/output interfaces and Source Buffer, as the bridge of the connection, make MAM and MDP work independently. These three interfaces are the output interface of MAM and the input/output interfaces of MDP. Whenever a new component implementing these interfaces is plugged-in the system, the system is able to process new media data. This framework provides a feasible and extensible mechanism to control various supported media streams.

2.3 Scene Graph Manager

2.3.1 Scene (Cache) Graph Management

ISO/IEC 14496-1 specifies numerous kinds of nodes and data types. In order to manage these nodes with a graph data structure, namely **Scene Graph**, a general form is introduced to represent each node. To accelerate the rendering of the scene graph, each node is attached with a corresponding cache render resource, in which pre-computed information is cached. These cached render resources constitute another graph, namely **Scene Cache Graph**. Only when the values of some fields are changed, a node needs to recompute related information and update its corresponding cached resource. In this way, considerable amount of computational resource are saved. The performance of SCGM will be shown in Part 5.

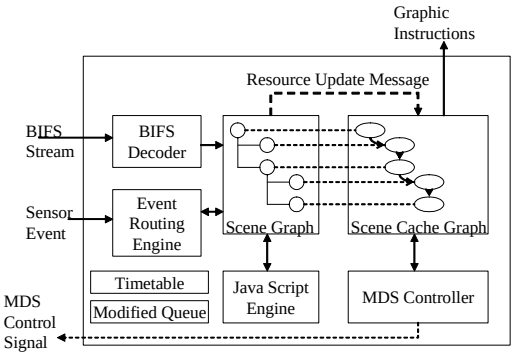


Figure 3. Architecture of Scene Graph Manger

2.3.2 MDS Controller

The MDS Controller, which is the controller of the Media Decoding Framework, constructs necessary MDSs for decoding media to the composition buffer. A MDS is constructed by connecting an appropriate media access module and a decoding pipeline. The MDS Controller also drives the data retrieving thread and decoding thread in Media Access and Media Decoding Pipeline modules, respectively.

3. SYNCHRONIZATION OF HETEROGENEOUS MEDIA

In this section, we introduce semi-pull model to solve the issues of intra-media synchronization for media of each type and inter-media synchronization between different media streams.

3.1 Categories of Media

Media objects supported by MPEG-4 can be classified into three categories according to their sampling time length: (1) **Instant Time Media** are the media data that can be computed immediately. They are usually processed in EMP (will be discussed in 4.1). For example, the event generated by a *TimeSensor* is such kind of media. (2) **Discrete Time Media** are the media whose sampling lengths are long enough so that computers can accomplish the processing. Most of these media are visually related, such as video and animation. For example, a video with frame rate 29.97 fps has media sampling time about 33 ms, which is a long time for the devices. (3) **Continual Time Media** are those media whose sampling time is very short. These kinds of media are usually auditory. For example, the sampling rate of an audio sound is 44100, its sample time is 0.000023 sec.

3.2 Semi-Pull Model

We propose semi-pull model to solve both intra-media and inter-media synchronization problems in a uniform way. In pull model, which is a conventional model to decode and render a media

stream, the render thread accomplishes all the tasks required to present a media stream, such as getting source data, decoding, and rendering. Therefore, the timing to render a decoded frame is controlled by the render thread. Pull model is an appropriate model for the simple case that only one video and audio stream is presented concurrently. However, for the more complicated case when more continual media objects in an MPEG-4 scene should be presented in a synchronized way, pull model does not apply.

In our semi-pull model, a new thread, decoding thread, is created for decoding. The decoding threads deliver decoded data through composition buffers, as shown in Figure 2. Each decoded data sample is associated with a composition timestamp, which indicates the time at which this sample should be rendered. A composition timestamp is created with respect to a global reference clock generated by scene graph manager. We will discuss how the semi-pull model achieves the heterogeneous media synchronization.

3.3 Intra-media Synchronization

The goal of intra-media synchronization is to maintain the on-time presentation of the data samples so that acceptable user perception at playback is ensured. Without intra-media synchronization, the presentation of the media may be interrupted by pauses or jitters. For instant time media, synchronization can be achieved if the EMP execution rate is kept high enough. For other media types, semi-pull model ensures that the presented data sample is the most updated and rendered on time. Thus, intra-media synchronization of each kind of media is accomplished.

3.4 Inter-media Synchronization

Inter-media synchronization can be categorized into two cases: one is for the synchronization between time-variant (discrete and continual) media and the other is for the synchronization between instant and time-variant media. A typical example of the first case is the lip-synchronization between video and audio streams. The first case can be resolved by forcing all media streams synchronized to the global clock. In this way, the synchronization can be accomplished, except for media streams that have their own reference clocks. Currently in MPEG-4, only audio streams have this problem. Audio streams are rendered by a sound card, which has a local clock. Inevitably, skew between the local and global clocks may be intolerable after a long-time playback. For this problem we have to synchronize the audio stream to the global clock by adjusting the amount of waveform data that is to be fed into frame buffer on the sound card. The algorithm adopted in our implementation is:

```
Initialize temp buffer
While more decoded waveform data {
    Copy decoded data to temp;
    TimeOffset =
      SystemAbsoluteTime – BufferPlayingTime;
    If( abs(TimeOffset) < Threshold )
       Smoothly drop or insert sample;
    Else {
       Reset audio device;
       Drop a segment of samples;
    }
    Output temp data to audio device buffer;
}
```

For the second case, the data of instant time media are processed when current frame is about to be rendered. Because of the high execution frequency of EMP, the delay between the time at which the data is created and processed is restricted within a tolerable time interval. Therefore instant time media can be synchronized to the global clock, which is also referenced by time-variant media. The inter-media synchronization is achieved.

4. IMPLEMENTATION AND OPTIMIZATION

4.1 Adaptive Frame Rate Mechanism

There are two render threads in our system. One is for rendering visual frames; the other is for rendering audio streams. The visual render thread is responsible for driving scene graph manager and graphic engine. It takes a large part of the limited computational resources of the system. Thus, an optimization mechanism is proposed to reduce the workload of this thread.

	Procedure Type	Overhead	Execution Freq. (times / s)	Exec. Freq.
Visual render thread	EMP	Low	30	Fixed
	SRP	High	0~30	Adaptive

Table 1. The 2 procedures of visual render thread

The job of visual render thread are divided into two procedures (see Table 1.), Event Monitoring Procedure (EMP) and Scene Render Procedure (SRP). The first one detects, routes, and handles instant time media, such as events. The second one traverse scene cache graph and drives graphic engine to render 2D/3D graphics. The overhead of EMP is quite low and the execution frequency determines the response time for incoming events of the system. Hence, the execution frequency of EMP is fixed. SRP has much higher overhead and its execution frequency determines frame rate of the scene. Thus, we have to make the execution frequency of SRP as low as possible, while preserving the visual quality. To solve this problem, the **Adaptive Frame Rate** (AFR) mechanism is proposed. In this mechanism, a scene runs in either *active state* or *inert state*. The scene runs in active state when at least one MovieTexture or TimeSensor is active. Otherwise it is in inert state. When the scene runs in inert state, render thread adopts lazy-render strategy, with which the only periodic routine task is to execute EMP. In other words, render thread executes SRP only when some events occur and somewhere in the scene has been modified. When the scene runs in active state, to continually render the active movie textures, render thread executes SRP in every specified period $\mathbf{P_a}$, in addition to what it does in inert state. We propose an algorithm to adaptively estimate the appropriate period $\mathbf{P_a}$.

Definition: $\mathbf{F} = \{f_1, f_2 \ldots f_n \mid f_i$: the frame rate of MovieTexture i$\}$

Input: $\mathbf{F}$ Output: $\mathbf{P_a}$

```
If "Fa not init" or "F changed"
  Fa = Min( Median(F), MAX-FPS );
If (Average-CPU-Utilization > threshold)
  Fa *= FPS-DECREASE-FACTOR;
Else
  Fa += FPS-INCREASE-FACTOR;
Pa = 1/Fa;
```

More computational resource can even be saved. If the determined scene frame rate 1/Pa is lower than video frame rate, some frames will be dropped by the frame-dropping filter embedded in the video decoder.

4.2 2D/3D Graphic Engine

The engine supports primitive 2D/3D rendering functionalities comprising transformation, frustum clipping, viewport, lights, and optimization of texture processing. Two essential hardware supports are not present on the target platform, including

hardware acceleration for 3D rendering and floating-point number processing unit. To overcome these problems, the engine applies fixed-point arithmetic when decimal arithmetic is required and performs all 3D rendering functionalities by software emulation.

4.3 MPEG-4 Video Decoder Optimization

4.3.1 Optimized Inverse DCT Algorithm

A large part of IDCT calculation needs floating-point operations. XScale CPU doesn't have a dedicated floating-point unit, and all the floating-point operations are implemented by software simulation, which is not efficient. Thus, we have to implement the IDCT module with only integer operations. Our implementation also adopts Feig.'s inverse DCT Algorithm [2], which significantly reduces the amount of multiplicative and additive operations. In addition, de-quantization operation is also combined with IDCT module. The most multiplications are further eliminated with this combination. Moreover, unnecessary memory accesses on temporary buffers that store dequantized data before IDCT are reduced.

4.3.2 Optimization of Motion Compensation

In the recommended decoding procedure, the predictive error for motion prediction is derived from IDCT and written in a temporary buffer. The error values are then read and added with the result of motion compensation in another buffer. In our implementation, motion compensation is directly done upon the result of IDCT in the same buffer. A temporary buffer is thus eliminated and the number of memory access is reduced.

4.3.3 Optimization of Color Conversion

The color conversion from YCbCr to RGB is optimized by table-lookup. Each floating-point operation is replaced by a single memory addressing.

4.4 Buffer Management

This module maintains two buffers to hold decoded data. While a decoding thread writes data to one buffer, the render thread reads data from the other one. To avoid the race condition, both threads may lock the buffers. When the decoding thread finishes writing to the buffer, it locks and toggles the two buffers. When the render thread reads data from the buffer, it also issues a lock. Since it takes very short time for the toggling and reading operations, the potential waiting time can be minimized.

4.5 Script Engine

To enhance the interactivity of BIFS scene, the system supports JavaScript mechanism specified in MPEG-4. The script engine parses all the script codes in the scene and transforms them into a syntax tree for speeding up the script execution. When an event is routed to a Script node, the corresponding function is evoked.

4.6 Video Streaming

A media access module is implemented for MPEG-4 video streaming over RTP. The module maintains the RTP/RTCP session of an MPEG-4 stream, feedbacks the network status, manages the buffer, controls transmission errors and composes the fragmented data during transmission. The module adopts forward error correction (FEC) to recover data when a packet is lost. At the sender side, the XOR operator is used to generate a redundant packet among a group of packets. At the receiver side, if a single packet is lost, the lost packet can be recovered by applying XOR operation on those correctly received packets.

5. EXPERIMENTAL RESULTS

Figure 4. Snapshots of MPEG-4 content

We developed the system on Toshiba e740 Pocket PC with Intel XScale PXA 250 CPU and built-in 802.11b wireless module.

Figure 4 shows two snapshots of MPEG-4 content. In the left one, the screen of the "monitor" (the 3D model placed in the middle) binds an active movie texture. The experimental results of the performance of SCGM for this scene are described as follows. When the rendering frame rate is fixed at 12 fps, the average CPU load is 98.65% without SCGM. When SCGM is enabled, the CPU load is improved toward 6.37%. In other words, roughly 15 times of the computational resource is saved.

In the right of Figure 4, there are three objects, each of which is bound with a movie texture. The elliptical object in the middle of the scene is set with 0.3 transparency and randomly moving around in the scene. These videos are in QCIF and frame rate 15 fps. Two of them are streamed from remote server, while the third one is retrieved from local file. With AFR mechanism, the highest frame rate of the scene is about 14 fps.

To measure the performance of AFR, we consider the lifetime of the battery with power consumption varies from 100% to 20%. The experimental result shows that AFR gains 118% improvement of the lifetime in the first scene. Under the same condition, only 27% improvement of the lifetime is gained in the second scene. The reason that AFR performs better in the first scene is that most part of it is in inert state, while AFR works more efficiently in that state.

6. CONCLUSIONS

We have successfully developed a MPEG-4 interactive player on a resource constrained PDA. Our system adopts four novel mechanisms, namely adaptive frame rate, scene cache graph management, media decoding framework and semi-pull model for resolving the problems including computational resource reduction, low power consumption, comprehensive media format support and heterogeneous media synchronization. For further research we will improve the efficiency of our system and develop new functionalities such as MPEG-4 Multi User Worlds.

7. REFERENCES

[1] ISO/IEC 14496 AM 1, Part 1: Systems, Part 2: Visual, Part 3: Audio, Part 6: DMIF, International Organization for Standardization, 2000.

[2] E. Feig, S. Winograd, "Fast Algorithms for the Discrete Cosine Transform," IEEE Trans. on Signal processing, vol.40, no. 9, pp. 2174-2193, 1992.

[3] RFC 3016, Y. Kikuchi, T. Nomura, S. Fukunaga, Y. Matsui, and H. Kimata., "RTP Payload Format for MPEG-4 Audio/Visual Streams", November 2000

[4] Meng-Jyi Shieh, Chien-Feng Huang, Cha-Dong Duh, Wei-Teh Wang, Wen-Chin Chen, "Implementation of an efficient MPEG-4 2D/3D Mixed Renderer for Visual Editing and Interactive Applications", 3rd MPEG-4 Workshop and Exhibition.

A Visual MPEG-4 Scene Editor

Yi-Chin Huang, Meng-Jyi Shieh, Chien-Feng Huang, Ching-Che Kao*, Shu-Min Yang, Wen-Chin Chen*
Communication and Multimedia Lab.,
Department of Computer Science and Information Engineering,
National Taiwan University, Taipei, Taiwan
Digimax Inc.*
*E-Mail: {yichin, feitian, sasquatch, wcchen}@cmlab.csie.ntu.edu.tw, {cardy, cckao}@ms.digimax.com.tw**

ABSTRACT

The Communication and Multimedia Laboratory (CML) has developed the MPEG-4 editor for 5 years since 1998. In 2001, Digimax Production Ltd. joined this project and addressed ourselves to make this system more complete. We have implemented a visual MPEG-4 scene editor for creating 2D/3D mixed scene, with features of event routing mechanism, visual editing, and friendly user interface. With our system, one can produce highly interactive MPEG-4 content easily.

1. INTRODUCTION

To meet the emerging multimedia applications as well as the rapid hardware advance such as CPU power, memory size, wire and wireless network bandwidth, the MPEG committee developed the MPEG-4 standard aiming at bringing various multimedia applications onto all kinds of software/hardware platforms. An MPEG-4 compliant application may contain in one single content several different objects such as text, pictures, audio, video clips, vector-based graphics, 2D/3D meshes, virtual reality, and so forth. These are the multimedia forms we are familiar with. However, the lack of a standardized framework to integrate these various multimedia formats makes it difficult to develop highly-interactive multimedia applications. Furthermore, existing multimedia systems cannot be migrated among multiple software/hardware platforms, making it even harder to develop a general multimedia application for embedded systems.

The MPEG-4 standard is specified by ISO/IEC 14496. It builds upon three application areas: digital television, interactive graphics applications and interactive multimedia. It aims at establishing a universal, efficient coding of different forms of audio-visual data, which we call audio-visual objects. Moreover, it tries to offer a new natural and synthetic interactive multimedia. This goal will be attained by defining two basic elements:

1. A set of coding tools for audio-visual objects supporting different functionalities such as object-based interactivity and scalability, efficient compression, and error robustness;
2. A syntactic description of coded audio-visual objects, providing a formal method for describing the coded representation of these objects and the methods used to code them.

Unlike the traditional frame based multimedia standards, MPEG-4 introduced the concept of "object-based" multimedia architecture, which enables the integration of various multimedia components. To realize sophisticated multimedia applications, MPEG-4 specifies a mechanism for describing the spatial and temporal relations among the different multimedia components of the applications. This mechanism is called the "MPEG-4 Scene Description." MPEG-4 Scene Description allows users to describe not only the spatial and temporal relations among various media, but also the interactive behavior and the logical relationships. It is the key that makes MPEG-4 suitable in various multimedia systems.

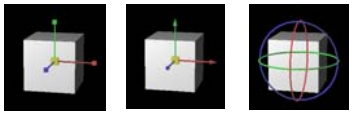
However, creating MPEG-4 Scene Description is not intuitive, but instead, is a very complicated and tedious process. The task of developing a large scale MPEG-4 application would be thus quite difficult. Therefore, in 1998, the Communication and Multimedia Laboratory (CML) of National Taiwan University began to develop the MPEG-4 Scene Editor. In September, 2001, Digimax Product Ltd. joined the project and introduced into the scene editor some fancy functions that fulfill the professional designers' requirements. To demonstrate the power of the editor, several large-scale and sophisticated MPEG-4 multimedia applications were developed.

In this demonstration, we present the MPEG-4 Scene Editor that allows users to visually edit a 2D/3D MPEG-4 scene. This article is structured as follows. Section 2 lists the features and functionalities to facilitate the editing process. Section 3 presents the system overview and several important components. The forth section describes the some MPEG-4 applications created by using our systems.

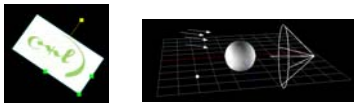
2. FEATURES AND FUNCTIONALITES

Our system has the following features:

- "What You See Is What You Get (WYSIWYG)" visualized editing;
- A variety of visualized controls: (a)Translation, (b) Rotation, (c)Scaling, (d) 2D control, and (e) Lighting Controls

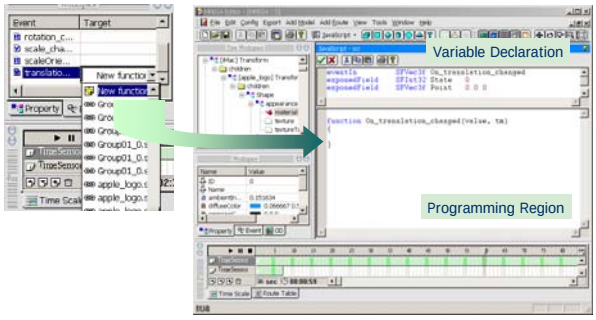


(a) (b) (c)



(d) (e)

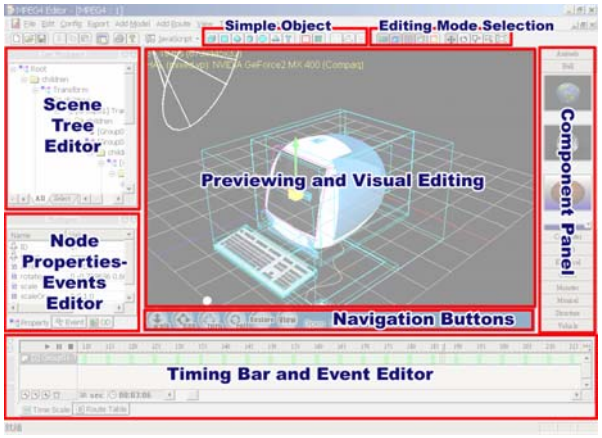
- Capability of using mouse to control the objects in the preview window;
- Simple user interface for various MPEG-4 functionalities;
- MPEG-4 File Format Exporting Wizard;
- JavaScript in PME (Property-Method-Event) style.



Our system supports most main MPEG-4 functionalities, including:

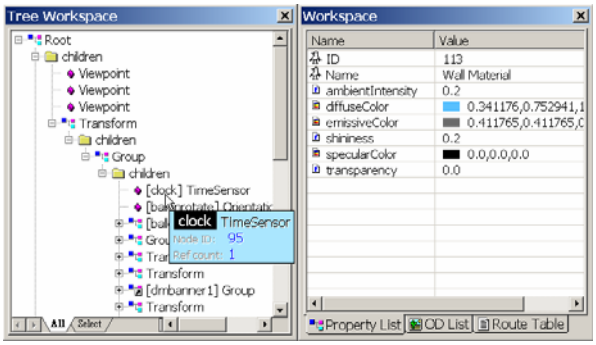
- Nodes defined in MPEG-4 specification version 2;
- Event Routing, ECMA-JavaScript ;
- BIFS-Command, BIFS-Animation, BIFS-Audio.

3. SYSTEM OVERVIEW

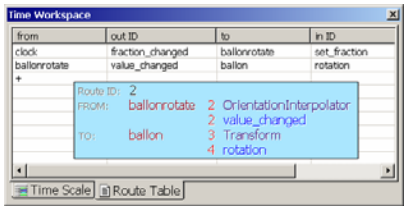


From the system operational aspect, the Scene Editor consists of six major components as follows:

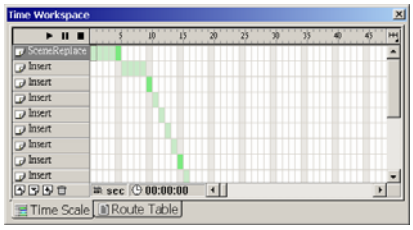
- Scene Tree Editor ;



- Node properties editor;
- Event (routing) editor;



- BIFS-Commands editor;



- Previewing and visual editing window;
- Component panel.

4. RESULTS

We implemented our system in Windows 2000 with VC++ 6.0, BCG UI library, and Direct3D/OpenGL library. The followings give some snapshots of MPEG-4 applications created by using our system. The first three pictures is a MPEG-4 game. This game mixed 2D and 3D within a single scene, and using JavaScript to control the logical interaction. The lower right one is a MTV content and contains some video clips and 2D pictures.

5. REFERENCE

[1] Meng-Jyi Shieh, Kuo-Luen Perng, Wen-Chin Chen, “The Design and Implementation of a Visual MPEG-4 Scene-Authoring Tool”, 2nd MPEG-4 Workshop and Exhibition.

[2] Meng-Jyi Shieh, Chien-Feng Huang, Cha-Dong Duh, Wei-Teh Wang, Wen-Chin Chen, “Implementation of an efficient MPEG-4 2D/3D Mixed Renderer for Visual Editing and Interactive Applications”, 3rd MPEG-4 Workshop and Exhibition.

Implementation of an efficient MPEG-4 2D/3D Mixed Renderer for Visual Editing and Interactive Applications

Meng-Jyi Shieh, Chien-Feng Huang, Cha-Dong Duh*, Wei-Teh Wang, Wen-Chin Chen*
Communication and Multimedia Lab.,
Department of Computer Science and Information Engineering,
National Taiwan University, Taipei, Taiwan
Digimax Production Ltd.*
*E-Mail: {feitian, bearw, wcchen}@cmlab.csie.ntu.edu.tw, {cardy, edgar}@ms_digimax.com.tw**

ABSTRACT

To push MPEG-4 to be used widely in interactive applications, such as games or interactive TV, an efficient MPEG-4 2D/3D mixed renderer is needed. In this paper, the architecture and some mechanisms of our implementation are introduced for making the MPEG-4 renderer more efficient and correct in playing and editing mode. Those mechanisms include the render resource management, 2-pass scene tree traversal, and video-texture optimization. Some MPEG-4 applications are also created to prove such architecture is workable.

1. INTRODUCTION

MPEG-4 [1] is the succeeding standard of MPEG-1 and MPEG-2 video standards. ISO standard committee designed this standard in 1998. Instead of using the current frame-based video technology, it adopts the object-oriented concept, which integrates the existing multimedia technologies, such as 2D/3D graphic, animation, video codec, multimedia streaming, interactive, and programmatic environment into single architecture.

The MPEG-4 standard is specified by ISO/IEC 14496. It is built upon three application areas: digital television, interactive graphics applications and interactive multimedia. It aims at establishing a universal, efficient coding of different forms of audio-visual data, which we call audio-visual objects. Moreover, it tries to offer a new natural and synthetic interactive multimedia.

To build an interactive scene at the terminal, it needs not only the audio and visual media but also additional information in order to combine these media into a presented scene. This information, called scene description, determines the placement of audio-visual objects in space and time. The scene description is represented using a parametric approach (BIFS - Binary Format for Scenes). The description consists of an encoded hierarchy (tree) of nodes with attributes and other event information. BIFS is based on the Scene-Graph concepts employed in VRML [3] and adds features such as Facial Animation, Enhanced Audio, native 2D primitives, Streaming and Animation streaming protocols.

In order to allow users to interact with the presented scene and make the scene dynamic and animated, MPEG-4 provides some interactive mechanisms, which are Event-Route, BIFS-Command, BIFS-Anim, and JavaScript mechanisms. Using those mechanisms, many interactive applications can be realized.

Interactive applications sometimes are not efficient enough because too many events might modify the properties of nodes unexpectedly, such as user actions, BIFS mechanism, or the JavaScript engine. That makes the resource management becomes more difficult. Similarly, a content author also modifies various properties of nodes more often in a MPEG-4 authoring tool. However, the performance of the renderer would be pretty low if it always reconstructs the needed resource. Therefore, an MPEG-4 renderer requires some acceleration mechanisms to meet the need of interactive applications, such as 3D game applications.

The basic concept to improve the performance of a renderer is based on the following principles:

1. To reduce the computation in each rendering frame.
2. To decrease the data transfer between system memory and graphic adapter.
3. To use the supported functionalities of the hardware accelerator as possible.

In this paper, we present the design and architecture of our implementation to meet the above principles, and we also show some MPEG-4 interactive applications at the end of the paper.

2. THE RENDERER ARCHITECTURE

Hundreds of nodes are defined in ISO/IEC 14496-1, and some nodes will be finalized in the following version. Thus, an ideal MPEG-4 rendering kernel should be flexible and extendable to handle increasing node definitions. Moreover, not only Built-in nodes but also **PROTO** nodes and **Script** nodes should be processed in the renderer. In general, content authors could declare user-defined fields within **PROTO** nodes and **Script** nodes. Therefore, a general form in our rendering module represents each node. This general form is composed of a node type ID and some fields that are defined in ISO/IEC 14496-1. To accelerate each node process, the renderer attaches the different render resource to each node according to its node type (as shown in Figure 1), and setups the resource by querying the value of each field.

As shown in Figure 2, the renderer holds all the render resources and traverses them to render the whole scene. To keep the render resources and the BIFS scene tree coincident, the BIFS scene tree sends the render resource messages (this will be mentioned in Chapter 4) when the properties of a node is modified by

JavaScript engine, Event Route, BIFS-Anim, and BIFS-Command. The renderer contains a reference clock to synchronize the media decoder module. By referencing this clock, the decoder module is required to decode the media and transfer into the composition memory in time, which will be acquired by the renderer. Thereby, the renderer can get the needed textures and audio raw data in next rendering frame.

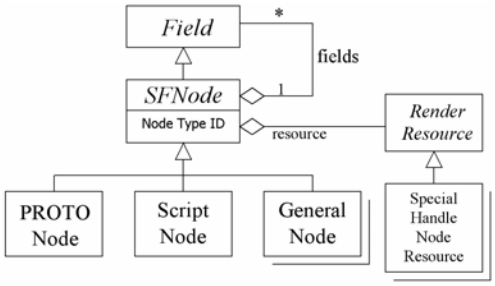


Figure 1: The abstract node class, SFNode, holds some fields and its type ID. The renderer attaches a special render resource according to the type ID.

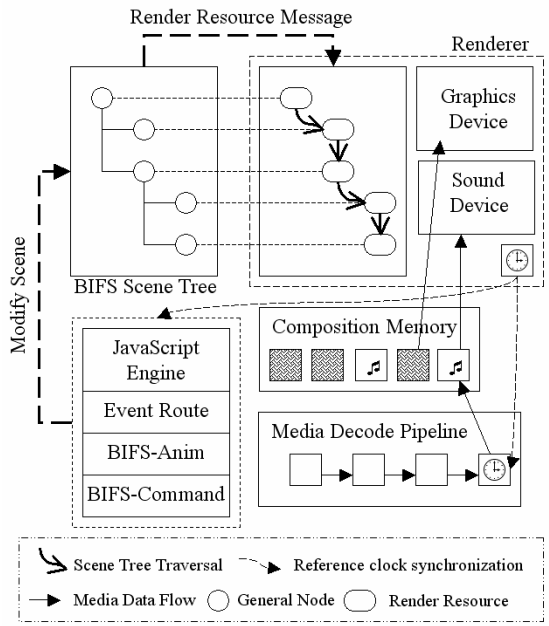


Figure 2: The renderer architecture.

3. THE RENDERING PROCESS

3.1 Rendering Pipeline

The rendering pipeline is shown in following, and we will explain some details in next section.

```
Traverse the scene tree for collecting following information:
    1. The transformation of each node
    2. The bounding box of each geometric object
For each layer
{
    1. Set the viewpoint and lights according to their ancestor transformation nodes
    2. Bind the active bindable nodes
    3. Set the clipping region by the size of the layer
    4. Clear the Z-Buffer
    5. Traverse the sub-tree within the layer
       For each geometric object
       {
          If the bounding box of the geometric object is in the visible region of the active viewpoint
          {
           If the geometry is transparency
                Push this one to post-render list
           Else
                Render it
          }
       }
    6. Sort the geometric object in the post-render list by the distance from the viewpoint, and then render them
}
```

3.2 2-Pass Scene Traversal

A scene may include some **Layer3D** and **Layer2D** nodes that indicate some transparent, rectangular rendering region where a 2D/3D scene is drawn. A layer node may bind some bindable nodes, such as **Viewpoint** and **Background**. The renderer should render each layer separately and combine each layer at a single frame.

We find that a single traversal of a BIFS-tree might encounter some problems:

1. Some special nodes, such as **Viewpoint**, **PointLight** and **Fog** might appear in middle of a BIFS-tree and can be influenced by its ancestor transformation. However, these nodes have to be processed before rendering other objects.
2. Transparent geometric objects, such as objects with an alpha texture or transparent material, must be rendered after all other non-transparent objects. And those transparent objects must be rendered according to their distance from the active viewpoint.

Therefore, it is hard to render the scene correctly just in one single traversal. Our implementation uses two-pass process. In the first pass, the renderer computes the transformation of each node and collects the special nodes. In the second pass, the renderer renders the geometric objects by traversing the scene tree again. The transparent geometries will be pushed to a list and render them after all other non-transparent geometries being rendered.

3.3 Frustum Clipping and Space Partition

There are two possible ways to eliminate objects overdrawing, which lie totally outside the viewport or be occluded by other object. The first approach is view frustum clipping, object lies outside viewport will be clipped by using the bounding box

information. The second approach is space partitioning, which requires preprocessing. We have to calculate a PVS (potential visibility set) table of objects and store those data in a **Script** node, and write some script code to handle the visible states of the geometric objects with some **Switch** nodes.

3.4 Editing Supporting

The renderer can also be embedded into a scene-authoring tool as a preview window. So in the editing mode, the renderer will show the auxiliary controls to facilitate the objective of visual editing. However, geometric objects will hide the part of the controls if the objects and the controls are in the same 2D/3D coordinate system. And the content author would not be able to click and select the controls conveniently. In our implementation, the renderer shows the auxiliary controls on a new layer so that no objects can occlude them (as shown in Figure 3). Now, most of 3D tools also use this style to show their visual editing controls.

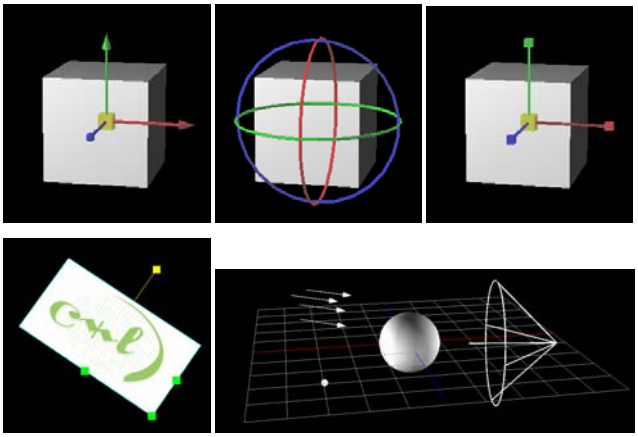


Figure 3: Some snapshot of auxiliary controls in our renderer. From top-left to right shows Move, Rotate, Scale, 2D control and some light auxiliary control. Most of editing controls are placed on the top layer and would not be occluded by geometric objects.

No matter what these geometric objects is, **Transform**, and **Transform2D** nodes indicate its transformation. Therefore, the renderer searches the nearest **Transform** or **Transform2D** node to modify the transformation if a user would like to modify the geometric layout in a scene.

4. RENDER RESOURCE MANAGEMENT

The best approach to archive the highest performance is to setup all the render resources and assign some resources to hardware at the beginning. Each rendering frame uses those pre-computed resources and uses hardware accelerations, which gives the maximum performance. That is because the computation and data transferring are reduced in each rendering frame. Unfortunately, the MPEG-4 scene is not static and might change over time by the BIFS Route mechanism, the BIFS-Anim stream, the BIFS-Command mechanism, and the JavaScript engine (as shown in Figure 2). Hence, we cannot put all the geometric resource into the graphic hardware and cannot calculate all the data only once. Therefore, a better mechanism of the renderer must be designed to solve above dilemma.

To improve the rendering performance, a render resource management is brought into our implementation. As shown in Figure 1, an abstract interface ***IRenderResource*** is designed to handle the resources management for each node. This interface accepts some render messages send from the BIFS kernel when the scene tree is modified. In other words, if some resources need to be rebuilt, the BIFS kernel sends the associated message to the render resource object. Any render resource instance, which inherits the *IRenderResource* interface, will implement the handle functions for following messages:

Message	Description
INIT	Triggered when the node is created.
DESTROY	Triggered when the node is destroyed.
MODIFY	Triggered when some fields of the node are modified So part of the render resource must be reinitialized.
TIME_SEEKING	Triggered when a user seeks the time line through the UI.
STRUCTURE	Triggered when the scene tree structure is modified. Maybe some nodes are added or removed.
TEXTURE_RATE	Triggered when the texture of a geometric object is changed. To handle this message, the renderer checks whether the texture coordinates of the geometric object need to be adjusted.
SWITCH_OUT	Triggered when a switch node change its selected child. This message will send to the previous selected child.

Table 1: The render resource messages are sent from BIFS kernel to a render resource instance.

Table 1 shows the render resource messages that would be send from the BIFS kernel to a render resource instance. For instance, the kernel will send **MODIFY** message when some fields of a node are modified. To take account of the performance, our system will not send the **MODIFY** message until all the fields are set. In our implementation, each field has a dirty bit to record its up-to-date states, and the renderer will rebuild the affected resources by referencing this dirty bit.

The height and width of a video frame in MPEG-4 are not power of 2 usually, but the graphic libraries (D3D or OpenGL) restrict height and width of a texture to be power of 2. One solution is to expand the video frame with an up-sample algorithm in each rendering frame. But, this method will reduce the rendering performance wildly. In our solution, the renderer adjusts the texture coordinates of the object instead of expanding the video frame. Therefore, the kernel will send message **TEXTURE_RATE** that contains two parameters, **Rate_U** and **Rate_V** (as Figure 4) when the texture of an object is changed. The texture address (Rate_U, Rate_V) indicates the ratio of the

video frame size to the real texture size. The renderer scales the texture coordinates of the geometric object according to **Rate_U** and **RATE_V** when it accepts this message. This trick is simple and increases the performance highly. However, if the texture wrap of the geometric object is using tiling method, this mechanism will fail. But in most cases, using tiling texturing with a video texture is not much meaningful.

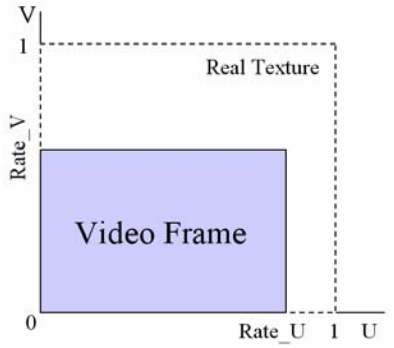


Figure 4: The message **TEXTURE_RATE** contains two parameters, Rate_U and Rate_V. If Rate_V is negative, then the image is inversed.

5. RESULTS

Figure 5 shows some of the MPEG-4 applications generated by our authoring tool and presented by our renderer. We use the JavaScript mechanism to achieve the goal of dynamical scene and user interactions. Some of scenes combine 2D and 3D objects and allow users to interact with the contents. The game logic is created by JavaScript, which we consider it more intuitive than using the BIFS-Command mechanism or the Route mechanism. That is because these two mechanisms are hard to record the game states and hard to handle the game interactive logic.

Figure 5: Some snapshots of MPEG-4 applications. The first three pictures is a MPEG-4 game. This game mixed 2D and 3D within a single scene, and using JavaScript to control the logical interaction. The last picture is a MTV scene and contains some videos.

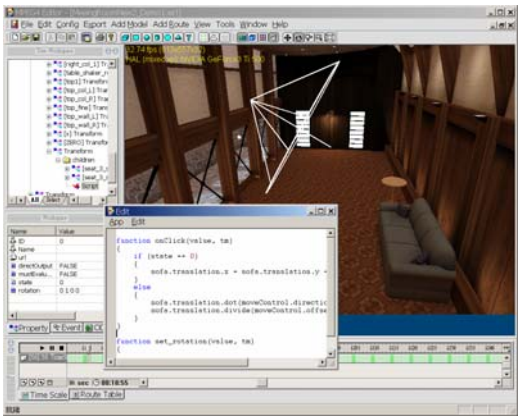


Figure 6: The snapshot of the scene-authoring tool [2][4]. As you can see, a content author is editing a JavaScript function.

6. FUTURE WORK

There are some virtual characters with facial and body animation in our interactive applications. Our system uses some nodes that are defined in AFX (Animation Framework eXtension), which makes our system mix part of ISO/IEC 14496-1 and AFX, and become incompatible with any standard profile. To solve above problem, we still have much work to do.

7. ACKNOWLEDGEMENT

The National Science Council (NSC) has supported this project for 4 years, and we appreciate their promotion (NSC 90-2622-E-002-008). We appreciate digimax production dept. for the beautiful and interesting demos, and also the bug reports. And all partners also try their best in the sub-modules of this project, such like Mei-Yu Fan (BIFS-Audio), Yu-Ju Tseng (BIFS-Anim), and I-Chieh Hsu (Mesh Coding). And last, we thank to Prof. Ja-Ling Wu, our laboratory leader, who always encourages us to keep on improving this system.

8. REFERENCE

[1] ISO/IEC 14496 AM 1, Part 1: Systems, Part2: Visual, Part 3: Audio, Part 6: DMIF, International Organization for Standardization, 2000.

[2] D.R. Liu, M.J. Shieh, Y.C. Lee, W.C. Chen, “On the Design and Implementation of an MPEG-4 Scene Editor”, ICCE 2000 Digest of Technical Papers.

[3] VRML (IS 14772-1), Virtual Reality Modeling Language, April 1997.

[4] Meng-Jyi Shieh, Kuo-Luen Perng, Wen-Chin Chen, “The Design and Implementation of a Visual MPEG-4 Scene-Authoring Tool”, 2nd MPEG-4 Workshop and Exhibition.

The Design and Implementation of a Visual MPEG-4 Scene-Authoring Tool

Meng-Jyi Shieh, Kuo-Luen Perng, Wen-Chin Chen
Communication and Multimedia Lab.,
Department of Computer Science and Information Engineering,
National Taiwan University, Taipei, Taiwan

E-Mail: {feitian, perng, wcchen}@cmlab.csie.ntu.edu.t

ABSTRACT

We have implemented an MPEG-4 scene-authoring tool for complete profile in our laboratory. This tool is the extension of the result of our laboratory last year. We changed the system architecture and design for 3D scene, the event in/out mechanism, visual editing, and friendlier user interface. We believe that such a tool can allow users to produce MPEG-4 contents easily and thus can push the MPEG-4 standard to be widely used. In this paper, we describe the design and implementation of an MPEG-4 scene-authoring tool, and we also introduce some necessary components for an MPEG-4 scene-authoring tool.

1. INTRODUCTION

The MPEG-4 [1] standard is specified by ISO/IEC 14496. It builds upon three application areas: digital television, interactive graphics applications and interactive multimedia. It aims at establishing a universal, efficient coding of different forms of audio-visual data, which we call audio-visual objects. Moreover, it tries to offer a new natural and synthetic interactive multimedia. This goal will be attained by defining two basic elements:

1. A set of coding tools for audio-visual objects capable of providing support for different functionalities such as object based interactivity and scalability, and error robustness, in addition to efficient compression.
2. A syntactic description of coded audio-visual objects, providing a formal method for describing the coded representation of these objects and the methods used to code them.

The MPEG-4 scene composition defined in a language called BIFS (Binary Format for Scenes). BIFS is based on the Scene-Graph concepts employed in VRML [4]. It is a superset of VRML with adding features such as Facial Animation, Enhanced Audio, native 2D primitives, Streaming and Animation streaming protocols.

In order to allow users to interact with the presented scene, ISO/IEC 14496-1 provides Interactivity mechanisms, which are integrated with the scene description information in the form of linked event sources and targets. These event sources and targets are parts of the scene description nodes, and thus allow close coupling of dynamic and interactive behavior with the specific scene at hand.

A BIFS scene is not likely to be static. It may be added, deleted, or modified later on. An MPEG-4 content server can send a sequence of commands to update BIFS scene. We call those commands BIFS-Commands. A BIFS-CommandFrame consists of some BIFS-commands. However, ISO/OEC 14496 does not specify the storage form of BIFS-Command in server-side. It just specifies the transmit format between server-side and client-side. This implicates that a server-side program may dynamically create a BIFS-CommandFrame and than send it to client.

In this paper, we present the design and architecture of the authoring tool that allow users to visually edit a 2D/3D BIFS scene. Furthermore we introduce some necessary functions and tools that an MPEG-4 editor must hold. The propose system is an extension of our previous work [2] and we will still improve this system in future work.

2. MOTIVATION

MPEG-4, which includes the most of the newest generation of multimedia technologies, is the most complex coding standard in the world now. Owing to its enormous architecture and many system components, which are across many domains, it becomes hard to implement and has no real applications. To push this standard to be a success and widely used, it needs not only a complete implementation of the whole system but also abundant MPEG-4 contents. In order to let content providers create variant, interesting, and interactive contents easily and rapidly, we decided to develop a system, called MPEG-4 scene-authoring tool, for the content editing use. Following are the major issues of our MPEG-4 scene-authoring system design:

- What you see is what you get
- General and fewer restriction
- User friendly

Especially, the first and the third issue have great influences on the users. A system with these characteristics gives the users direct senses about the system manipulation and let users spend less time on learning how to use this system, so the training costs reduced.

From the point of view of system design, the second issue is usually opposite to the third one, because users need to control more detail parameters with operating a general system, and since the system need to provide more user interfaces to access there parameters, it won't be user friendly. To solve this dilemma,

we made many efforts to balance the weights of building a general system and making the system user-friendlier.

3. SYSTEM OVERVIEW

From the system operational aspect, the authoring tool consists of several major components. We briefly describe each component as follows:

3.1 Scene tree structure editor

The normal way to build an MPEG-4 scene is to use a VRML like description language. By using this description language, users can adjust all the detail options of a scene node. Unfortunately, the MPEG-4 standard defines exactly one hundred types of nodes with different functionalities. In addition, the more details defined in the language, the more difficulties for general users to learn and use it.

For this reason, we provide a visualized Scene Tree Structure Editor. By using this tool, users do not need to know any grammar of the description language and the hierarchically relation of the scene objects. This tool allows users to construct a BIFS scene with only a little BIFS concepts. Therefore, it is necessary for the MPEG-4 scene-authoring tool to provide such a tool.

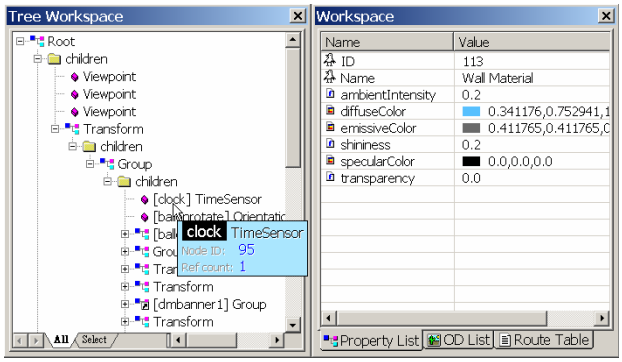


Figure 1. Snapshots of scene the tree structure editor window and the node properties editor window

3.2 Node properties editor

A scene node consists of some fields. Each field has its own data type. Sometimes, these data types are not intuitive and may raise difficulties for uses to assign their data values. For this reason, our system provides a node properties editor. It allows users to see the detail of a node and edit the attributes of the node. The editor provides different interface for assigning the data values to different fields.

3.3 Event (route) editor

Users may provide the “route entry,” which is an interactive relation between two nodes in the scene to enable user interactivity. A node may have several “event in” and “event out” fields. The definition about which fields can “event in” or “event out” in a node is defined in ISO/IEC 14496-1:1999 Annex H. There are exactly 100 nodes, and every node has its own special event definitions. Users may get confused with it. Therefore, we must provide a way to give hint to users as to what they can choose.

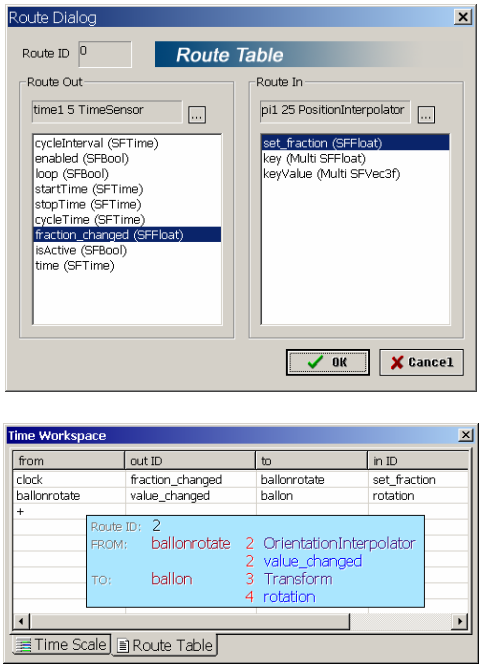


Figure 2. A snapshot of editing event entry and route table

3.4 BIFS-Commands editor

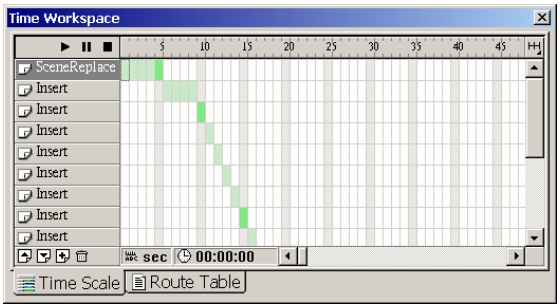


Figure 3. Snapshots of BIFS-Commands window – Every row defines a BIFS-Command, and its column indicates the showing time.

BIFS-Commands are used to modify a set of properties of the scene graph, its nodes and behaviors at a given time. Commands are grouped into CommandFrames in order to be able to send several commands in a single access unit. There are four basic commands, as follow:

1. Replacement of an entire scene
2. Insertion
3. Deletion
4. Replacement

We provide a BIFS-Commands editor to allow users adding BIFS-Command information in a scene. By using this tool users can specify a time to add some new nodes or to modify some attributes of a node in a scene. That can appear more plentiful variation in a scene than just monotone movement.

3.5 Preview and interactive window

In order to achieve the goal of "What you see is what you get," a preview window is provided. Users can directly see the visualized scene result instead of the scene description language scripts. Besides, they can control objects through this window. They may adjust position, size, and rotation angle of the objects by intuition. However, not all the objects in the scene are visible, these invisible nodes cannot be controlled by this way.

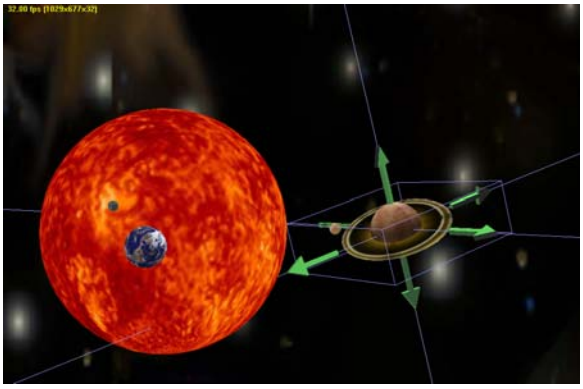


Figure 4. Snapshots of preview and interactive window – In the scene of the solar system, a planet is been editing by user mouse actions, and the bounding box and some control points are appearing in the preview window.

3.6 Import object

There are several reasons why we must provide such functionality. First, an author may have previously created some 3D objects in a different format. With this functionality, he may reuse those 3D objects in a new scene. Second, an author may be accustomed to certain other 3D model editors, such as Maya or 3D Studio Max. He can still edit a static scene with his custom 3D model editor, and then imports this static scene into our system.

3.7 Export to server side storage format

ISO/IEC 14496-1 does not specify the scene storage format of a scene at the server side. It only specifies the format of the bitstream, which is transmitted from the server-side to the client-side. This implicates that the bitstream of BIFS may be produced dynamically by an MPEG-4 server-side program. Therefore, we are developing an interactive MPEG-4 streaming server. In this project, the streaming server uses server-side JAVA script to produce the bitstream of the BIFS-CommandFrame. The current authoring tool saves the scene data into the format of the new interactive MPEG-4 streaming server.

4. DESIGN OF THE SYSTEM ARCHITECTURE

ISO/IEC 14496-1 specifies 31 Node data types, 100 kinds of nodes, and 20 basic data types. This is not a simple case for object-oriented programming designed for MPEG-4 scene editing and rendering. Not only many classes need implementing, but also some mechanisms need designing in kernel.

4.1 The system modules

In high-level view, our architecture consists of three modules: **BIFS Editor**, **Scene Render**, and **Scene Graph Manager,** as illustrated in Figure 5. The actions of the user will affect the organization of the **BIFS scene tree** by going through both **BIFS Editor** and **Scene Render**.

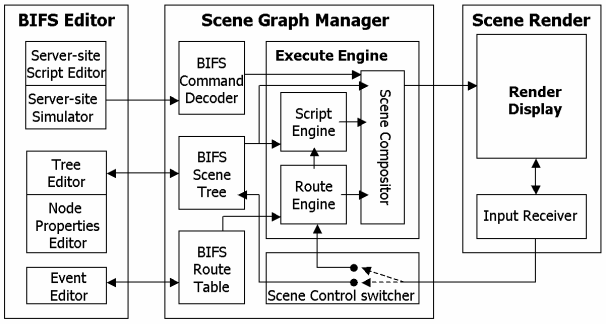


Figure 5. The System Architecture

The **BIFS Editor** module can show BIFS Scene Description in a user-friendly way. It provides a visual, non-language, and instantly error-checked mechanism. It contains 4 submodules: **Tree Editor**, **Node Properties Editor**, **Event Editor**, and **Server-site Script Editor** modules. Users can directly read and modify scene tree organization in **Tree Editor**, and modify the attribute of the selected node in **Tree Editor** by using **Node Properties Editor**. In addition, the entry of **BIFS Route Table** can be illustrated and modified by using **Event Editor**.

The **Scene Graph Manager** module, which consists of some data and controllers, is the kernel of our system. While the data of **BIFS Scene Tree** and **BIFS Route Table** are the most critical in this module, **Execute Engine** is the most important controller in this module. **Scene Compositor** would duplicate the data of **BIFS Scene Tree,** which will be used or modified later by other controllers. Once an event has occurred, **Route Engine** will route this event to its target. Meanwhile, the data of **Scene Compositor** might be modified or some script function executing in Script Engine could be triggered.

To fulfill What-You-See-Is-What-You-Get, the scene tree will be sent to the **Scene Render** module for previewing after composing by **Scene Compositor**. The action of users will be sent to **Scene Control Switcher**, which will then deliver the event to different block according to the current editing mode. On one hand, if the editor is in the preview mode, the UI message will be sent to **Route Engine**; on the other, when editor is in edit mode, the UI message will be sent to **BIFS Scene Tree.**

Moreover, **BIFS Scene Tree** could be modified according to the user actions.

4.2 Event route mechanisms for virtual editing

To realize the objective of virtual editing, some mechanisms must be added to the scene tree node classes and the render module. An object in a render window consists of some nodes. Therefore, if a user wants to move an object or resize an object by using the mouse in the preview window, the system may adjust field values in different nodes.

To make this problem more general, our system uses event route mechanisms. In editing mode of preview window, when users choose a node in the BIFS scene tree presentation window or click an object in the preview window, a sub-tree will be chosen, and we say this sub-tree is in focus state. After that, the preview window computes the bounding box and then shows some control points on the screen, as in Figure 4 and Figure 7. Every control point has its individual significance in different editing modes, such as **Move**, **Resize**, and **Rotation**, and so forth. When a user drags a control point, the control point will throw an event with some bearing information to the focused sub-tree. The event will be delivered recursively until a node accepts it. This node will decide to apply this event to its attributes or selectively deliver this event to its children nodes.

Figure 6 is an example of a box object. If an event arrives to the transform node that is the root of the sub-tree, when this event is **Move**, **Scale**, or **Rotation**, the transform node will accept this event and apply this to its attributes. But if this event is **Resize**, transform node will deliver this event to its children nodes. After that, the shape node accepts this event and simply delivers it to the geometry node. Finally, geometry node receives it and applies to size field.

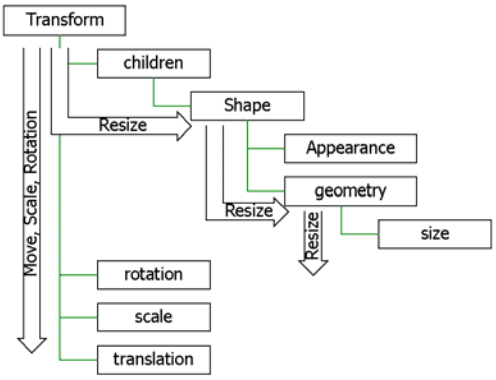


Figure 6. An example of the edit message route mechanism – This diagram shows the scene tree of a box, which receives an editing message and decides to send this message to children or hold it and applies it.

5. RESULTS

We implemented our system in Windows 2000 with VC++ 6.0, BCG UI library, and Direct3D/OpenGL library. Figure 7 is a Screen shot of the result. At the ending final, the authoring tool allows users to save these scene data into the data format of the MPEG-4 interactive server, which was also developed in our laboratory.

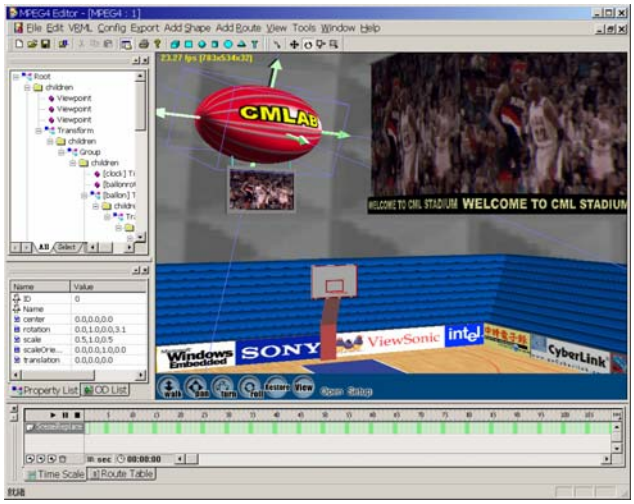


Figure 7. A snapshot of our system – You can see a hot balloon is in selected state and some control points are showing.

6. CONCLUSION AND FUTURE WORK

A new interactive authoring tool is currently under development. It allows editing client-site and server-side JAVA-script and simulating its executing result. We will use JAVA-Script solution to support BIFS commands and the user interactivity.

7. REFERENCE

[1] ISO/IEC 14496 AM 1, Part 1: Systems, Part2: Visual, Part 3: Audio, Part 6: DMIF, International Organization for Standardization, 2000.

[2] D.R. Liu, M.J. Shieh, Y.C. Lee, W.C. Chen, “On the Design and Implementation of an MPEG-4 Scene Editor”, ICCE 2000 Digest of Technical Papers.

[3] Y.S. Tung, J.L. Wu, and C.C. Ho, “Architecture Design of an MPEG-4 System”, ICCE 2000 Digest of Technical Papers.

[4] VRML (IS 14772-1), Virtual Reality Modeling Language, April 1997.

分項計畫二：MPEG-4/7媒體內容分析與視訊串流技術 (III)

Content Analysis and Streaming Technology for MPEG-4/7 (III)

成果報告

計畫編號：NSC91-2622-E-002 –002

全程計畫：民國 90 年 08 月 01 日至民國 93 年 07 月 31 日

本年度計畫：民國 92 年 08 月 01 日至民國 93 年 07 月 31 日

計畫主持人　：吳家麟　台灣大學資訊工程系教授
共同主持人　：黃肇雄　台灣大學資訊工程系教授
陳文進　台灣大學資訊工程系教授
歐陽明　台灣大學資訊工程系教授
周承復　台灣大學資訊工程系助理教授
陳炳宇　台灣大學資訊管理系助理教授

一、摘要

本分項計畫名稱為「MPEG-4/7 媒體內容分析與視訊串流技術」，共分為三項技術研發分項：(1)【影音特徵值粹取工具集】(audio/video features extraction tools)，(2)【內容分析與摘要模組】(content analysis and summarization module)與 (3)【資訊串流模組】(content streaming module)，用以支援總計畫**【媒體內容工程: MPEG-4/7 相關技術之研發】**中所提出的**”互動資訊媒體服務系統（Interactive Information Media Service System）”**與**”家庭媒體中心（Home media center）”**兩項核心應用。

【影音特徵值粹取工具集】技術分項之主要目的為：藉助現今電腦的快速運算能力，輔以使用者的輸入與幫助，以達到有助於多媒體資料精簡與搜尋，並於 MPEG-7 標準所函括之低階層特徵值之粹取。在【內容分析與摘要模組】此技術分項內，則進一步研究具有意義資訊（semantic）的取得與多媒體資訊之摘要精簡化（summarization）。在【資訊串流模組】技術分項中，則針對具可調整性與容錯誤性的壓縮方式，做深入的研究與探討，並整合串流伺服器與即時串流協定等技術，完成 MPEG-4 多媒體資訊及時串流的實驗平台。本計畫將以個人電腦(PC)為發展平台，分三年完成上述技術模組的研發以及兩項與媒體內容相關的系統。

The project --『Content Analysis and Streaming technology for MPEG-4/7』, is composed of three technical development modules, including (1) 【Features extraction tools】, (2) 【Content analysis and summarization module】 and (3) 【content streaming module】。These three modules support the two kernel applications "（**Interactive Information Media Service System**）" and "（**Home media center**）" proposed in the project 【 MPEG-4/7 based Content Engineering 】。

The 【Features extraction tools】 module will develop techniques of the MPEG-7 defined low level features extraction for video and audio contents。The developed techniques can be useful for the summarization of multimedia content and the content-based query by the computation power and appropriate user's assistance。The purpose of the 【Content analysis and summarization】 module is to further develop the techniques about the extraction of the semantic information and the summarization of multimedia content。The final module -- 【content streaming module】is focused on the research of the scalable and error resilience codec. Integrating the streaming server and real-time streaming protocol, we will finish the experimental platform for supporting the real-time MPEG-4 multimedia content streaming。The project is to be developed on PC platform. It is scheduled to be completed in three years.

二、計畫緣由與目的

隨著 Internet 的蓬勃發展、網路寬頻技術的進步、以及個人電腦計算能力的大幅提升，資訊的產生，已經由過去的只由傳播媒體（報紙、電視）提供，普遍到現在的可由一般資訊接受者、資訊使用者來產生。時至今日，存在有各種不同來源的影音（audio-visual）多媒體資訊，它們正以各種不同的形式顯示與儲存。而隨著科技的進步，使得一般大眾也能輕易創造多媒體資訊；資訊的數量，也自然地呈倍數般地成長。因此，如何有效地利用這些資訊，如何快速地找到所需的資訊，以及如何將這些雜亂無章的資訊轉化成有價值的知識，都是隨著資訊量快速成長所衍生出來的重要研究課題。值得注意的是，此處所指的資訊，並不局限於文字模式（text-base），更包括了所有多媒體形式的資訊。

以往對多媒體的各種應用而言，如何有效率地處理多媒體內容的產生、儲存、傳送、編輯、呈現，都是最後是否成功的重要關鍵。因此 MPEG-1、MPEG-2 標準的制訂，成功地促成了 VCD、DVD 等多媒體應用的發展。而 MPEG-4 標準的產生，更刺激了許多雙向式多媒體的應用。然而，在邁向數位資訊公路、文件（Document）革命和內容（Content）革命的當下，許多亟待思考與解決的新課題也隨之產生。審視新一代 MPEG 國際標準，多媒體實驗革命的精神與實踐，儼然已包含在近年來所大力制訂的 MPEG-4、MPEG-7 與 MPEG-21 的標準當中。總括來說，多媒體文件的形式，屬於 MPEG-4 的範疇，而多媒體內容使用的變革，則是 MPEG-7 與 MPEG-21 致力的重點。

為能更清楚的呈現本計畫緣由與目的，以下的段落中，我們將分別就 MPEG-4 及 MPEG-7 兩項標準的制訂精神與技術特色做重點式的說明。

MPEG-4 技術之特色包括：

1. 可調整性（Scalability）

   為了包含各種多媒體應用的需求，MPEG-4 在訂定標準的過程中，特別注重「可調整性」，希冀對於不同的網路環境、計算能力、使用者需求，都能提供適當的支援。對於編解碼器來說，所謂可調整性包括了下列項目：

   ✧ 空間解析度上的可調整性
   ✧ 時間解析度上的可調整性
   ✧ 畫面呈現品質的可調整性
   ✧ 壓縮位元率的可調整性
   ✧ 計算能力需求的可調整性
   ✧ 以物件為基礎的可調整性

2. 互動性（Interactivity）

   MPEG-4 對於使用者與應用程式的互動，也做了適當的定義。在編解碼器執行的過程中，使用者擁有比以往更多的決定權。舉例而言，使用者有權決定：

   ✧ 每秒呈現的畫面數目（Frame Rate）
   ✧ 整體或各物件的畫面之大小及位置
   ✧ 整體或各物件的視覺及聽覺品質

3. 支援多重網路環境（Supporting of Heterogeneous Networking Environment）

MPEG-4 在網路傳輸方面，不著重特定的網路架構，而是以訂定共通介面的方式，讓多媒體應用程式可以適用在不同的網路環境。

4. 支援真實與合成場景（Supporting of Natural and Synthetic Scenes）

MPEG-4 的另一特色，是讓真實物件（如影片、自然語音）與合成物件（如 VRML 物件、人工合成語音）共同呈現在同一場景中。相對應地，編解碼器便要能處理自然與合成場景的混合編解碼。

5. 可擴充性（Extensibility）

MPEG-4 期許成為多媒體應用在編解碼方式上的最終標準。因此對於擴充性格外注意。透過其所定義之 SDL（Syntactic Description Language）語言，MPEG-4 可以容納新的編解碼工具。

有別於 MPEG 之前所推出的標準，MPEG-7 被設計與定位成與媒體檔案共存共榮的好夥伴。雖然聽起來簡單，但 MPEG-7 所勾勒出的可是一個基本而實在的概念。MPEG-1 到 MPEG-4 提供的都是與音訊和視訊壓縮和解壓縮有關的標準，MPEG-7 則為“多媒體內容描述介面”或簡稱為 MCDI（Multimedia Content Description Interface）。

MPEG-7 試圖將足以代表多媒體內容的種種特徵（features）標準化、量化，成為電腦系統可處理的共通資料結構，包括術語上所謂的：

- 敘述詞（Descriptor）
- 描述結構(Description Scheme)
- 描述結構定義語言（the Description Definition Language (DDL)）

敘述詞（Descriptor）用以形容某個具獨立屬性的特徵，例如：一張圖像（Image）的主要形成顏色（Dominant color）；描述結構（Description Scheme）則是包含了數個描述詞或小集合的描述結構，用已形容某些具關連性的特徵。描述結構定義語言（the Description Definition Language (DDL)）用以詳細指明描述結構的階層關係，並允許使用者以此語言產生新的描述結構。總括來說，MPEG-7 其主要精神與技術特色包括：

➢ 敘述詞（Descriptor）部分：

- 混和形式（Cross-modality）

  MPEG-7 定義含括影像，音訊及其他種類的敘述詞（Descriptor），允許資料不同形式間的轉換及搜尋。

- 資料適應（Data adaptation）

  MPEG-7 定義有關內容的敘述詞（Content Descriptor），可以根據此資訊所反映出來使用者裝置的能力、網路資源、使用者的喜好、使用者的環境等等，來賦予其最有意義的內容資訊。

- 唯一性（Unique identification）

  MPEG-7 提供唯一辨別資料的機制，換言之，它可以為所要描述的資料與敘述詞定下清楚的關連性。

➢ 描述結構(Description Scheme) 部分：

- 描述結構關係（Description scheme relationships）

  MPEG-7 定義高階層的描述結構(Description Scheme)，用來表示出其組成敘述詞

間的相關性。

- 階層性與可調整性（Hierarchy and scalability of descriptors ）

  MPEG-7 定義高階層的描述結構(Description Scheme)，提供階層性與可調整性的特點，使得搜尋的動作更有效率。以階層性來說，N 層的 descriptors 可補 N-1 層的不足；以可調整性來說，N 層的 descriptors 可視為 N-1 層的功能加強與精緻化。

➢ 描述結構定義語言（the Description Definition Language (DDL)）部分：

- 創新能力（Compositional capabilities）

  MPEG-7 定義描述結構定義語言（the Description Definition Language (DDL)），允許建立新的敘述詞（Descriptor）與描述結構(Description Scheme)，並允許針對現存的敘述詞（Descriptor）與描述結構(Description Scheme)作修改及增強。

- 多類型的媒體格式（Multiple media types）

  MPEG-7 定義描述結構定義語言（the Description Definition Language (DDL)），提供一個廣泛的機制，將敘述詞（Descriptor）、描述結構(Description Scheme)與所要描述資料之間的相關性詳細指明。此處所描述的資料包含了多類型的媒體格式，可包括影像，音訊，文字，甚至是文字敘述的描述資料，或上述格式的結合。

- 作業平臺獨立性（Platform independence）

  描述結構定義語言（the Description Definition Language (DDL)）必須符合應用層面與作業平臺互相獨立。

- 符合文法規則（Grammar）

  MPEG-7 定義描述結構定義語言（the Description Definition Language (DDL)），必須遵守一個不會造成混淆性但容易解析的文法規則。

三、研究方法及成果

A. 【以使用者注意模型為基礎之視訊興趣區自動決定系統】

1. 前言與研究目的

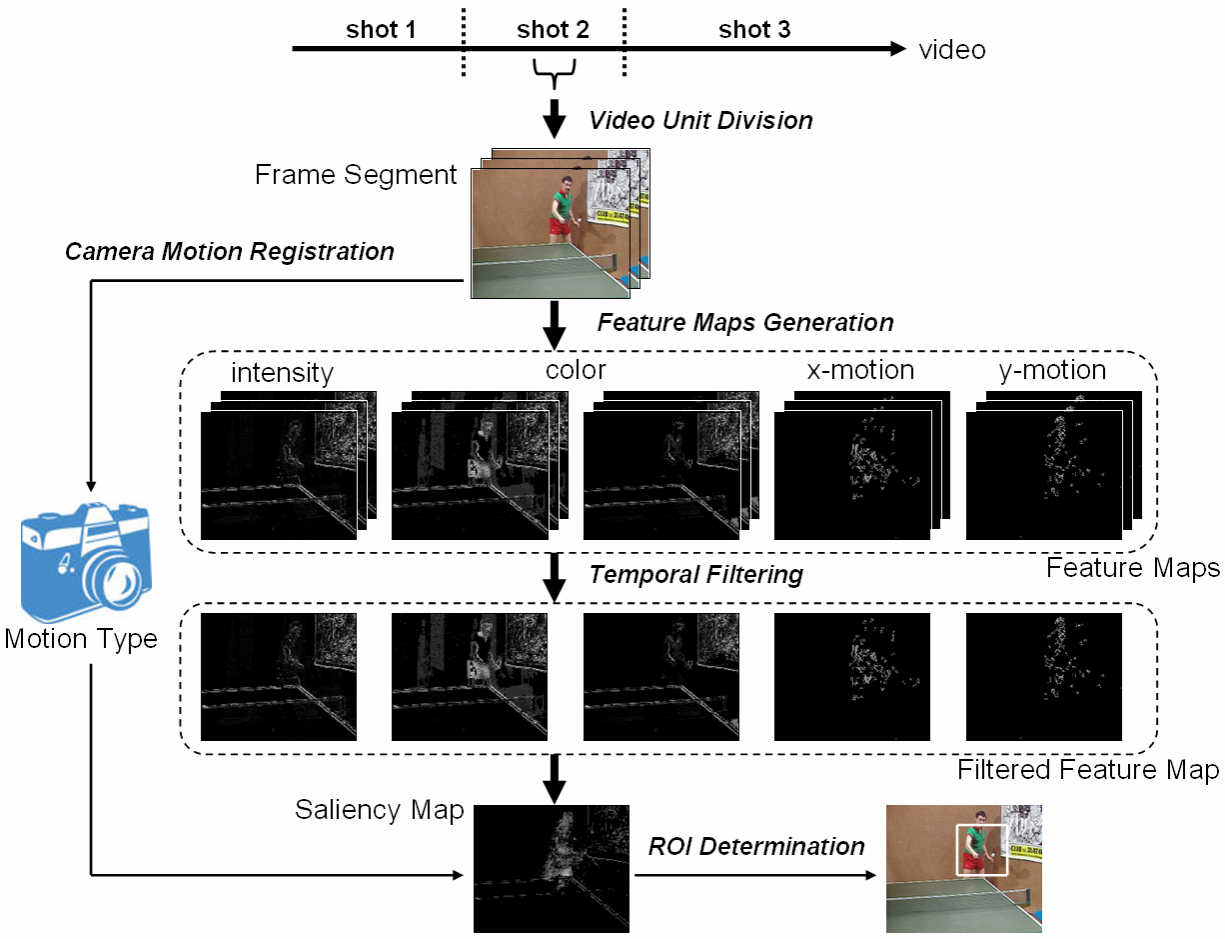
近年來，隨著多媒體產生技術的快速進步，加上網際網路對於資訊傳播的便利性，造成了多媒體文件 (multimedia document) 在數量上呈現急遽的增加。面對如此大量的多媒體資訊，越來越多的使用者希望能僅獲得該些文件簡要的內容精華，而非原始的完整文件。而另一方面，雖然在影像處理 (image processing) 、人工智慧（artificial intelligent）及電腦視覺（computer vision）等相關領域已有許多重要的研究成果，然而，其對於如何能自動分析出多媒體文件內容之語義 (semantics) 仍無法解決。因此，如何能依照人類知覺上的感知，來發展出一套符合人類知覺需求的興趣區 (Region-of-Interest, ROI) 決定系統，也就是可找出多媒體文件中對於觀察者較具意義或重要性的部份，將是數位化時代多媒體管理及研究當前一個重要的課題。

傳統的興趣區分析主要著重於兩種多媒體型式：影像 (image) 與視訊 (video)。然而，在視訊方面的研究成果卻遠落後於影像的相關研究。這種情形肇因於忽略了許多視訊獨有的特性。第一、視訊可視為由一連串的訊框 (frame) 沿時間軸所構成，而越相近的訊框彼此間具有的相關性越高，而非互相獨立存在的影像；第二、在視訊的產生過程中，拍攝者常會使用運鏡 (camera motion) 的技巧來強調視訊中的重點或藉以引導觀眾的注意力。因此在分析視訊興趣區時，我們將考慮「應用媒體美學」 (applied media aesthetics)，也就是利用視訊拍攝時慣用而遵循的基本準則，來增加興趣區分析時的準確度與代表性。

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。在這個研究中，視訊的注意特徵值 (attentive features) 及應用媒體美學的知識都被同時考慮且利用。本研究將成為達到更高階具意義性視訊分析的一個重要基礎。

圖(一) 包含一個興趣區(白色方框內)之影像。

2. 研究方法

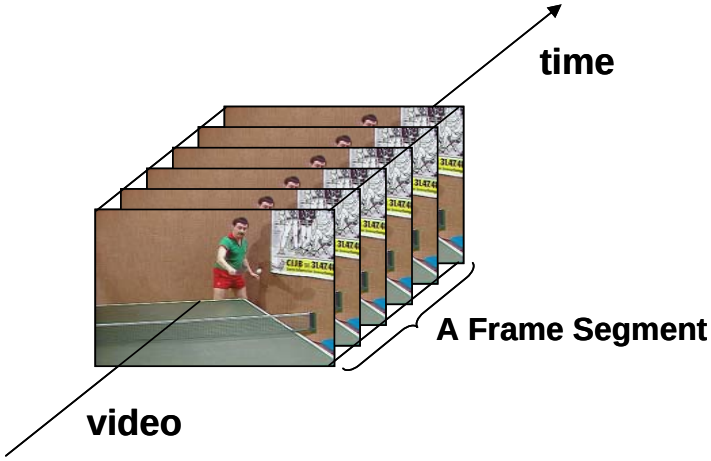


圖(二) 視訊興趣區決定之所提架構。

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。

[1] 視訊之使用者注意呈現

視訊和影像兩者間一個最主要的差異即是視訊具有時間的概念，而影像則無。然而，傳統的視訊興趣區分析仍將視訊視為由各自獨立的影像所構成，而以該些個別影像為基礎進行分析。以此種方式所找出的興趣區，從個別影像的角度來看似乎還相當令人滿意，但以整體從視訊的角度出發，我們可明顯發現所找出的興趣區互不連貫，常有不合理的甚至可視為是雜訊的情況產生。為了提高視訊興趣區分析的正確性及合理性，我們改以一小段連續的訊框，稱之為訊框切片，做為我們分析的基礎單位。在實驗中，目前我們是以 0.5 秒長度的訊框量組成一個訊框切片。其最大的好處為，不但可以在考慮平面上的資訊時，同時掌握到時間軸上的資訊，更可以大幅降低雜訊所造成的影響。



圖（三）訊框切片實例。

我們利用使用者注意模型來決定訊框切片中每一張訊框之相關視覺注意資訊以獲得其對應之特徵值映圖，而其所包含的使用者注意子模型包含三大類，分別為亮度、顏色及運動。亮度子模型是負責呈現訊框中各像素 (pixel) 之亮度對比性資訊 (*IC*)，而顏色子模型則負責呈現訊框中各像素之顏色對比性資訊，其可再進一步細分為兩類，即紅綠對比性資訊 (*RGC*) 及藍黃對比性資訊 (*BYC*)。另外，運動子模型則負責呈現訊框中各像素的移動強度資訊，其亦可細分為水平向移動資訊 (x-motion) 及垂直向移動資訊 (y-motion)。亮度及顏色對比性的三種資訊分別定義如下:

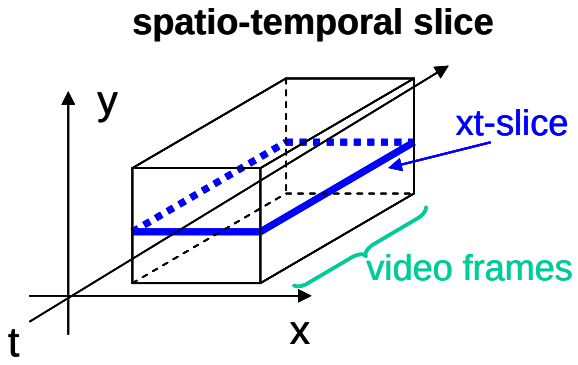
$$IC(p) = \max_{p' \in w} |I(p) - I(p')|，$$

$$RGC(p) = \max_{p' \in w} |R(p) - G(p')|，$$

$$BYC(p) = \max_{p' \in w} |B(p) - Y(p')|。$$

其中 p 代表該像素之位置向量，w 為以該像素為中心之一方形區域，而 I、R、G、B 及 Y 分別為取出該像素之亮度、紅色、綠色、藍色及黃色成份之函數。而運動子模型之水平向及垂直向移動資訊則分別由二維結構張量 (2-D structure tensor) 所決定。我們以水平向移動資訊之計算為例，我們可將一訊框切片視為是一時空薄片 (spatio-temporal

slice)，接著我們在所有的水平薄片 (xt-slice) 上分別以二維結構張量對每一像素進行計算，即可得到其對應之水平向移動資訊。另一方面，以同樣方式計算出垂直向移動資訊後，每一像素即可得到其對應的完整移動資訊。使用二維結構張量計算移動資訊之最大好處為可直接以像素本身之亮度資訊進行計算，同時可以得到該移動資訊之可信度數值 (*cm*)。二維結構張量 (*J*) 與其相關計算式定義如下。



圖（四）時空薄片及水平薄片之實例。

$$J = \begin{bmatrix} J_{xx} & J_{xt} \\ J_{xt} & J_{tt} \end{bmatrix},$$

$$J_{xx} = \sum_{x',t' \in w} \hat{H}_x^2(x-x',t-t'),$$

$$J_{xt} = \sum_{x',t' \in w} \hat{H}_x(x-x',t-t')\hat{H}_t(x-x',t-t'),$$

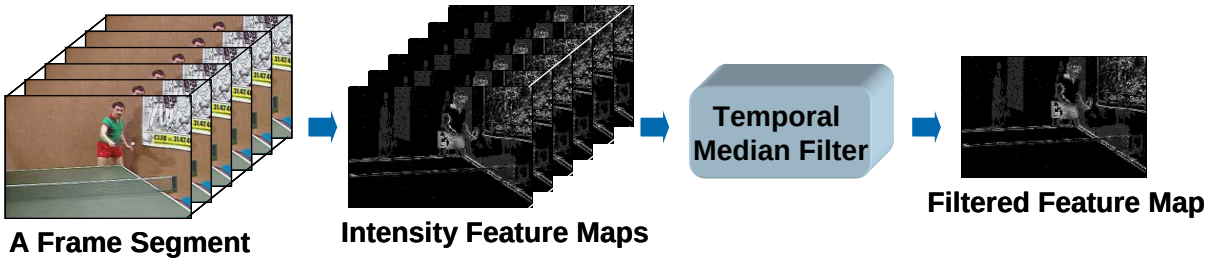
$$J_{tt} = \sum_{x',t' \in w} \hat{H}_t^2(x-x',t-t'),$$

$$\hat{H}_x = \partial(G*H)/\partial x \ , \ \hat{H}_t = \partial(G*H)/\partial t.$$

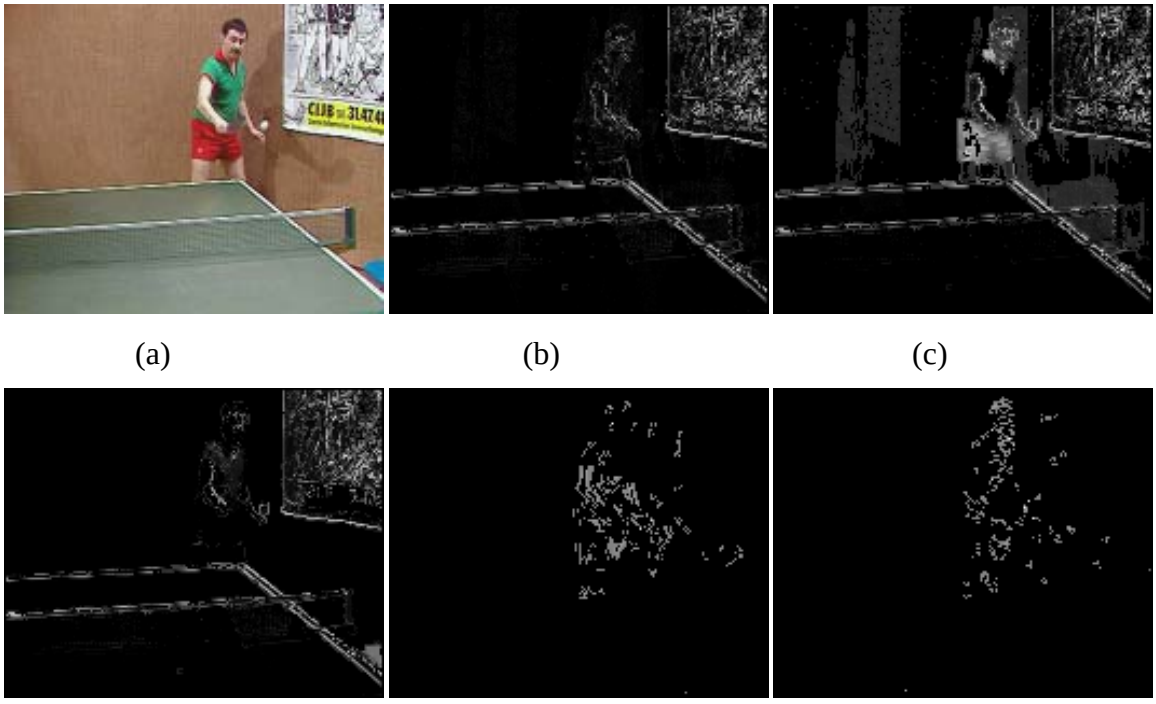
其中 *w* 為以(*x*,*y*)坐標為中心之3×3像素區域，*G* 為高斯核心。

$$\theta = \frac{1}{2}\tan^{-1}\frac{2J_{xt}}{J_{xx}-J_{tt}} \quad , \quad cm = \frac{(J_{xx}-J_{tt})^2+4J_{xt}^2}{(J_{xx}+J_{tt})^2}, \ 0 \le cm \le 1.$$

而θ即為該像素所對應之水平向或垂直向移動資訊，其值介於[-90^o, 90^o]，以水平向移動資訊為例，0^{o}代表無任何移動，而正值與負值分別代表向右方及左方移動，且其值越大，移動強度越強。



圖(五) 亮度已過濾特徵值映圖之產生流程示意圖。



圖(六) 單一原始訊框及以其獲得之特徵值映圖實例。其分別為：(a)原始訊框，(b)亮度特徵值映圖，(c)紅綠對比特徵值映圖，(d)藍黃對比特徵值映圖，(e)水平向移動映圖，及(f)垂直向移動映圖。

在獲得訊框切片中每張訊框所對應之各類特徵值映圖後，我們對每一類注意特徵值之所有映圖均以一時間平均過濾器將其過濾為單一之已過濾特徵值映圖。已過濾特徵值映圖在視訊興趣區分析中扮演了相當重要的角色，其包含了某訊框切片中某類特徵值的共同顯著區域，因此可以大幅提升興趣區分析的準確性。以亮度已過濾特徵值映圖為例，其可用以呈現在該訊框切片中整體的亮度特徵值資訊。

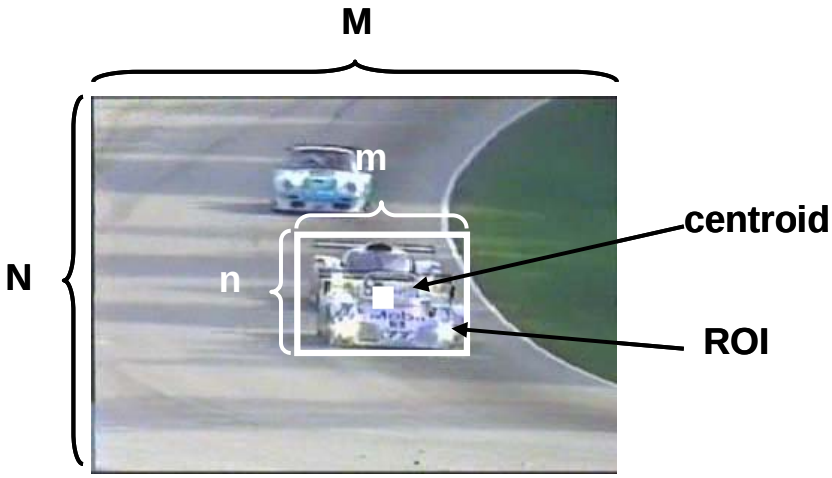
另一方面，在視訊中常會出現不同類型的運鏡類型，而這些類型由應用媒體美學的角度而言則不同程度的隱含了拍攝者想呈現給觀眾那個部份在視訊中是該視訊的重點或較為值得注意的部份。以變焦放大 (zoom-in) 的運鏡為例，拍攝者即著重於將訊框之中央區域呈現給觀眾，而若是使用向左搖鏡 (left-pan)，拍攝者則可能較著重於將向左方移動之某些物體呈現給觀眾。我們可發覺在視訊中，不同的運鏡類型將大幅提升或降低某些特徵值對於觀眾的影響力，在視訊之興趣區分析中，運鏡扮演了重要的角色。因此，

在獲得某一訊框切片之所有已過濾特徵值映圖後，我們需將其合併為單一之顯著映圖以代表整體之使用注意性，而我們即以該訊框切片所屬之運鏡類型來決定不同類之已過濾特徵值映圖之合併參數。

$$S(N)=\alpha_{c,1}\times FFM_1+\alpha_{c,2}\times FFM_2+\cdots+\alpha_{c,n}\times FFM_n\ ,$$

其中$S(N)$為第N張訊框之顯著映圖，而FFM_i與$\alpha_{c,i}$分別為第i類已過濾特徵值映圖及其對應之合併參數。

[2] 視訊之興趣區決定



圖（七）以顯著性加權規律矩決定興趣區中心位置及大小之實例。

在顯著映圖中，具有越高顯著值的像素代表越容易吸引使用者的注意，反之亦然。因此，在某訊框切片之顯著映圖後，我們採用顯著性加權規律矩 (saliency weighted regular moment) 來決定每個興趣區之中心位置 (centroid) 及其大小。

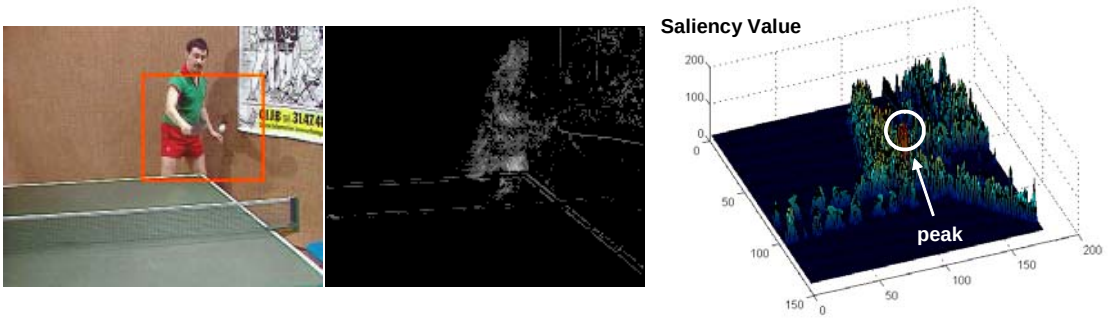
$$m_{pq}=\sum_{1}^{M}\sum_{1}^{N}x^p y^q s(x,y),\quad p,q=0,1,2,...,$$

$$\eta_{pq}=\frac{\sum_{1}^{M}\sum_{1}^{N}(x-\bar{x})^p(y-\bar{y})^q s(x,y)}{m_{00}^{(p+q+2)/2}},\ \bar{x}=\frac{m_{10}}{m_{00}},\ \bar{y}=\frac{m_{01}}{m_{00}},$$

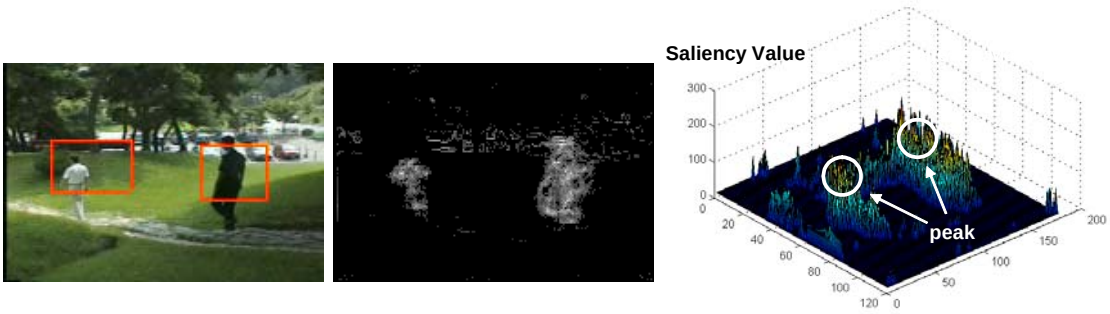
其中 M 及 N 分別為顯著映圖之長與寬，而 $s(x,y)$為對應於像素(x,y)之顯著值函數。而該興趣區之中心位置即定為$(\bar{x},\bar{y})=(m_{10}/m_{00},m_{01}/m_{00})$，而大小則定為$(k\sqrt{\eta_{20}})\times(k\sqrt{\eta_{02}})$，其中 $k=2$。

對於觀眾來說，當他們在觀賞一段視訊時，很自然的會隨著視訊的播放而在不同情況下感知到不同數量的物體或區域較吸引他們的目光。當然，在觀看不同視訊時更容易有這樣的情形發生，例如在觀賞一段網球公開賽時，我們很容易的就會注意到兩邊的球

員以及比賽中的網球，而在欣賞個人歌唱表演時，我們的視線很可能就被歌唱者所完全吸引了。換言之，一段視訊在不同時間點之訊框中，可能會有不同數量的興趣區，因此我們對於視訊興趣區數量，也必須能隨著視訊的播放而動態的決定並調整。我們發現，若將一個訊框切片的顯著映圖立體化，也就是說我們將顯著映圖上每個像素所對應的顯著值做為第三維的高度量值，則我們可以將其繪成如地形圖般的立體圖。將其與原來的訊框做比較，則在一般情況下立體圖中的高山山峰數目剛好會對應到吸引使用者目光的區域數量。因此，我們即可依此觀察動態的對每一訊框切片的顯著映圖進行興趣區數量決定。另外，我們也發現一訊框之興趣區數量通常不會超過三個，因為過多的興趣區數目代表的其實就是無主要的興趣區，他們任一均無法特別引起觀者的注意。



(a) (b) (c)



(d) (e) (f)

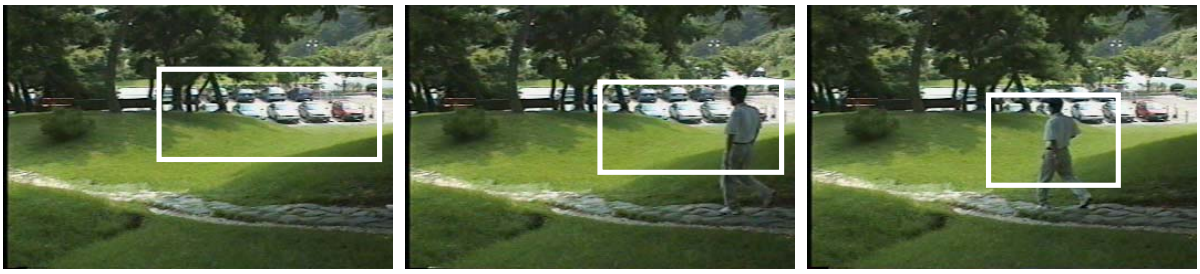
圖（八）分別具有一及二個興趣區之訊框切片實例(a)與(d)。圖(b)及(e)分別為其對應之顯著映圖，而(c)與(f)則為其對應顯著映圖之三維立體成像。

3. 實驗結果

表(一) 興趣區決定之使用者實驗結果。

	GOOD (%)	ACCEPT (%)	FAILED (%)
Data Set I	82	15	3
Data Set II	73	21	6
Avg.	78	18	4

為了檢驗本視訊興趣區決定架構之效能，我們對使用者實際進行實驗。由於在架構中採用了應用媒體美學的知識，因此我們挑選由攝影方面專家所拍攝的視訊做為我們的實驗測試影片。我們將測試影片分為兩大類，第一大類 (Data Set I) 包含電視節目、電影及廣告等，而第二大類 (Data Set II) 則包含以運動影片為主的視訊。每類均包含大約 10 部左右的視訊，而每部約 2~3 分鐘左右長度。接著我們請使用者在看完每部測試影片後，依所決定出之興趣區及本身主觀的滿意程度對每部視訊給定一個意見，可給定的意見有相當好 (good)、可接受 (accept) 及不滿意 (failed)。從實驗結果中可看出，我們無需瞭解各視訊的內容意義，以使用者注意模型找出使用者對訊框的顯著性觀感，再加入應用媒體美學的相關知識，即可找出令其相當滿意的興趣區，亦證明了本架構的有效性及其在模擬使用者顯著視覺上的準確性。



52 79 108

125 143 181

209 232 251

267 307 482

圖(九) 興趣區決定之實例。圖下之數字代表該訊框在視訊中之順序編號。

B.【應用內容知覺機制於音樂導向的視訊摘要系統】

1. 前言

隨著資訊科技的快速進步，數位攝影機（Digital Video Camcorder）的價格已逐漸能讓一般的使用者接受，並且在眾多的家庭中逐漸普及。相對於傳統的 V8 攝影，數位影片不但便於自行編輯、複製與傳播交換，而且每單位的製作與儲存成本也相對低廉。因此，使用者利用數位攝影機來記錄他們日常生活中的精彩片段以及零星的大小事。

對於使用者來說，拍攝影片是一件非常有趣的事，可是事後的剪輯工作卻令人感到非常的挫折。雖然市面上有許多專業的視訊剪輯軟體，例如：Adobe Premiere、Ulead MediaStudio 和 CyberLink PowerDirector 等等。但對於家庭使用者來說，這些軟體依舊過於複雜，並且不容易學習使用；更何況，原始長度一小時的影片，往往需要花費超過三、四個小時以上的時間去剪輯，即使有能力去使用這些軟體，卻沒有足夠的時間與心力去處理。最後，大多數擁有數位攝影機的使用者，都會留下一堆拍攝後尚未處理的母帶。然而，無法管理並善加利用的數位內容，我們認為本質上等同於不存在的數位內容。在本篇論文中，我們基於一些內容知覺的機制（Content-aware Mechanisms），完成一個應用於家庭視訊影片，音樂導向的視訊摘要系統。在論文中我們探討了許多音訊/視訊的特徵，以用來輔助分析輸入的音訊和視訊資料，並使用這些分析後的描述資料，將輸入的音訊和視訊資料結合成我們的音樂影片。音訊與視訊的結合是將視訊的節奏搭配音訊的節奏而完成。

2. 研究目的

根據經驗，人們對於沒有劇情、沒有旁白，以及沒有配樂的影片最缺乏耐心，而這些正是一般家庭影片所具備的特色，也因為這些原因導致未經過剪輯的家庭影片不但無趣且乏善可陳。另一項在 MIT 的研究報告也指出，使用者對於同樣的影片，搭配不同的配樂，較好的配樂會讓人們覺得影片的內容較佳，儘管都是同一段影片。此外，我們認為，在影片與配樂上做良好的搭配（讓影片配合音樂的節奏作剪輯），可以收到顯著的加成效果。

因此，在解決家庭影片前述所提到的問題上，我們提出一個全自動化的視訊摘要系統，允許使用者選擇自己喜愛的音樂，讓系統根據音樂的節奏，以及影片本身的重要性，配合音樂的長度自動化摘要的剪輯出一段音樂性的影片（像電影的預告片一樣）。配合著音樂的節奏，且加入適當的轉場特效，剪輯出具有專業質感的影片。也由於整個過程是全自動的，可以讓使用者免除複雜的視訊剪輯軟體學習過程。

3. **研究方法**

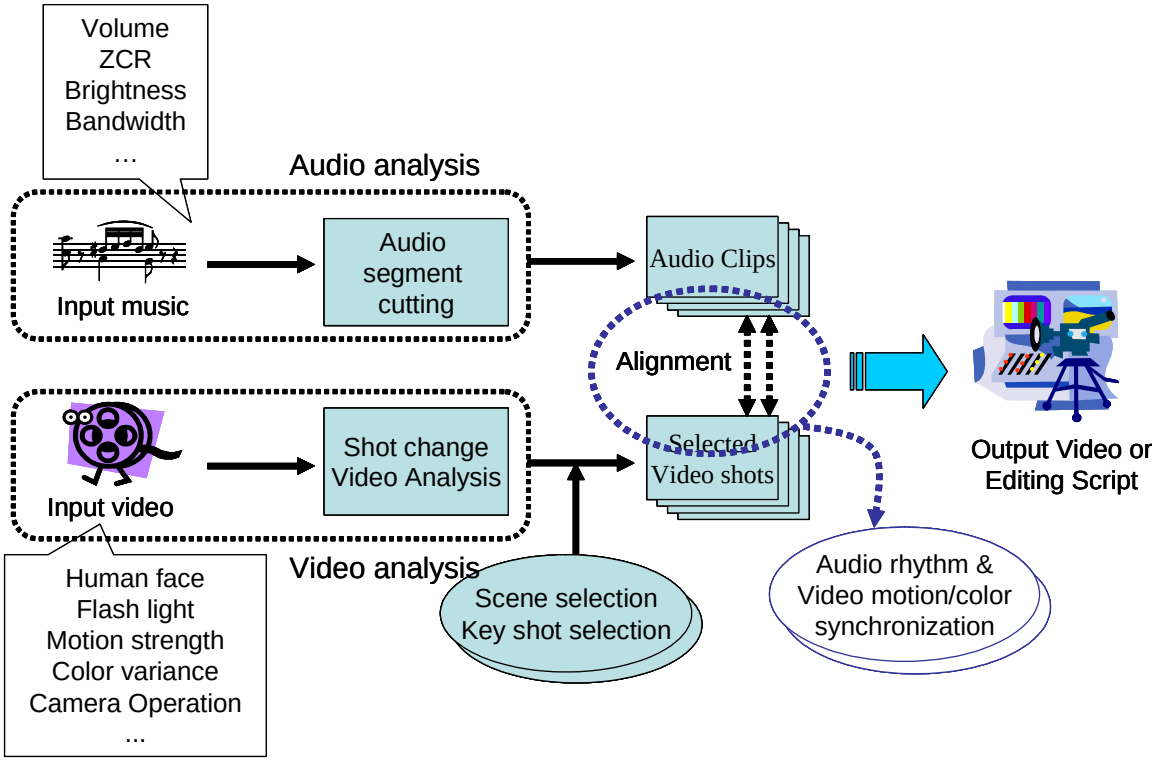
首先我們提出一個通用性的系統架構（如圖十）。在這個架構下，我們將整個流程細分為三個部分，列舉如下：

[1] 媒體分析（Media Analysis）：這部分包含針對輸入的家庭影片以及所選擇的配樂進行分析，利用媒體內容分析領域所發展出來的一些方法及工具，分析出影片中的重要片段，以及配樂裡的音樂節奏。

[2] 媒體同步結合（Media Synchronization）：利用前一部份所得到的分析資訊，我們提出兩種不同的演算法，配合參數的設定，將選出影片重要片段配合音樂的節奏加以剪輯，並加上適當的轉場特效。

[3] 媒體產出（Media Production）：第二部分中所產生的僅是存在電腦記憶體中的結構化資訊，在媒體產出的這部分，我們可以針對不同的需求，使用編輯腳本語言（Editing Scripting Language）直接輸出成影片—全自動化的產出方式。也可以輸出成目前市面上視訊編輯軟體的專案格式，再由這些編輯軟體進行更細部的編輯動作—此為半自動化的產出方式。

在本論文中，我們使用 Microsoft DirectShow Editing Service 作為全自動化的編輯腳本語言；而使用訊連科技 CyberLink PowerDirector 視訊編輯軟體的專案格式作為輸出，並使用 PowerDirector 作為半自動化的細部編輯軟體。

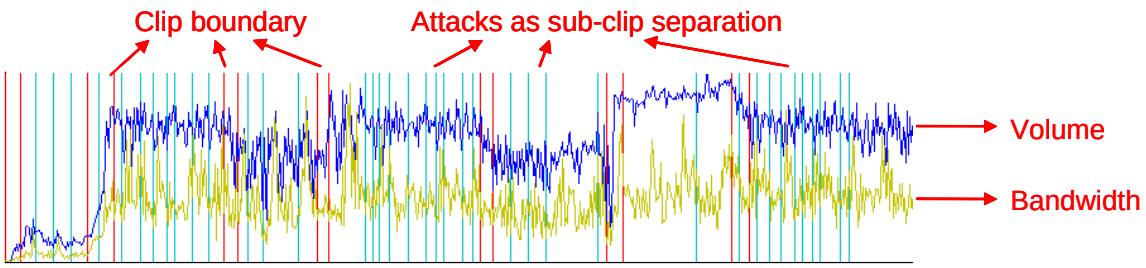


圖(十). 自動化音樂性家庭影片的製作系統架構

以下我們針對主要部分的細節分項說明：

媒體分析（Media Analysis）

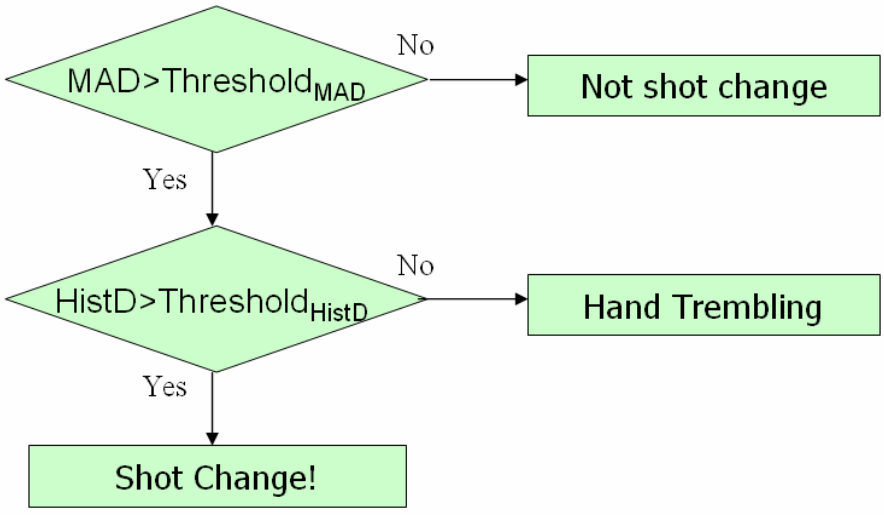
音訊分析：在音樂分析方面，首先我們根據音訊的音量（Volume）與過零率（ZCR, Zero Crossing Rate）將音訊進行切割。我們利用音量的顯著變化，切割出音訊的主要段落，再利用 ZCR 的突升（Peak）與突降（Valley）特徵部分，作為偵測音訊打擊點（Attack）的依據。依此方式我們可將輸入的音訊分析如圖十一：



圖(十一). 輸入音訊根據 Volume 與 Bandwidth 資訊進行分析之後切割出的結果

此外，我們利用單一音訊片段的 ZCR 密度，定義出音訊片段的動態性(Audio Dynamic)，以此資訊作為偵測不同節奏音樂的依據。

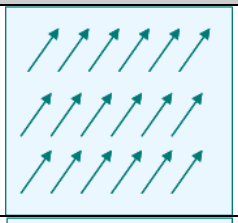
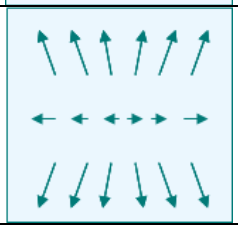
視訊分析：首先我們進行視訊的場景偵測（Shot Change Detection），我們採用混合像素式的最小絕對值差（Pixel-based Minimal Absolute Difference）與統計式的最小絕對值差（Histogram-based Difference）來做為我們偵測場景變換的方式（如圖十二）。根據我們試驗的結果，基於家庭影片的單純性，我們採用的場景偵測方式可以達到 90% 以上的正確性。



圖(十二). 混合式的場景變換偵測流程圖

此外，鏡頭動作（Camera Operation）例如伸縮（Zoom）與平移（Pan）廣泛出現在一般的家庭影片，所以我們也利用基於 MPEG 編碼的區塊移動向量（Block-based Motion Vector）資訊，來進行鏡頭動作的偵測（如表二）。區塊移動向量這樣的方法不但快速，而且有相當程度的準確性。

表(二). 分析區塊移動向量來進行鏡頭動作的偵測

	Motion Vectors	Magnitude of Average Motion Vectors	Average Magnitude
Pan		10 units	10 units
Zoom		0 units	8 units

然後，我們將所使用的視訊特徵，分成三種不同的層級：高階、中階與低階。在我們的分類上，高階的特徵表示最重要以及使用者最想看的片段。中階特徵表示較具活潑性與存在某些重點的畫面。低階特徵則是作為不良畫面的過濾作用。各種層級的特徵以及相對應的描述請見表三。

表(三). 高階、中階與低階視訊特徵與其描述

高階特徵	人臉資訊（Human Face）：偵測畫面是否有人物出現，通常家庭影片中人們最想看的部分幾乎都是人物的部分。 人聲事件（Speech）：偵測是否有人在講話。 閃光燈事件（Flashlight）：閃光燈通常來自於照相機的閃光，通常某些具有相當重要意義的場景會出現連續的閃光燈事件。
中階特徵	移動向量強度（Motion Vector Strength）：移動向量強的畫面較具有活潑性，人們觀看的時候比較喜愛動作變化顯著的畫面。 鏡頭動作種類（Camera Operation Type）：通常家庭影片中出現的鏡頭動作，都具有某種程度的語意意義，例如正在追蹤某個物體(Pan)，或是想要強調某個物體（Zoom）。
低階特徵	畫面亮度（Brightness）：在某些情況下，曝光不足或過當的畫面常出現在家庭影片中，過濾掉這些不良畫面，選擇曝光適當的畫面。 畫面彩度（Color Variance）：過濾掉色彩較不豐富的畫面，保留色彩豐富的畫面。

結合這些視訊特徵，我們提出兩種不同層級的重要性函數，分別是畫格層級的重要性函數（Frame-level Importance Function）以及場景層級的重要性函數（Shot-level Importance Function）。

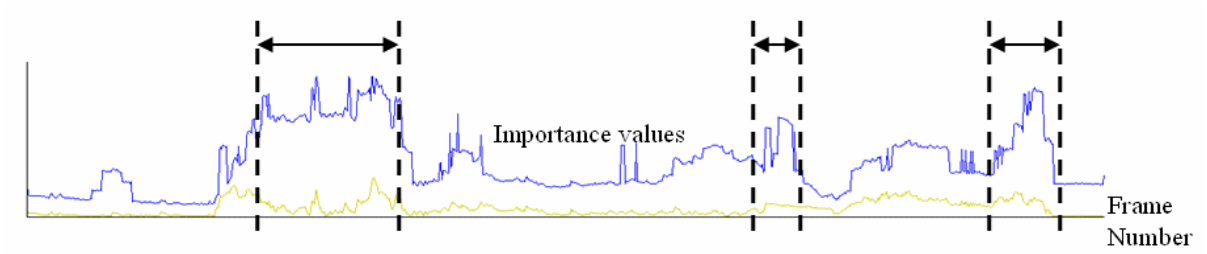
畫格層級的重要性函數：

$$Imp = Coeff_H \times (R_{face} + E_{flash}) + Coeff_M \times (R_{motion} \times S_a + CameraType) + Coeff_L \times (\left|\frac{Mean - 130}{255}\right| + R_{ClrVar})$$

場景層級的重要性函數：

$$Imp = Len \times (\frac{Num_{face}}{Len} + \frac{\sum CameraType}{Len}) \times (\frac{\sum Motion}{Len}) \times (\frac{\sum Heterogeneity}{Len}),$$

我們根據畫格層級的重要性函數來選出重要的視訊片段（如圖十三）。場景層級的重要性函數則是用來評估不同場景的重要性，在媒體同步結合階段給予較重要的場景較高的權重，使其能保有較多的視訊細節。

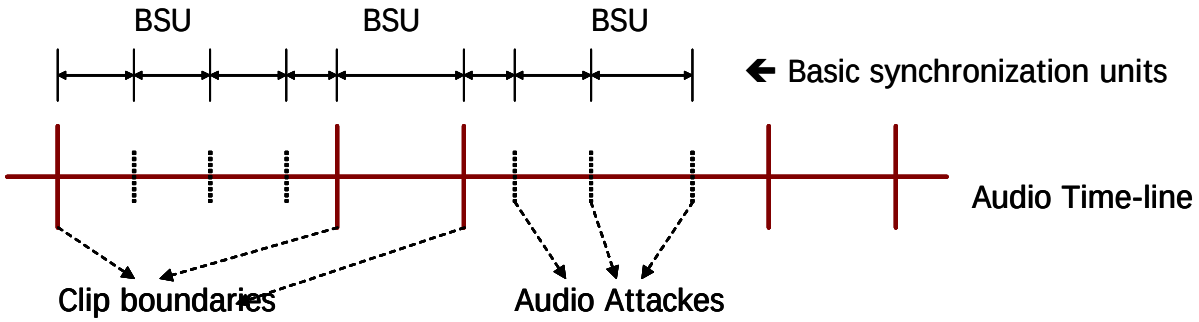


圖(十三). 選取較高的畫格層級重要性值來作為視訊摘要

媒體同步結合（Media Synchronization）

針對媒體的同步結合，我們提出兩種不同的組態方式，分別是**韻律的（Rhythmic）**以及**中規中矩的（Medium）**。

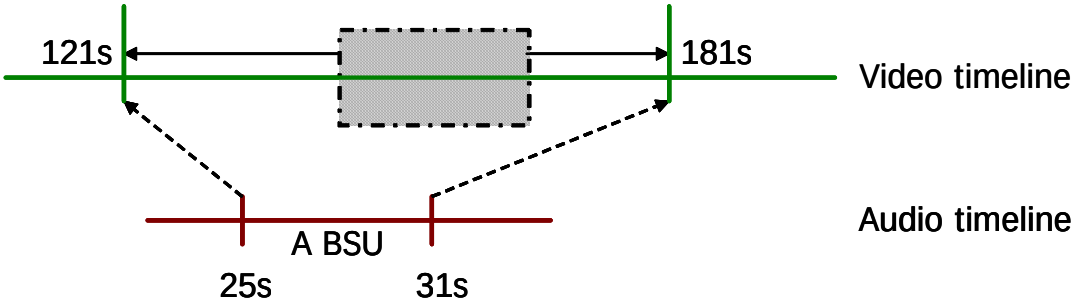
首先我們定義音訊的基本同步單位（BSU, Base Synchronization Unit）以及 LBSU（Larger BSU）（參見圖十四）。我們以此音訊的基本同步單位作為將來剪輯時同步的最基本單位。因為音訊的片段是根據音樂的節奏所分割出來的，所以我們所剪輯出來的片段可以搭配音樂的節奏。



圖(十四). 音訊的基本同步單位

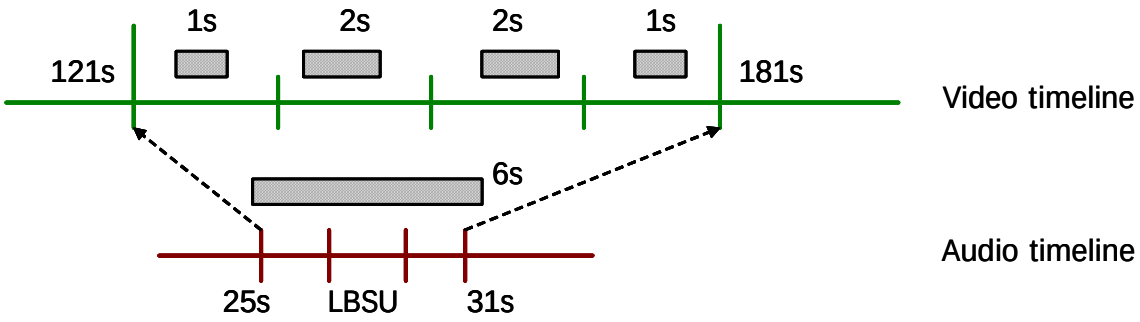
根據輸入影片與音樂的長度，我們可以計算出視訊縮減比率（Video Reduction Rate），再依照視訊縮減比率將音訊的 BSU 區間反投影回原本的視訊時間軸區間，在該區間內搜尋一段畫格層級重要性最高的片段，選擇該片段作為最後產出的影片片段，此即為韻律的（Rhythmic）組態方式（參見圖十五）。

『韻律的』本意即為『以音樂為主』，因為其使用的 BSU 為較小的單位，所以剪輯出來的影片隨著音樂節奏而改變的效果比較顯著，但卻有可能因為搭配音樂的關係，遺漏某些場景或是影片遺失調部分細節。因此，我們提出另一種中規中矩的（Medium）組態方式。



圖(十五). 韻律的（Rhythmic）組態方式同步示意圖

在中規中矩的（Medium）組態方式裡，我們採用較大的 LBSU 作為基本單位，因為其基本單位較大，在反投影回原本的視訊時間軸上，會覆蓋到較多的場景（如圖十六），在這樣的情況下，我們根據之前所提出的場景層級重要性函數為每一場景配給所應得的剪輯長度，依據每一場景的重要性。



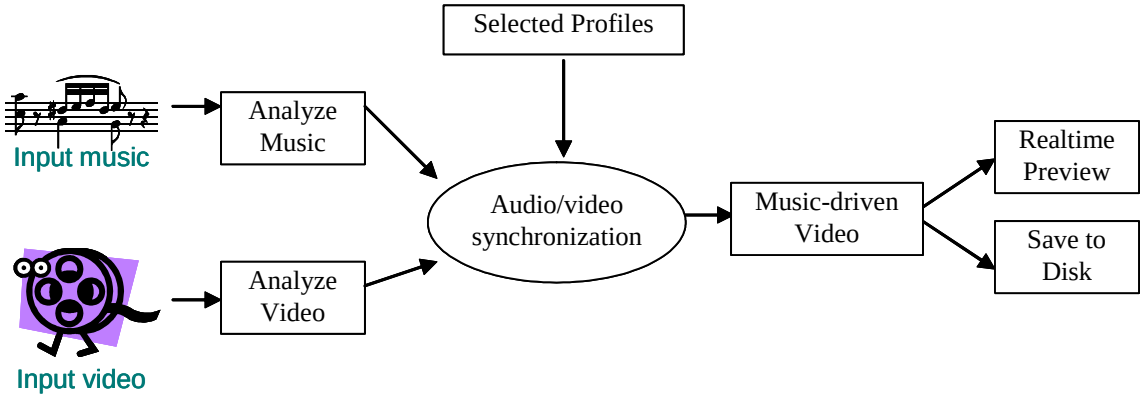
圖(十六). 中規中矩的（Medium）組態方式同步示意圖

在中規中矩的（Medium）組態方式裡，因為其考量到每一個場景以及其相對的重要性，所以比較不會遺失影片中的場景，然而因為其使用的是較大單位的 BSU 所以在音樂的搭配上會較韻律的（Rhythmic）來的比較不明顯。

簡單的說，這兩種不同的組態方式是讓使用者得以用簡單的方式，選擇以音樂為主，或是以影片本身為主的剪輯方式，給予使用者全自動化過程中部分的彈性。

4. 研究結果

本論文中所提出的音樂導向的視訊摘要系統，其流程可以簡化為圖十七所示。在本論文中，我們收集六種不同種類的家庭影片，包含畢業典禮、全家出遊、結婚典禮等等影片，針對二十位使用者（十位具有電腦學科背景，其餘無）進行客觀性測試。我們試驗所提出的兩種組態方式的差異性（表四）；以及相同的影片配上不同的配樂（一首快歌和一首慢歌）所呈現出來的不同效果，用來試驗影片與音樂之間的搭配關係（表五）；最後我們讓使用者自由操作我們的系統，選擇他們想看的影片以及喜歡的配樂，試驗他們對本系統的整體滿意度（表六）。試驗的方式都是透過相關的問題，讓使用者依照五種不同的程度選擇回答。



圖(十七). 音樂導向的視訊摘要系統簡化流程圖

表(四). 試驗使用者對於兩種組態之間的差異性

問題	非常同意	同意	無意見	不同意	非常不同意
請問你覺得 Medium 組態方式與 Rhythmic 組態方式之間的差異明顯嗎？	10%	50%	40%	0%	0%
請問你是否覺得在某些影片中，選擇 Medium 會有較佳的效果？	20%	80%	0%	0%	0%

表(五). 試驗使用者對於同樣影片搭配不同音樂的差異性

問題	非常同意	同意	無意見	不同意	非常不同意
請問在這兩種不同配樂的剪輯結果中，你覺得影片的剪輯方式與音樂節奏搭配得非常好嗎？	20%	60%	20%	0%	0%
你認為後者(搭配較快配樂的影片)所呈現出來的效果較前者更具動態性(較快的節奏，較多的動作）嗎？	90%	10%	0%	0%	0%

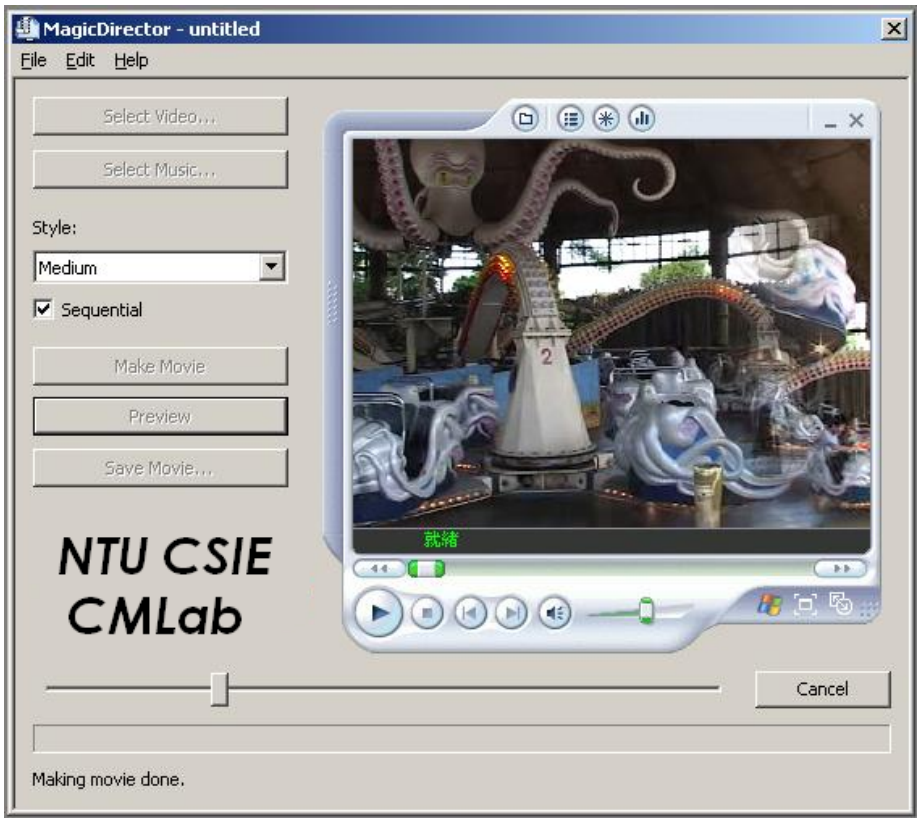
表(六). 試驗使用者對本系統的整體滿意度

問題	非常同意	同意	無意見	不同意	非常不同意
你認為摘要後的影片能將原本的原始影片描述得很好嗎？	55%	45%	0%	0%	0%
你認為這樣的自動化視訊音樂性影片摘要系統對於家庭使用者來說，在處理家庭影片上有著很大的幫助？	50%	50%	0%	0%	0%
你認為配上音樂的家庭影片在感官呈現上有比較好的呈現效果嗎？	100%	0%	0%	0%	0%

在這裡我們提供兩則經由我們系統全自動剪輯出來的影片，原始長度分別為 32 分鐘與 54 分鐘，配上適當的音樂之後剪輯成 4-5 分鐘左右。分別是：

- 六福村之旅
  http://www.cmlab.csie.ntu.edu.tw/~chenhsiu/research/6travel.mpg
- 結婚典禮
  http://www.cmlab.csie.ntu.edu.tw/~chenhsiu/research/marriage.mpg

最後，由於本研究中所實作出來的系統較貼近一般使用者，所以在效能上必須不能花費太久，根據我們在 Pentium 4 2.8 GHz 的機器上測試的結果，MPEG-1 視訊影片（352x288 解析度），原始長度 31 分種的影片只需要 5 分鐘的時間即可分析完成，對於輸入的音訊，5 分鐘 34 秒的 MP3 檔案只需要花費 11 秒的時間即可。這樣的效能對於消費端的應用程式而言是可以接受的。



圖(十八). 本論文中所發展的音樂導向的視訊摘要系統

第三年預期成果與具體成果之比較表詳列於下：

預期成果	實際成果	說明
具結構性之內容分析與摘要模組	內容分析與摘要模組整合於**以音樂導向的視訊摘要系統**中	符合。
二維/三維搜尋器之查詢方式及參數轉換	整合於總計畫中的**家庭媒體中心**(Home Media Center)中	符合。
內容意義察覺之資訊串流	完成**以使用者注意模型為基礎之視訊興趣區自動決定系統**之系統	符合。
融合業界最新技術，進行系統之錯誤更正與效能調整	完成。	符合。
與其他分項計畫整合，實作出總計畫第三年之應用系統	完成。	符合。
撰寫技術手冊與使用手冊	完成。	符合。
落實研發成果至合作廠商	完成**應用內容知覺機制於音樂導向的視訊摘要系統**	該系統以及其相關的演算法已被訊連科技 CyberLink 採納成為其核心技術。 此項技術已規畫為視訊剪輯商業產品 MagicDirector 以及應用於現有產品 PowerDirector 中的新技術 MagicCut 等實際商業應用。

四、相關論文完成

[1] 黃振修：應用內容知覺機制於音樂導向的視訊摘要系統（A Musical-driven Video Summarization System Using Content-awarded Mechanisms）- 碩士論文
[2] 沈至豪：同時保障買賣雙方權益的數位影像權利管理系統（A Digital Image Rights Management System which assuring both buyers' and sellers' rights）- 碩士論文
[3] 陳萱瑋：以節奏分析為基礎的動作電影分割與摘要（Action Movies Segmentation and Summarization Based on Tempo Analysis）- 碩士論文
[4] 鄭文皇：利用使用者注意模型決定視訊之興趣區（User Attention Model in Region-of-Interest Determination on Videos）- 碩士論文
[5] 葉人豪：新聞與運動節目中的廣告偵測（TV Commercial Detection in News and Sports Program Videos）- 碩士論文
[6] 梁致豪：棒球影片事件偵測（Semantic Event Detection in Baseball Videos）- 碩士論文
[7] 潘廷建:於因子分解圖形上解決積和問題的有效演算法（An Efficient Sum-Product Algorithm on Factor Graphs）- 碩士論文

五、國際期刊與國際會議論文

國際期刊論文

[1] Jin-Hau Kuo, Chia-Chiang Ho, Kan-Li Huang, Jim Shiu, and Ja-Ling Wu, “A Low-cost Media-Processor based Real-Time MPEG-4 Video Decoder,” IEEE Transactions on Consumer Electronics, vol. 49, no. 4, Nov. 2003, pp. 1488-1497.

[2] Chia-Chiang Ho, Jin-Hau Kuo, and Ja-Ling Wu, “Building MPEG-7 Transcoding Hints from Intrinsic Characteristics of MPEG Videos,” IEEE Transactions on Consumer Electronics, vol. 50, no. 1, Feb. 2004, pp. 306-311.

國際會議論文

[1] Jin-Hau Kuo, Chia-Chiang Ho, M-J Shieh, J-H Yeh, and Ja-Ling Wu, “An MPEG-4/7 based Architecture for Analyzing and Retrieving News Video Programs,” in Proceedings of IEEE International Conference on Consumer Electronics (ICCE’2003).

[2] Jin-Hau Kuo and Ja-Ling Wu, “An MPEG-7 Content-based Analysis/Retrieval System and Its Applications,” in Proceedings of SPIE International Conference on Visual Communications and Image Processing (VCIP ‘03), Lugano, Switzerland, July 2003..

[3] Chia-Chiang Ho, Wei-Ta Chu, Chen-Hsiu Huang and Ja-Ling Wu, "User-Oriented Approach in Spatial and Temporal Domain Video Coding," Proceedings of the 4th IEEE Pacific-Rim Conference on Multimedia, 2003.

[4] Wen-Huang Cheng, Wei-Ta Chu and Ja-Ling Wu, "Semantic Context Detection based on Hierarchical Audio Models," Proceedings of the 5th ACM SIGMM International Workshop on Multimedia Information Retrieval, 2003.

[5] Chia-Chiang Ho, Wen-Huang Cheng, Ting-Jian Pan, Ja-Ling Wu, “A User- Attention Based Focus Detection Framework and Its Applications,” International Conference on Information, Communications & Signal Processing Fourth IEEE Pacific-Rim Conference On Multimedia (ICICS-PCM’2003).

[6] Jin-Hau Kuo, Sung-Wen Wang, and Ja-Ling Wu,“ A MPEG-7 Contour-based Analysis /Retrieval System for Fish Images”, 2003 ICICS.

[7] Yuh-Jue Chuang, Ja-Ling Wu, “An SGH-Tree Based Efficient Huffman Decoding,” Fourth International Conference on Information, Communications & Signal Processing Fourth IEEE Pacific-Rim Conference On Multimedia, 2003.

[8] Yuh-Jue Chuang, Ja-Ling Wu, “Direct Splitting and Merging of 2-D DCT in the DCT Domain,” International Conference on Informatics, Cybernetics, and Systems, 2003.

[9] Chia-Wei Lin, Yuh-Jue Chuang, Ja-Ling Wu, “A Generic Algorithm for Constructing Reversible Variable Length Codes with Limited Maximum Codeword Length,” International Conference on Informatics, Cybernetics, and Systems, 2003.

[10] Chen-Hsiu Huang, and Ja-Ling Wu, "A User Attention Based Visible Video Watermarking Scheme," International Conference on Informatics, Cybernetics and

Systems, 2003
[11] Wei-Ta Chu, Wen-Huang Cheng, Ja-Ling Wu, and Hsu, Y.-J., "A Study of Semantic Context Detection by Using SVM and GMM Approaches," IEEE International Conference on Multimedia & Expo, 2004.
[12] Sung-Wen Wang, Ya-Ting Yang, Chia-Ying Li, Yi-Shin Tung, Ja-Ling Wu, “An optimization of H.264/AVC baseline decoder on low-cost TriMedia DSP processor,” SPIE’2004.
[13] Sheng-Ho Wang, Sung-Wen Wang, Yi-Chin Huang, Yi-Shin Tung, Ja-Ling Wu, “Boundary-energy sensitive visual de-blocking for H.264/AVC coder,” SPIE’2004.
[14] Wei-Ta Chu, Wen-Huang Cheng, Sheng-Fang He, Chia-Wei Wang, and Ja-Ling Wu, “A Unified Framework Using Spatial Color Descriptor and Motion-based Post Refinement for Shot Boundary Detection,” accepted by the fifth Pacific-Rim Conference on Multimedia, 2004.
[15] Hsuan-Wei Chen, Jin-Hau Kuo, Wei-Ta Chu, and Ja-Ling Wu, “Action Movies Segmentation and Summarization Based on Tempo Analysis,” accepted by the ACM SIGMM International Workshop on Multimedia Information Retrieval, 2004.
[16] Wei-Ta Chu, Wen-Huang Cheng, and Ja-Ling Wu, “Generative and Discriminative Modeling toward Semantic Context Detection in Audio Tracks,” accepted by the 11th International Multimedia Modelling Conference, 2005.

六、參考文獻

其餘參考文獻請參照分項計畫二計畫書參考文獻。

七、相關論文成果 (見後頁)

Building MPEG-7 Transcoding Hints from Intrinsic Characteristics of MPEG Videos

Chia-Chiang Ho, Student Member, IEEE, Jin-Hau Kuo, and Ja-Ling Wu, Senior Member, IEEE

Abstract —*MPEG-7 transcoding hints are defined to provide better quality or reduced complexity of transcoding applications. This paper describes application scenarios, hierarchies, and generation methods of transcoding hints that are based on intrinsic characteristics of MPEG videos.*[1].

Index Terms — MPEG-7 Transcoding Hints, MPEG characteristics.

I. INTRODUCTION

MPEG-7 defined media transcoding hints aiming at reducing complexity and improving the quality of general transcoding processes [1]. The basic assumption is that transcoding hints are generated in advance, and are stored together with the host video itself (or stored separately but with a proper linkage to the video). The possible applications of transcoding hints include but not limited to constant bitrate (CBR) to variable bitrate (VBR) conversion, encoding mode decision, frame rate and bitrate control, complexity reduction for motion re-estimation and video skimming and browsing.

Collected as one MPEG-7 description scheme (MediaTranscodingHint), transcoding hints defined in MPEG-7 include:

(i) *Motion Hints*:

MotionRange, consisting of four values (*xLeft*, *xRight*, *yDown* and *yUp*), indicates the recommended integer pel search range for conducting motion estimation to the left, right, bottom and top of a video frame, respectively;

Intensity describes the motion intensity, specifically, the variance of motion field, of a video segment;

Uncompensability describes the amount of new content in a video segment.

(ii) *Shape Hints*:

ShapeChange describes the amount of shape change (of a moving region) in a video segment;

NumOfNonTranspBlocks describes the average number of 16×16 blocks per frame containing at least one pixel with a non-zero alpha-map value. (Since shape hints are metadata used for object-based video coding and thus are not considered in this paper.)

(iii) *Coding Hints*:

AvgQuantScale describes the average quantization scale used to compress a video segment;

IntraFrameDistance describes the distance between intra-coded frames (I-frames);

AnchorFrameDistance describes the distance between anchor frames (I- or P-frames);

Difficulty describes the transcoding difficulty of one video segment, which is a normalized value (between 0 and 1), by taking video segments into consideration;

Importance describes the importance of one media segment with respect to the whole semantic media content. It is also a normalized value as the *Difficulty* hint does;

SpatialResolutionHint describes the maximum allowable spatial resolution reduction factor for preserving perceptibility.

Generality and flexibility are two central goals of MPEG-7 [2]. In MPEG-7, both generation and consumption of transcoding hints are not formally defined. The definition of "video segment" is also left for real applications. Focusing on MPEG videos, in this paper, we (1) clarify multimedia application scenarios within which transcoding hints are expected to be generated; (2) define practical transcoding hint hierarchies from both structural and semantic perspectives; (3) propose a compressed domain transcoding hint generator which takes advantages of intrinsic, low-level characteristics of MPEG videos. Our work provides a clear roadmap for implementations embedded on consumer electronics or multimedia softwares.

II. APPLICATION SCENARIOS AND HIERARCHIES OF TRANSCODING HINTS

Before devising hierarchies and generation schemes of transcoding hints, we would like to look into today's multimedia applications and figured out scenarios within which transcoding hints could be generated dependently. The success of MPEG-series standards makes MPEG videos scattering around the world. These videos reside in content archives of content providers, VCD and DVD disks, and personal collections (e.g., home videos), etc. Moreover, due to the widespread and cost reduction of digital camcorders, more MPEG videos are generated day by day. The generation of corresponding transcoding hints makes distribution and sharing of MPEG videos more easily and efficiently, even in today's heterogeneous network and computation surroundings. We classify MPEG video related applications into three different scenarios described as follows.

[1] This work was supported in part by the National Science Council, R.O.C., under the contract NSC91-2622-E-002-040 and 89-E-FA06-2-4-8.

The authors are with the Communications and Multimedia Laboratory, National Taiwan University, Taipei, Taiwan (email: {conrad, david, wjl}@cmlab.csie.ntu.edu.tw).

Contributed Paper
Manuscript received November 27, 2003

Scenario 1-- Raw Input, Compressed Output:
This scenario refers to those applications involving MPEG encoding. Examples are professional studio encoders, personal video recorders (PVR) and software MPEG encoders, etc.

Scenario 2-- Compressed Input, Compressed Output:
This scenario refers to those applications involving MPEG video processing. Examples include video post-processing, video editing and video content analysis for indexing, browsing, and/or searching. These applications take MPEG videos as inputs, and generate metadata and/or modified MPEG videos as outputs.

Scenario 3-- Compressed Input, Raw Output:
This scenario refers to those applications performing MPEG video decoding. Both hardware and software decoders, which decode MPEG videos coming from network or local storages, belong to this scenario.

With such application classifications in mind, we defined generic transcoding hint hierarchies for MPEG videos in an entity-relationship fashion, as shown in Fig. 1. Importance hint and spatial resolution hint are semantically determined through user interactions. The transcoding hint generator is responsible for generating other hints. Two styles of video representation are proposed. The *sequence-scene-frame* style coincides with human understanding of videos, and is the semantic one. The *sequence-GOP-frame* style coincides with internal structure of MPEG bitstreams, and is thus structural one. These two styles of video representation may help for improving efficiency and accuracy of transcoding in both systematic and user-oriented ways. The availabilities of different hints depend on the generation methods used in our transcoding hint generator, which will be illustrated later.

Another way to look at these hierarchies comes from the classification of hints, as given in Table I. Such classification affects the relationship between transcoding hints and video representations. For example, motion hints present both structural and semantic meanings. Associating motion hints with the structural video representation helps for those transcoders requiring re-estimation of motion vectors, while associating motion hints with semantic video representation helps for transcoders thinking semantically. This is because motion hints themselves somewhat stand for intrinsic importance of each individual scene, rather than those importance values set externally by the user (maybe the content provider).

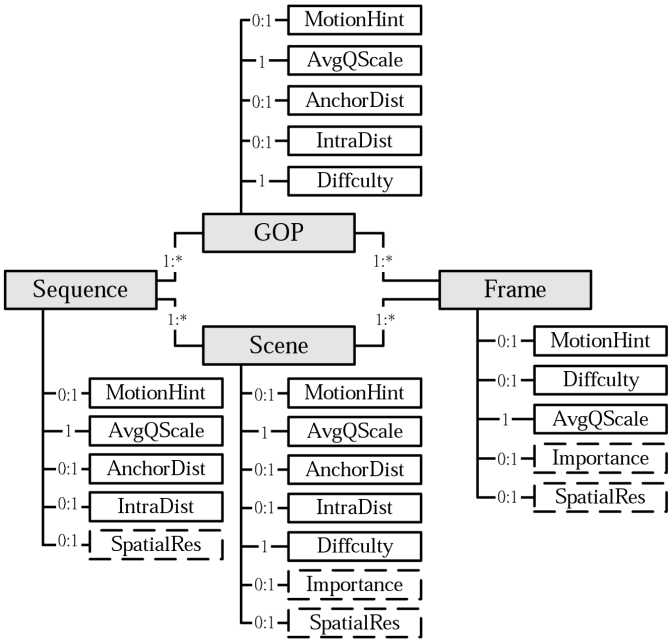


Fig. 1: The proposed transcoding hint hierarchy for MPEG videos. The names of hints have been abbreviated for clarity.

The proposed video representations and associated transcoding hint hierarchies can be easily implemented in XML forms and by using the MPEG-7 defined binary representations.

III. The Proposed Compressed Domain Transcoding Hint Generator

Applications belonging to the aforementioned scenarios 2 and 3 process the MPEG compressed videos for various usages, such as playback, browsing, editing and transcoding. For these applications, we proposed a compressed domain transcoding hint generator that is built upon an MPEG decoder, as shown in Fig. 2. All information required for generating video representations and transcoding hints is available just after performing the VLC decoding module. That is, no de-quantization, IDCT or motion compensation operations are involved, and thus the generator operates very efficiently.

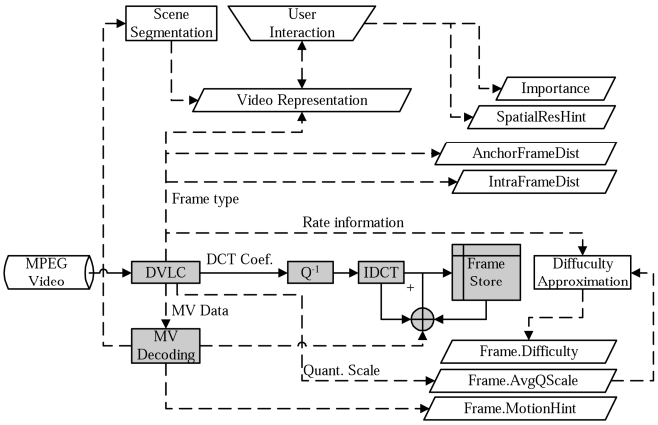


Fig. 2: The architecture of the proposed compressed domain transcoding hint generator. The gray-shaped modules constitute a typical MPEG decoder.

Table I: Classifications of transcoding hints

		Structural	Semantic
Motion Hints		Yes	Yes
Coding Hints	AvgQuantScale	Yes	Yes
	IntraFrameDistance	Yes	No
	AnchorFrameDistance	Yes	No
Difficulty		Yes	Yes
Importance		No	Yes
SpatialResolutionHint		No	Yes

The scene segmentation module divides the entire video sequence into separate scenes. We adopt a scene change detection scheme based on macroblock type statistics in B-frames [3]. The structural video representation is figured out by inspecting frame coding types and GOP information embedded in the bitstream. Then, transcoding hints belonging to individual frames, GOP's, scenes and the whole video are generated using the algorithms described below.

A. Algorithms for generating different transcoding hints

Motion range hint: For each predicted frame, we firstly check if its reference frames are in the same scene. If not, the motion range hint is not calculated for this frame. Otherwise, the maximal and minimal values of forward and backward motion vectors are figured out and denoted as Min_{x_F}, Max_{x_F}, Min_{y_F}, Max_{y_F}, Min_{x_B}, Max_{x_B}, Min_{y_B}, Max_{y_B}, respectively. For P-frames, Min_{x_B}, Max_{x_B}, Min_{y_B} and Max_{y_B} are set to zero. To account for different frame distances between the predicted frame and its reference frames, the above values are normalized as:

$$\begin{aligned} NMax_{x_F} &= Max_{x_F} / FD, \\ NMin_{x_F} &= Min_{x_F} / FD, \\ NMin_{y_F} &= Min_{y_F} / FD, \\ NMin_{y_F} &= Min_{y_F} / FD, \\ NMax_{x_B} &= Max_{x_B} / BD, \\ NMin_{x_B} &= Min_{x_B} / BD, \\ NMin_{y_B} &= Min_{y_B} / BD, \\ \text{and} & \\ NMin_{y_B} &= Min_{y_B} / BD, \end{aligned} \tag{1}$$

where *FD* is the frame distance between the current frame and the forward reference frame, and *BD* is the frame distance between the current frame and the backward reference frame. The frame-level motion range hint is computed as:

$$\begin{aligned} xLeft &= -MAX\left(NMin_{x_F}, NMin_{x_B}\right), \\ xRight &= MAX\left(NMax_{x_F}, NMax_{x_B}\right), \\ yUp &= -MAX\left(NMin_{y_F}, NMin_{y_B}\right), \\ \text{and} & \\ yDown &= MAX\left(NMax_{y_F}, NMax_{y_B}\right). \end{aligned} \tag{2}$$

For the whole video, one GOP or one scene, the corresponding motion range hint is figured out by taking the maximal value of available frame-level motion range hints of all frames in the range.

Motion Uncompensability hint: For each predicted frame, we calculate the ratio of intra-coded macroblocks over total macroblocks. This is based on the general rule of typical MPEG encoders, that is, choosing intra mode when the least error of motion compensation is still too large. However, the definition of Uncompensability is not precise in the frame level. Owing to different frame distances and different coding modes, different frames have different "capabilities" of motion compensation in nature. So, we calculate the Uncompensability hint only for higher levels. For a GOP, a scene, or the whole video, the value of the Uncompensability hint is the average of the predefined ratio values of each predicted frame in the range.

Motion intensity hint: For each P-frame, we calculate the average intensity of forward motion vectors. For each B-frame, average intensities of both forward and backward motion vectors are calculated. Similar to the generation of the motion range hint, the frame distance normalization is also performed here. For B-frames, the average of normalized forward and backward average motion intensities is then calculated. The motion intensity of one GOP or one scene is the average of motion intensities of all frames in the range. Finally, a sequence-wide normalization is performed such that the motion intensity value of the most active GOP or scene is 1.0.

AvgQuantScale hint: The calculation of AvgQuantScale hint for each frame is straightforward. For each frame, the AvgQuantScale is set to be the average of quantization scale values of all coded macroblocks. For one GOP, one scene or the whole video, the AvgQuantScale value is the average of quantization scales of all frames in the range.

IntraFrameDistance and AnchorFrameDistance hint: These two hints are figured out for every GOP or scene, by checking frame coding patterns. If the coding pattern is not regular, they are just not available. If IntraFrameDistance or AnchorFrameDistance is not consistent over the entire video, it is not available in the sequence level.

Difficulty hint: We approximate coding difficulty of each frame as

$$Difficulty = Consumed\_Bitcount * AvgQuantScale. \tag{3}$$

GOP-wide, scene-wide and sequence-wide normalizations are then performed to calculate Difficulty values of all GOP's and scenes.

IV. EXPERIMENTAL RESULTS

To validate the applicability of the proposed transcoding hint generator, we cascaded four novel test sequences, namely, "container", "mobile", "foreman" and "mother and daughter" into a single video sequence with totally 1200 frames (each scene is of 300 frames). Fig. 3 shows sample snapshots of these four scenes. Then MPEG-1 encoding was performed with the following settings: The number of frames in a GOP is 30 (so each scene spans ten GOP's); the I/P frame distance is 3; the bitrate is 1,500,000 bps; the motion search ranges are ±31 pixels. Let we call the generated MPEG video as "4Seq". Table II shows the summarization of generated scene-level transcoding hints.

Fig. 3: Sample snapshots of the four scenes in the sequence "4Seq".

Table II: The summarization of scene-level transcoding hints of the sequence "4Seq".

		container	mobile	foreman	mother and daughter
Motion Hints	xLeft	31	31	31	31
	xRight	31	31	31	31
	yDown	31	31	31	31
	yUp	31	31	31	31
	Uncomp	0	0	0	0
	Inten	0.089	0.337	1.0	0.398
Coding Hints	avgQ	2.69	12.8	4.72	2.03
	intraF	30	30	30	30
	anchorF	3	3	3	3
Difficulty		0.199	1.0	0.368	0.170

As for the motion intensity hint, Table II reveals that the scene "foreman" has stronger motion activities than other scenes do, while the scene "container" has weakest motion activities. This coincides with the feeling of human observers. Fig. 4 shows the GOP-level motion intensity hints. The GOP with the highest motion intensity resides in the scene "foreman", at the time that the man waved his arm to the right and the camera also moved rapidly.

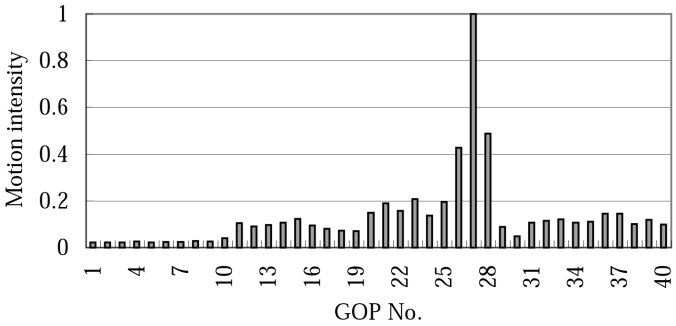


Fig. 4: The GOP-level motion intensity hints of the sequence "4Seq".

The motion range hints of the four sequences were detected to be the same, i.e. 31, for all four directions. This means that at each scene, there is, more or less, some macroblocks with motion vectors exceeding the maximal allowable values (31 here). This can be verified by taking frame-level motion range hints of the scene "container" as an example, which is shown in Fig. 5.

The IntraFrameDistance and AnchorFrameDistance are correctly detected as 30 and 3, respectively. As for the average quantization scale, the scene "mobile" has the largest value while the scene "mother and daughter" has the least one. Since the adopted encoder utilizes a constant bitrate rate control algorithm, such diverse values of average quantization scales directly correspond to the values of the Difficulty hint (generated according to (3)). From Table II, it can be seen that the scene "mobile" is the most difficult to encode, while "container" and "mother and daughter" are comparably much more easier to encode. Fig. 6 shows GOP-level difficulty hints. It can be observed that GOP's of the second half of the scene "foreman" have higher difficulty hint values than GOP's of the first half do. This is because the second half of 'Foreman" is with larger motion and more complex textures.

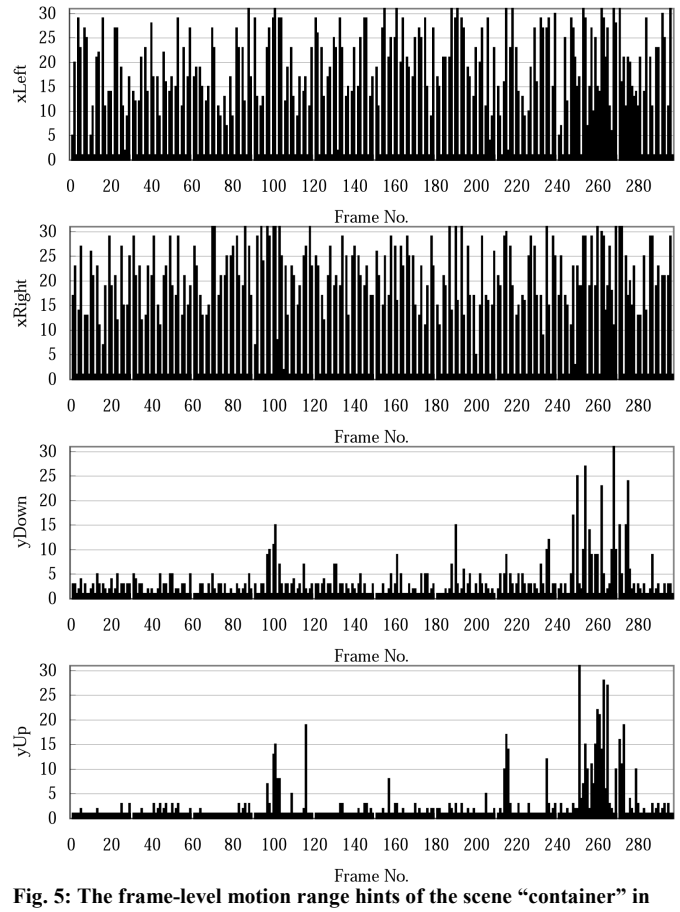


Fig. 5: The frame-level motion range hints of the scene "container" in the video "4Seq".

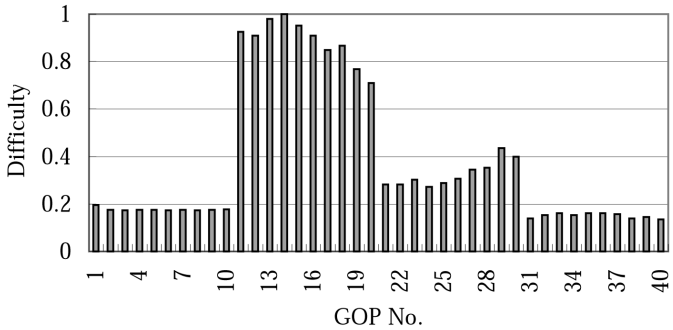


Fig. 6: The GOP-level difficulty hints of the sequence "4Seq".

V. APPLICATIONS

Due to the wide-spreading of DVD and the trend of HDTV/DTV, most existing video contents are still encoded in MPEG-2 form. Therefore, the MPEG-2-to-New video standard (e.g. AVC/H.264) transcoder or vice versa, is believed to be an indispensable functional module of future video signal treatments. As shown in [4], the main functionality addressed by the Media Transcoding Hint DS is to provide transcoding hints to support the transcoder. This transcoding hints enable the transcoder to reduce the computational complexity and to assist the transcoder in (re-) encoding parameter decisions to preserve visual quality through the transcoding process. The corresponding applications include the image resolution reduction for heterogeneous client devices (e.g. making use of importance and spatial resolution hint), video transcoder and real-time transcoder server for video streaming.

VI. CONCLUSION REMARKS

The advantages of using transcoding hints have been confirmed by experimental reports revealed in MPEG meetings. It is interesting that all transcoding hints are defined as optional, and this fact shows that the applicability of transcoding hints is very different for different media formats and target applications.

Defining transcoding hints is not to solve obstacles of transcoding, but to increase quality and/or to reduce complexity. An intelligent transcoder should have two alternative configurations: one is without the transcoding hint and the other uses transcoding hints wisely. Generating metadata to aid the task of transcoding is of course a positive idea. It is our belief that generating media-format-dependent metadata may help more for the task of transcoding than those general transcoding hints may do. The previous work done in our laboratory had transferred this idea into a practical and useful system [5].

REFERENCES

[1] Information Technology-- Multimedia Content Description Interface, Part 5 Multimedia Description Scheme, ISO/IEC 15938-5, 2001.

[2] S.-F. Chang, T. Sikora, and A. Purl, "Overview of the MPEG-7 standard," IEEE Trans. Circuits Syst. Video Technol., vol. 11, no. 6, pp. 688-695, June 2001.

[3] J.-H. Kuo and J.-L. Wu, "An efficient algorithm for scene change detection and camera motion characterization using the approach of heterogeneous video transcoding on MPEG Compressed Videos," IEEE 3rd Int. Conf. On Information, Communications and Signal Processing, Oct 2001.

[4] Anthony Vetro, Peter Kuhn, Teruhiko Suzuki, John R. Smith, Ana B. Benitez, Charilaos Christopoulos, "CE Report on the Media Transcoding Hint DS," ISO/IEC JTC1/SC29/WG11 MPEG/M6168, July 2000.

[5] P.-K. Hsiao, Y.-S. Tung, Ja-Ling Wu, K.-L. Huang and H.-S. Chen, "A frame-based MPEG characteristics extraction tool and its application in video transcoding," IEEE Trans. Consumer Electronics, vol. 48, no. 3, pp. 522-532, Aug. 2002.

Chia-Chiang Ho received the B.S. degree in computer science and information engineering from National Taiwan University, Taipei, Taiwan, in 1996. Currently, he is a Ph. D student in the same department. His research interest includes video coding, video streaming and multimedia system design.

Jin-Hau Kuo received the B.S. degree in information and computer engineering from Tatung University, Taipei, Taiwan, in 1997. He is currently a Ph.D. student in the department of computer science and information engineering, National Taiwan University. His current research interests include data compression, digital image processing, content-based retrieval, code optimization, and multimedia systems.

Ja-Ling Wu received the B.S. degree in electronic engineering from TamKang University, Tamshoei, Taiwan, R.O.C. in 1979, and the M.S. and Ph.D degree in electrical engineering from Tatung Institue of Technology, Taipei, Taiwan, in 1981 and 1986, respectively. From 1986 to 1987, he was an Associate professor of the Electrical Engineering Department at Tatung Institute of Technology, Taipei, Taiwan. Since 1987, he has been with the Department of Computer Science and Information Engineering, National Taiwan University, where he is presently a Professor. He was also the Head of the Department of Information Engineering, National Chi Nan University, Puli, Taiwan, from Aug. 1996 to July. 1998.

Prof. Wu was the recipient of the 1989 Outstanding Youth Medal of the Republic of China, the Outstanding Research Award sponsored by the National Science Council, from 1987 to 1992, and the Excellent Research Award from NSC, 1999.

Prof. Wu has published more than 200 journal and conference papers. His research interests include algorithm design for DSP, data compression, digital watermarking techniques and multimedia systems.

A Low-Cost Media-Processor Based Real-Time MPEG-4 Video Decoder

Jin-Hau Kuo, Chia-Chiang Ho, Kan-Li Huang, Jim Shiu, Ja-Ling Wu, *Senior Member, IEEE*

Abstract — *Due to the high-speed development in semi-conductor technology, general-purpose computers can process digital video, audio and graphics easily. Nowadays, the similar functionality has shifted from the general-purpose computer to the so-called media-processor one. Since the processor is targeted on mass-produced consumer electronics devices, the price and the ease of use are the main concern. The latest MPEG-4 video coding standard, which consumes relatively low bit-rate but yields acceptable quality, meets the target exactly. In this paper, by realizing an MPEG-4 codec using both general-purpose processor and media-processor, we investigated some issues related to the use of a media-processor for satisfying various multimedia compression and processing requirements. The investigated issues include the speedup of the MPEG-4 decoder based on a media-processor DSP chip, a multimedia co-processor or hardware accelerator, the Very Long Instruction Word (VLIW) and the Single Instruction Multiple data stream (SIMD) programming techniques. Moreover, a media-processor based MPEG-4 solution for set-top box applications is also included*[1].

Index Terms — MPEG-4, SIMD, VLIW, Performance measurement, Embedded system.

I. INTRODUCTION

The progress of video compression techniques is getting mature after years of efforts on developing many standards and de-facto codecs. The well-known moving picture expert group (MPEG) has worked out the newest version of video codec, the MPEG-4 [1], which consumes relatively low bit-rate but yields acceptable quality, as compared with the widely adopted MPEG-2 counterpart. Recently, lots of researches [2]-[6] have focused on how to meet various multimedia compression and processing requirements based on a media-processor platform. Among them, realization of the MPEG-4 decoder using a media-processor is of great interest. The compelling reasons for adopting this approach are its high flexibility and low cost. A media-processor based terminal device has to provide cost effective (based on commodity pricing) execution environment for video applications because its silicon resources are limited. This leads to many challenges in designing and programming of media-processors.

In this paper, by implementing a **realistic** MPEG-4 simple profile plus B-VOP decoder over a specific media-processor system [7], we investigate a series of important issues related to the use of a media-processor for satisfying various multimedia compressions and processing requirements. The issues we investigated including the speedup of the MPEG-4 decoder based on a specific media-processor DSP chip (such as a multimedia co-processor or hardware accelerator), the Very Long Instruction Word (VLIW), and the Single Instruction Multiple data stream (SIMD) programming techniques. And finally, a media-processor based MPEG-4 solution for set-top box applications is also included. Our optimized implementation of an MPEG-4 decoder on the given DSP chip can support various resolutions up to 4CIF (720 x 576 pixels) at frame rate 30 frames per second (fps).

The paper is organized as follows: In section II, an overview of the MPEG-4 standard is presented. For comparison, section III presents a general-purpose processor based MPEG-4 video codec. The corresponding performance and experimental results are also included. Based on these results, the investigation of a media-processor based real-time MPEG-4 video decoder is provided in Section IV. The corresponding experimental results are shown in Section V. Finally, section VI concludes this write-up and points out the directions of our future work.

II. THE MPEG-4 NATURAL VIDEO CODING

Generally speaking, multimedia can be viewed as a framework for interaction with information from different types of sources, such as image, video, computer-generated animation, speech and music. In the last decade, some international organizations, such as ISO and ITU, had developed various international standards for representing, delivering, and accessing different kinds of multimedia sources. However, due to the fast growing of Internet and web-based interaction, together with the trend of technical convergence of computer, communication, and consumer electronics areas, more and more emerging multimedia applications demand system level integration. It's evident that an open and timely standard is needed. The MPEG committee has foreseen this need and started the process of standardization of MPEG-4 in 1993. After six-year development, the first version of MPEG-4 had been adopted as the ISO/IEC's standard in October 1998. MPEG-4 Version 2, an amendment of various parts of version 1, had also been finished in December 2000.

The MPEG-4 standard, with the formal title "Information Technology –Generic Coding of Audio-Visual Objects", includes the following six parts:

[1] This work is also partially funded by the Ministry of Education (MOE, Taiwan) Program for Promoting Academic Excellence of Universities under the grant number 89E-FA06-2-4-8, and the National Science Council (NSC, Taiwan) under the grant number NSC90-2622-E-002 –008.

Jin-Hau Kuo, Chia-Chiang Ho, Kan-Li Huang and Ja-Ling Wu are with Communication and Multimedia Laboratory, Department of Computer Science and Information Engineering, National Taiwan University, Taipei, 106, Taiwan, R.O.C(e-mail: david@ cmlab.csie.ntu.edu.tw).

Jim Shiu is with [2]SetaBox Technology Corp. (STC), Taipei, Taiwan.

Contributed Paper
Manuscript received September 19, 2003

- Part 1: Systems.
- Part 2: Visual.
- Part 3: Audio.
- Part 4: Conformance Testing.
- Part 5: Reference Software.
- Part 6: Delivery Multimedia Integration Framework.

In short, MPEG-4 provides the standardized technological elements enabling the integration of production, distribution and content access paradigms of the three fields: digital television, interactive graphics applications, and interactive multimedia applications. For getting more detailed information about MPEG-4, references [9]-[11] are highly recommended.

MPEG-4 part 2 (Visual) specifies the coded representations of natural and synthetic visual objects. There are four types of visual objects defined in MPEG-4: video objects, still texture objects, mesh objects, and face objects. Aiming to be a generic standard, MPEG-4 is defined as a toolbox consisting of various tools such that different applications can utilize different sets of tools to accomplish their objectives. In [8], those tools for natural video coding are designed based on the following three considerations:

(i) Compression efficiency:
This is also the leading principle for developing the MPEG-1 and the MPEG-2 standards. In MPEG-4, more advanced compression technologies are added to encourage the developments of new applications.
(ii) Content-based interactivity:
Instead of using traditional frame-based video coding, object-based video coding is adopted to provide more efficient representation, manipulation, editing and scalability of video data.
(iii) Universal access:
Error resilience tools enable MPEG-4 content to be accessed through error-prone communication environment. Object-based temporal and spatial scalability tools enabling MPEG-4 contents to be presented in platforms of various computing or power-consumption capabilities.
These tools are needed for achieving MPEG-4 simple profile and will be discussed in section 3. The information about other tools can be found in [8].

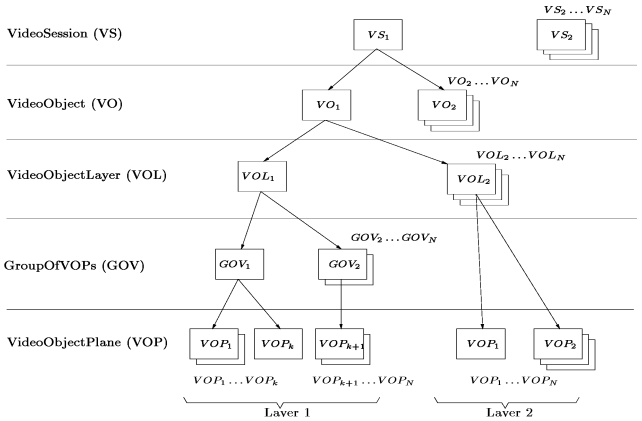


Fig. 1. The logic structure of an MPEG-4 video bitstream [8].

The syntax of MPEG-4 natural video coding [11] is consisted of a hierarchical structure, as shown in Fig. 1. The major layers of the hierarchy are:

■ **Visual Object Sequence (VS)**:
This layer provides global configuration information, referring to the whole group of visual objects that will be simultaneously decoded and composed by a decoder.
■ **Video Object (VO)**:
Video object is one type of visual object. A video object can be frame-based, or it can have time-varying arbitrary shapes.
■ **Video Object Layer (VOL)**:
Each video object can be represented in non-scalable or scalable form. For scalable form, a video object is encoded using at least two video object layers: one base layer and one enhanced layer. Both temporal and spatial scalabilities have been defined.
■ **Video Object Plane (VOP)**:
Each video object is sampled in time, each time sample of a video object is called a "video object plane". Video object planes can be grouped together to form a "group of video object plane", which is optional. A "frame" in MPEG-1 and MPEG-2 can be viewed as a rectangular VOP in MPEG-4.
■ **Video Packet, Macroblock (MB), and Block**:
Video packet is similar to "slice" defined in MPEG-1 and MPEG-2, or "GOB" defined in H.263. Variable number of macroblocks compose a video packet and fixed number of blocks composes a macroblock. And each block consists of 0 to 64 (8x8) color samples.

In MPEG-4 video, macroblock information is divided into three parts: shape information, motion information and texture information. These three parts are coded separately. A simplified video decoding process is shown in Fig. 2.

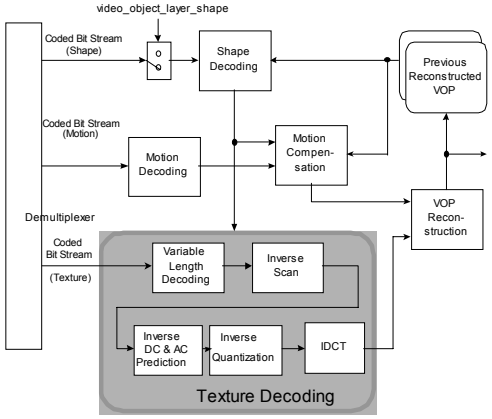


Fig. 2. A simplified MPEG-4 Video Decoding Process.

III. An Implementation of MPEG-4 Simple Profile Natural Video Codec on General-Purpose Processors

Before investigating and implementing an MPEG-4 simple profile decoder on a specific media-processor, we first realize it on a general-purpose processor (such as the CPU of personal computer). This work is of two purposes: First, most of the media-processors lack for a user-friendly debugging-environment. This fact causes difficulties in programming and debugging. Thus, we first implement a modularized MPEG-4

natural codec using Microsoft Visual C++ 6.0 on Intel Pentium Processor, with Microsoft Windows operation system. The Microsoft Visual C++ provides a convenient and powerful programming environment. It can assist us in developing a prototype of MPEG-4 video codec, in algorithm level, very quickly. Second, from the performance profile (in terms of computation complexity) of each module, we can find the bottleneck of the implementation, and then, optimize the corresponding modules, in algorithm level, so as to speed up the overall system.

The realized simple profile natural video coder consists of three sets of coding tools, namely, short header tools, basic tools, and error resilience tools. The simple profile codec provides efficient and error resilient coding for rectangular video objects, and therefore, is suitable for applications in mobile environment.

3.1 Short Header Tool

Short Header tool provides the minimum set of syntax for coding rectangular video objects. This syntax is very similar to that of the H.263. Other tools defined in the MPEG-4 simple profile are disabled when Short Header tool is used.

3.2 Basic Tools

Basic tools of the simple profile address the issue of compression efficiency.

3.2.1 I-VOP and P-VOP

In the simple profile, B-VOP's are forbidden. Advantages of using only I-VOP's and P-VOP's include small memory requirement, low end-to-end delay and low complexity, which are desirable features for applications reside in low computing and low power consumption surroundings. Only two frame buffers are required for both encoder and decoder: one for decoding and another for motion estimation/compensation referencing. For coding efficiency, we have implemented the B-VOP tool. Adding this feature needs one more buffer to store the averaged reference frames.

3.2.2 AC/DC Prediction

Not surprising, each 8x8 block is transformed into DCT domain in MPEG-4, followed by a quantization process. MPEG-4 specifies two methods for quantization. The first one is similar to that of the MPEG-1/2, using two separate quantization matrices respectively for inter and intra blocks. The second one is similar to that of the H.263, using the same quantization step for all AC's. If Short Header tool is used, DC's are quantized using a fixed quantization scale of value 8; otherwise, non-linear scales, based on the quantization steps used for all AC's, are used to quantize the DC's.

Intra DC's are coded differentially, as were done in many popular image or video coding standards, to increase compression efficiency. In MPEG-4, the predictor for one DC is adaptively selected, based on the comparison of horizontal and vertical DC gradients, as shown in Fig. 3.

MPEG-4 also defines the predictors for the first row or the first column of AC coefficients. The predictors are taken from the co-sited coefficients of the candidate blocks. The candidate block is the same block of the DC predictor. Fig. 4 illustrates this idea. To compensate for differences of the quantization steps used in the current block and the predicting block, appropriate scaling of prediction coefficients is required. To implement AC/DC Prediction tool, both our encoder and decoder buffers some information of each block in the last horizontal block line of the current frame. The needed information includes the DC, the prediction direction of DC, the first row and the first column of AC's, and of course the quantization step size.

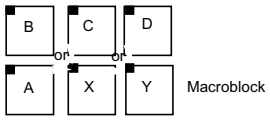


Fig. 3. If the difference between DC(A) and DC(B) is smaller than that between DC(B) and DC(C), the predictor for DC(X) is DC(A); otherwise the predictor is DC(C).

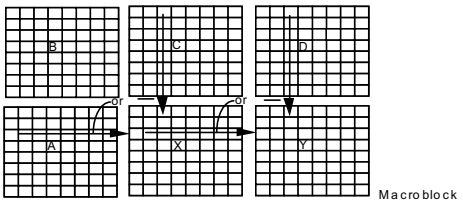


Fig. 4: If the DC predictor of block X comes from block A, the first column of AC's in X are predicted by the co-sited AC's come from block A; otherwise, the first row of AC's in X are predicted by the co-sited AC's come from block C.

3.2.3 4-MV and Unrestricted MV

Motion compensation (estimation) in the simple profile is nothing new, as compared to what MPEG-1/2 and H.263 are specified. The 4-MV tool gives every 8x8 block a motion vector, while the Unrestricted MV tool permits reference macroblocks or blocks to be not fully located in a frame region. Motion vectors are differentially coded, and each predictor is generated by taking the median of three motion vectors coming from neighboring macroblocks (or blocks, if the 4-MV Tool is used). The selection of candidate motion vectors in MPEG-4 is the same as that of the H.263, but the median operations defined by MPEG-4 and H.263 are slightly different from each other.

To implement the Unrestricted MV tool, the frame buffers are allocated two macroblocks larger, in both width and height, than the actual size of each frame. Proper stuffing is done for outliner pixels. The enlarged frame buffer is then used for motion estimation.

To implement the 4-MV tool, the decoder takes 'block' as the motion compensation processing unit, while the encoder performs traditional motion estimation algorithm for each macroblock, followed by a "local search around" process to

find the proper motion vector of each block.

To implement differential coding of motion vectors, the motion vectors of the macroblocks (blocks) in the last encoded macroblock (block) line are buffered.

3.3 Error Resilience Tools

3.3.1 Resynchronization

In MPEG-4, resynchronization is based on video packets. Each VOP is divided into video packets; each video packets can be decoded independently, that is, without the need of any information from any other video packets in the same VOP. This seems similar to the slice or the GOB approaches adopted in MPEG-1/2 and H.263, respectively. Notice that, MPEG-4 inserts resynchronization markers into the video packet headers of the bitstream, when approximately a constant number of bits are generated. That is, owing to variable bit-rate nature when encoding macroblocks in a frame, high motion regions (which consume more bits) will be protected more strictly, as compared to those low motion regions. The experiments performed on MPEG-4 showing that this arrangement is more efficient than that of the MPEG-1/2 or the H.263.

To realize the Resynchronization, our encoder keeps in mind how many bits have been generated since the start of encoding of the current video packet. The bit count is checked every time when one macroblock is encoded: when the pre-defined budget is reached, the current video packet is terminated and a new video packet is started. The field of header extension code (HEC), in the header of the new video packet, is set to one so that the information necessary for independent video packet decoding is added.

3.3.2 Data Partitioning

The Data Partitioning tool separates the motion and macroblock header information away from the texture information. An additional resynchronization marker is inserted between these two parts of information. If the texture information is attacked by errors, the motion information can still be used to reconstruct the macroblocks, which do not have texture information, by using motion compensation only.

A typical video encoder encodes macroblocks one by one, but this paradigm won't work when data partitioning is a must. To implement the Data Partitioning tool, our encoder buffers motion information, macroblock header information and texture information of each macroblock in the same video packet, by managing a linked list of macroblock information buffers. To handle the enabling of the Resynchronization, a pseudo-encoding phase is needed for each macroblock, so that we can know how many bits will be generated when encoding that macroblock. This idea is realized in the VLC module, which accumulates only bit counts during pseudo-encoding phases, but not generating real bitstreams. When the number of macroblocks of the current video packet is determined by the Resynchronization, the encoder can then start generating real bitstreams by encoding macroblock contexts.

3.3.3 Reversible VLC

The Reversible VLC (RVLC) tool is a data recovery tool, whose codewords are designed such that they can be decoded from both forward and backward directions. In MPEG-4, RVLC is applied only to the quantized DCT coefficients. When the Data Partitioning tool is enabled and some errors occurred in the texture information portion of a video packet, the RVLC tool helps to decode more valid texture information located after the last error, by using backward decoding mode.

The enabling of RVLC is defined in the syntax of a VOL header. Our encoder supports texture encoding, using RVLC table, by providing function pointers to select from two kinds of VLC's.

3.4 Modularized Codec Implementation

Our codec is implemented by modularizing primary functional blocks used in the encoder and the decoder.

The modules specifically defined for the encoder include:

- **Input raw file manipulator**:

Feeds raw images one by one into the frame buffer of the encoder.

- **Output bitstream writer**:

Collects bits generated by VLC encoder and syntax generator, and writes out bitstream into file upon proper time.

- **VLC encoder**:

Performs VLC encoding using internal fixed VLC tables.

- **DCT module and Quantizer**:

Performs DCT transformation and quantizes the transformed results to yields coefficients for each intra or inter block.

- **Motion estimator**:

Performs motion estimation for each macroblock (or block).

- **Syntax generator**:

The driven module of the encoder. Syntax generator calls other modules to prepare data for filling fields defined in the syntax of MPEG-4, to control the data flow, and to write out bitstream into the output file.

The modules specifically defined for the decoder include:

- **Input bitstream file dispatcher**:

Manipulates input bitstream file and provides specific bits required by VLC decoder and syntax parser.

- **Frame display**:

Displays decoded frames on the screen.

- **VLC decoder**:

Performs VLC decoding using table-lookup.

- **Syntax parser**:

The driven module of the decoder. Syntax parser calls other modules to decode field values, to reconstruct macroblocks, and to display decoded frames.

The modules shared by the encoder and the decoder are:

- **IDCT module and Dequantizer**:

Performs dequantization and IDCT to yields spatial domain pixel values or residue data.

- **Motion compensator (MC)**:

Performs motion compensation for each macroblock (or block).

- **Macroblock information buffer manager** (MIBM, activated when Data Partitioning tool is enabled):

MIBM allocates and de-allocates buffers for dealing with

macroblock information. It also maintains a link list of buffers so as to provide necessary information in order, when bitstream generation (encoder side) or macroblock reconstruction (decoder side) is needed for processing each macroblock of the current video packet.

3.5 Codec Optimization

In our realization, four time-consuming functional blocks, namely, DCT, IDCT, motion compensation (MC), and SAD (sum of absolute difference) calculation used in motion estimation module, are implemented using MMX and SSE technologies provided by the Intel Pentium II and Pentium III CPUs [12].

3.6 Experimental Results

Our MPEG-4 natural codec plus B-VOP is implemented using Microsoft Visual C++ 6.0, and is tested on a PII-400 PC platform, with Microsoft Windows operation system. The tested sequences are taken from typical H.263, MPEG-1 and MPEG-4 test sequences, such as "Grandma", "Football", "Table Tennis", and "Akiyo".

3.6.1 Encoder

A simple user interface of our encoder is shown in Fig. 5. The encoder accepts raw files (Y:Cb:Cr, 4:2:0; QCIF/CIF/4CIF in size) as input, and outputs related bitstream files. User can enable or disable some tools defined in the simple profile. Table I shows the resultant file sizes when encoding different input files by using different coding modes. "Grandma" and "Akiyo" are head-and-shoulder sequences, "Table Tennis" is a low motion sequence, and "Football" is a high motion one. It can be observed that enabling Data Partitioning and RVLC tools will increase the bitrate by no more than 5%, and the increased amount gets smaller when the bitrate getting lower. The encoding of "Football" sequence requires larger bit-counts, owing to its fast motion nature. Constant bit-rate control of the encoder is under development, which will help us for reducing bitrate while maintaining compromising video quality. In this way, we can generate lots of testing bitstream files that will be used as inputs for testing the performance of the following media-processor based MPEG-4 video decoder.

Table I: Bitstream sizes (KBytes) of 4 test video sequences under two kinds of coding methods and three different average quantizer steps.

Sequence	Coding mode					
	Short Header disabled, No data partitioning and RVLC			Data partitioning and RVLC		
	AvgQ = 4	AvgQ = 12	AvgQ = 20	AvgQ = 4	AvgQ = 12	AvgQ = 20
Grandma (QCIF, 150 frames)	235	79	47	247	82	49
Football (CIF, 125 frames)	3743	1310	763	3601	1356	789
Table tennis (CIF, 112 frames)	1746	613	345	1824	638	360
Akiyo (4CIF, 300 frames)	2919	1184	858	3041	1210	874

3.6.2 Decoder

A simple user interface of our decoder is shown in Fig. 6. The decoder accepts bitstream files generated from the encoder, and playbacks the video without frame rate control. The top title bar of the decoder shows the average frame rate of the current decoding bitstream file. Table II shows the execution time for decoding various bitstream files. It can be seen that our decoder can do real-time decoding for typical CIF sequences, and have the ability of decoding 4CIF videos at about 10 frames per second. Notice that only the execution time of the decoding kernel is taken into consideration. The experimental results reveal that we need a Pentium II level general-purpose processor plus a media-processor to decode 4CIF video sequence in real-time.

Table II: Average frame rates (frame/sec.) of the realized decoder for decoding various bitstreams. For each x/y in the table, x is the total elapsed time, and y is the time not taking the execution time of color conversion and frame displaying into account.

Sequence	Coding mode					
	Short Header disabled, No data partitioning and RVLC			Data partitioning and RVLC		
	AvgQ = 4	AvgQ = 12	AvgQ = 20	AvgQ = 4	AvgQ = 12	AvgQ = 20
Football (CIF)	15/22	24/41	30/57	14/21	24/39	29/52
Table tennis (CIF)	20/32	32/65	38/83	19/31	30/57	39/87
Akiyo (4CIF)	11/26	13/37	13/40	10/20	11/24	11/24

A simple profile video codec founds the basis of an MPEG-4 video coding application. To implement a robust MPEG-4 decoder, the error resilience tool is the only challenging work. Efficient error detection and data recovery require elaborate consideration. As for the encoder, 1) appropriate bit rate control scheme, 2) fast motion estimation algorithm for advanced prediction mode, and 3) code-level optimization are the keys to implement a robust and efficient MPEG-4 encoder. In general, such codecs should be implemented with modularization concept, enabling easy integration with additional tools defined in other profiles and other parts of MPEG-4. So that many multimedia applications, based on MPEG-4, can be constructed accordingly.

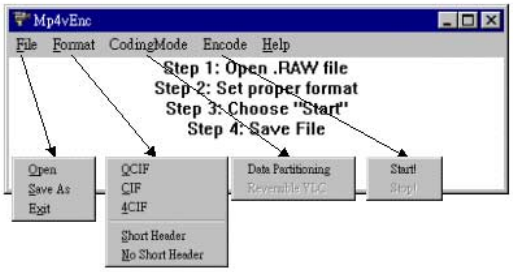


Fig. 5: The user interface of the proposed MPEG-4 video encoder.

Fig. 6: The user interface of the proposed MPEG-4 video decoder.

IV. Implementation of an MPEG-4 Simple Profile Video Decoder on a Media-Processor

Two years ago, we investigated several issues centered on the MPEG-1, 2 video decoders equipped with the Intel MMX, SSE instruction sets [12]. According to our previous findings and experiences, we try to speed up the MPEG-4 decoder over a specific media-processor. Since our goal is targeted on a low-cost media-processor based real-time MPEG-4 video decoder, the cost, the performance, and the ease of use are the main concerns. Obviously, the specific media-processor is cheaper than the general-purpose processor because of its limited functionalities. Nowadays, there are lots of documents, supports, and even discussion groups, to exchange the technical information and concrete experiences concerning on the usage of various media-processors, in the Internet [7], [13]. Those documents and Web-sites provides useful information for us to investigate issues related to the development of multimedia applications based on media-processors.

4.1 Overview of the Media-Processor

4.1.1 Hardware Architecture of the DSP Chip

The core of the target chip is a CPU with VLIW architecture [7]. It is a 32 bits CPU with 128 registers running under 100MHz or 150MHz clock rate. It has 32K instruction cache and 16K data cache using eight-way associative cache organization. Due to hardware simplicity, it is designed as a load-store machine. Besides the core VLIW CPU, the processor also has an Image Co-Processor to deal with the image interpolation and color conversion, an Video In/Out module to deal with the video digital-to-analog and analog-to-digital conversions, an Audio In/Out module whose function is the same as that of the Video In/Out module, and some other core modules for simplifying the developments of various multimedia applications.

4.1.2 Instruction Set Architecture

Three Level operators are supported by the processor: **Instructions**, **Operations** and **RISC operations**. At most five operations can be issued in a single instruction per cycle (c.f. Fig. 7). Because of limited function unit and instruction constraints, operations issued in a single instruction cycle usually less than five. The computational operation limits and classifications are shown in Fig. 8. In order to support multimedia applications, there are some built-in SIMD operations on the chip to enhance its performance.

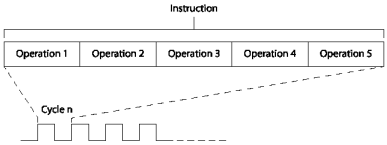


Fig. 7: The characteristics of the Instruction Hierarchy.

Functional Unit	Quantity	Latency[A]/Delay[B]	Recovery[C]	Slot Assignment				
				1	2	3	4	5
Constant	5	1	1	✓	✓	✓	✓	✓
Integer ALU	5	1	1	✓	✓	✓	✓	✓
Load/Store	2	3	1				✓	✓
DSP ALU	2	2	1	✓		✓		
DSP MUL	2	3	1		✓	✓		
Shifter	2	1	1	✓	✓			
Branch	3	3	1		✓	✓	✓	
Int/Float MUL	2	3	1		✓	✓		
Float ALU	2	3	1	✓			✓	
Float Compare	1	1	1			✓		
Float sqrt/div	1	17	16		✓			

Fig. 8: Operations' Classifications and Limits [7]

4.1.3 Memory Architecture

The memory architecture of the target processor consists of a two level memory hierarchy, and there is only one cache between the CPU and the main memory. The main memory uses SDRAM and the cache is on-chip. The memory system uses 32 bits to address data in cache or in memory. There are two separate caches: instruction cache and data cache. They have different sizes (32K bytes for instruction cache and 16K bytes for data cache) and access ports (2 for data cache and one for instruction cache). On the other hand, there are also some common characteristics, such as 8-way associative, 64 bytes per block, 1 valid bit and 1 dirty bit per block, between them. In more detail, data cache has two ports which can support two accesses concurrently, and uses copy-back, allocate-on-write, critical-word-first transferring, and hierarchy LRU algorithms as cache miss and replacement policies.

4.1.4 Software Development Environment

The target processor provides a set of software development tools including compiler, performance tuner, scheduler, assembler, linker, loader, data books, simulator, profiler, and some utilities applicable for software development.

This development kit [7] supports some programming styles, such as C/C++, decision tree, and assembly. We only use C/C++ and assembly to re-write and evaluate our MPEG video decoder. In our work, we adopt some tools in that development kit, such as compiler driver, simulator, and profiler to develop and evaluate our task. The text editor we used to write code is just as simple as that of the UltraEditor.

4.2. The Performance Optimization of the Media-Processor based MPEG-4 Video Decoder

4.2.1 The Speedup of the MPEG-4 decoder

The steps of acceleration can be considered as hierarchical stages, shown in Fig. 9. From top to bottom are the stages of *Algorithm, Instruction Set* and *Assembly Levels*, respectively. As expected, based on our previous research [12], the IDCT and the MC are the two most processing-intensive tasks. The performance profile is obtained by porting a pure C-language coded non-accelerating MPEG-4 decoder on the media-processor, directly.

In the *Algorithm Level*, we focus on the methodologies of developing fast algorithms for time-consuming modules (e.g., IDCT), parallel computing techniques (e.g., VLIW, SIMD instructions), and some other C code optimization skills to increase the performance gain. For IDCT, we apply the fast DCT-SQ scheme [14] and extend it to 2-D as what has been done it [12]. IDCT is a time-consuming module done for every block (there are 8x8 pixels per block). The following gives a general idea about its complexity.

For one I-VOP of a video sequence with picture size 720x480 pixels (4:2:0), the number of blocks is 8100 (720x480/64 + 720x480/256 + 720x480/256). For the brute-force 1-D and 2-D IDCT algorithms [15], the required numbers of multipliers and additions are 64, 56, and 1024, 896, respectively. For the fast 1-D and 2-D IDCT algorithms [14], the corresponding numbers of multipliers and additions are 13, 29, and 60, 462, respectively. The performance gain of the fast IDCT algorithms is quite amazing.

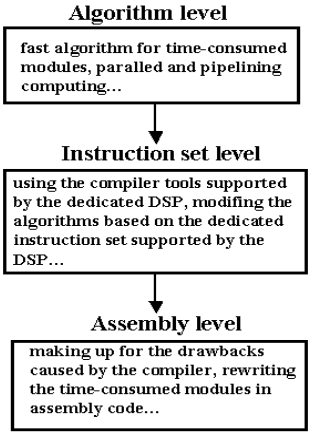


Fig. 9: The hierarchy stages of acceleration and the related methodologies

A 1-D DCT-SQ based IDCT algorithm needs 8+5 multiplications (8 for constant pre-multiplication with 8 input data), and 29 additions. We extend it to 2-D IDCT [11], and then optimize it further using VLIW and SIMD programming skills. The resultant 2-D IDCT algorithm just needs 60 multiplications and 467 additions. This part will be described in more detail in the stage of *Instruction Set Level*. On the other hand, for matrix transposing during execution of 2-D IDCT, we adopt in-place matrix transpose, instead of doing real matrix transposition.

Since the target media-processor possesses a VLIW core, which can issue 5 operations in each cycle. Which means, usually, there is enough room for filling in some time slots, without impacting on the total number of cycles. In some cases, it may be more efficient to schedule two unrelated operations in the same instruction cycle than the one with a longer latency and no-operations (nops). Thus, doing unrelated operations concurrently would increase the parallelism of computation, and therefore, improving the overall performance.

The existence of a branch is a big factor that will affect the performance. Branch may arise in a function call or one decision point (e.g., IF, While conditional operations). When a branch occurs, all present status need to be stored in order to continue the execution correctly when the branch terminates. In order to lower the number of conditional branches, one can use the scheduler to add a predicate field to any assembly operation. The field is implemented by using a flag, usually named *guard*. The *guard instruction* is a ``meta-instruction'' that does not modify the state of the CPU directly but instead may modify the subsequent instructions read by the system. The operation executes only if the *guard flag* is true. With guarding, one can also allow or prevent the execution of an operation using the same condition that would determine the outcome of a conditional branch. For performance gain, we try to lower the occurrence of branches by merging some related functions together and replacing the branch operations by guard operations. For a C-level operation 'IF v E1, else E2', if the guard is known to depend on the first order bit, one can use the machine level pseudo operation 'mux' instead. For example, if the variable v is constrained to be 0 or 1, one can replace MUX(v != 0, E1 , E2) by mux(v, E1 , E2), for saving a cycle. We also merge some functions, such as functions for processing AC coefficients into one function ac_rescale_recon_store, to reduce the branch overheads caused by function calls. The performance gain of this approach is impressive.

In the *Instruction Set Level*, we use high-level instruction sets, named *customer operations (custom ops)*, instead of writing assembly codes directly. The *customer ops* are C-level functions and can be mapped to low-level assembly operations directly. That means, by using *customer ops*, we can force the compiler or assembler to use the corresponding assembly instructions correctly. The same capabilities of C-level based MMX and SSE functions are also supplied by Intel. Therefore, we can conduct the optimization task more easily and discard operation schedulings, like stall, issue, and debugging problems.

As prescribed in the *Algorithm Level*, we apply the VLIW and SIMD instruction sets to optimize the IDCT module [11], further. In order to use SIMD instructions (two 16-bits operations concurrently), we adopt a 2-byte signed integer data type to calculate the IDCT. Thus the floating-point operations can be eliminated, but the precision of the output data is also reduced.

MC is also a time-consuming module of the decoding process. MC is carried out for every no-intra coded macroblock (16x16 pixels per macroblock). Due to half-pixel motion compensation is

also supported by the MPEG-4 video standard, some interpolations must be executed during the decoding process. However, it is also a processing-intensive task. For this module, we adopt the loop unrolling and the SIMD instructions to improve the performance. By unrolling MC module, we can schedule it onto VLIW architecture more compactly. That means the five operation slots are almost fully issued. The following is the pseudo code for block motion compensation and interpolation of the MC module.

```
CopyBlock(Src, Dst, Stride){
int dy;
long *lSrc = (long *) Src;  //for SIMD operation
long *lDst = (long *) Dst;  // for SIMD operation
for (dy = 0; dy < 8; dy++) {
    lDst[0] = FUNSHIFT(lSrc[1], lSrc[0]);
    lDst[1] = FUNSHIFT(lSrc[2], lSrc[1]);
    lSrc += Stride;
    lDst += Stride;
    }
}
CopyBlockH(Src, Dst, Stride){
int dy;
long *lSrc1;
long * lSrc = (long *) Src;  //for SIMD operation
long *lDst = (long *) Dst;  //for SIMD operation
for (dy = 0; dy < 8; dy++) {
   lSrc1=lSrc+Stride;
   lDst[0]=QUADAVG(FUNSHIFT(lSrc[1],lSrc[0]),
FUNSHIFT(lSrc[1], lSrc[0]));
   lDst[1]=QUADAVG(FUNSHIFT(lSrc[2],lSrc[1]),
FUNSHIFT(lSrc[2],lSrc[1]));
   lSrc=lSrc1;
   lDst+=Stride;
   }
}
CopyBlockV(Src, Dst, Stride){
int dy;
long t0, t1, t2, *lSrc1;
long *lSrc = (long *) Src;  //for SIMD operation
long *lDst = (long *) Dst;  //for SIMD operation
for (dy = 0; dy < 8; dy++) {
   lSrc1 = lSrc + Stride;
   t0 = QUADAVG(lSrc[0], lSrc1[0]);
   t1 = QUADAVG(lSrc[1], lSrc1[1]);
   t2 = QUADAVG(lSrc[2], lSrc1[2]);
   lDst[0] = FUNSHIFT(t1, t0)
   lDst[1] = FUNSHIFT (t2, t1)
   lSrc = lSrc1;
   lDst += Stride;
   }
}
CopyBlockHV(Src, Dst, Stride){
int dy; long t0, t1, t2, *lSrc1;
long *lSrc = (long *) Src;
long *lDst = (long *) Dst;
for (dy = 0; dy < 8; dy++) {
   lSrc1 = lSrc + Stride;
   t0 = QUADAVG(lSrc[0], lSrc1[0]);
   t1 = QUADAVG(lSrc[1], lSrc1[1]);
   t2 = QUADAVG(lSrc[2], lSrc1[2]);
lDst[0]=QUADAVG(FUNSHIFT(t1,t0),FUNSHIFT(t1,t0))
lDst[1]=QUADAVG(FUNSHIFT(t2,t1),FUNSHIFT(t2,t1))
   lSrc = lSrc1;
   lDst += Stride;
   }
}
```

Table III shows the SIMD instructions we used in IDCT and MC modules, respectively. The functionality and purposes of the corresponding SIMD instructions are also shown in Table III. As compared to the well-known multimedia instruction sets, such as MMX and SSE, some frequently used SIMD instructions but not provided by the media-processor are also specified in Table III. Some overheads are resulted from the incompleteness of SIMD instructions. The comparisons and experiment results can provide us a hint towards the solution of designing a single chip MPEG-4 decoder.

Table III: The SIMD instructions we used in (a) IDCT and (b) MC modules. The functionality and purpose of the corresponding SIMD instructions are also shown. The gray rows specified some frequently used SIMD instructions but not provided by the media-processor, as compared to the well-known multimedia instruction sets, such as MMX and SSE, provided by the Pentium II and III CPUs.

Data Type : 16 bits signed integer Operation			
Operation (customer ops)	Yes	No	Purpose
DSPIDUALADD	√		Dual clipped add of signed 16-bit halfwords
DSPIDUALSUB	√		Dual clipped sub of signed 16-bit halfwords
PACK16MSB	√		Pack most-significant 16 bits
PACK16LSB	√		Pack least-significant 16 bits
DUALASR	√		Dual-16 arithmetic shift right
DUALASL		√	Dual-16 arithmetic shift left (use bit-shift operation to simulate this operation)
IFIR16	√		Sum of products of signed 16-bit halfwords
DUALMULMSB		√	Dual clipped and multiply of signed 16-bit halfwords and dual get most-significant 16 bits (use IFIR16 and PACK16MSB to simulate this operation)

(a)

Data Type : 8 bits signed or unsigned integer Operation			
Operation (customer ops)	Yes	No	Purpose
QUADAVG	√		Unsigned byte-wise quad average
FUNSHIFT1	√		Funnel-shift 1-byte
FUNSHIFT2	√		Funnel-shift 2-byte
FUNSHIFT3	√		Funnel-shift 3-byte
DSPUQUADADDUI	√		Quad clipped add of unsigned/signed bytes
DUALICLIPI	√		Dual-16 clip signed to signed
QUADD		√	Unsigned byte-wise quad add (not supported by TriMedia™ DSP chip)
QUADAVGN		√	Unsigned byte-wise quad average without add one for rounding (not supported by TriMedia™ DSP chip)

(b)

In the *Assembly Level*, we inspect the assembly codes generated by compiler and assembler. By making use of the register renaming method, we can save more execution cycles to improve performance.

4.2.2 The co-processor based acceleration

The Image Co-Processor (ICP) of the target media-processor can be considered as an on-chip data highway to

perform SDRAM block read and write actions. It usually connects to the PCI interface to allow block write transactions across PCI. In the media-processor we selected, the ICP provides the functions of filtering, resizing (scaling) and YUV to RGB conversion of the source image. Filtering supports image enhancement. Scaling generates a new image that is larger or smaller than the original image. The YUV to RGB conversion is used to generate an RGB version of the image for outputting a frame buffer through the PCI interface or to a SDRAM. On account of all the ICP functions, data can be moved and transformed from memory to memory or from memory to the PCI bus. Hence, the CPU of the media-processor can use the ICP in a time-sharing manner to simultaneously achieve the prescribed functions. We make use of the ICP to scale the video fitting the size of monitor screen or TV. By utilizing the ICP functions, we can reduce the load of the media-processor CPU and gain performance.

V. EXPERIMENTAL RESULTS

Table IV and V show the execution time for decoding various bitstream files without and with optimization, respectively. The Max value denotes the maximal time for decoding one frame in the corresponding sequence. It occurs usually in I-VOP and IDCT modules. By observing the Max values, we can evaluate the performance gain of the IDCT module. The Kernel value means that only the execution time of the decoding kernel is taken into consideration. It can be seen that the proposed decoder can do real-time decoding for typical CIF sequences, and have the ability of decoding typical 4CIF sequences at about 29 frames per second.

Table IV: The execution time for decoding various bitstream files without optimization. The Max value denotes the maximal time for decoding one frame in the corresponding sequence. The Kernel value means that only the execution time of the decoding kernel is taken into account.

Video Sequence	fps	Max	Average	Kernel
Outbreak (4CIF, 1001 frames)	1001/73.07s =13.69fps	187ms	72ms	72544ms
CominBack (512x384 2174 frames)	2174/95.93s =22.66fps	153ms	42ms	90424ms
Crash (512x384 1754 frames)	1754/73.72s =23.79fps	105ms	40ms	71208ms
LastStop (512x384 2760 frames)	2760/109.38s =25.23fps	136ms	37ms	103178ms
Radius (640x352 2400 frames)	2400/145.86s =16.45fps	160ms	54ms	130537ms

Table V: The execution time for decoding various bitstream files with optimization.

Video Sequence	fps	Max	Average	Kernel
Outbreak (4CIF, 1001 frames)	1001/33.84s =29.58fps	58ms	20ms	20922ms
CominBack (512x384 2174 frames)	2174/72.48s =29.99fps	48ms	13ms	29289ms
Crash (512x384 1754 frames)	1754/58.45s =30.01fps	29ms	13ms	22881ms
LastStop (512x384 2760 frames)	2760/92.02s =29.99fps	41ms	12ms	33538
Radius (640x352 2400 frames)	2400/80.30s =29.89fps	51ms	16ms	40473ms

VI. APPLICATIONS

Over the last decade, the slow adoption of broadband as well as the expense and quality of content delivery has hampered the efforts of offering high-quality network services. But in recent years, as the broadband era is coming, more and more information including videos will be transferred among various networks. The cable and ISP industries have sought to offer high-quality video-on-demand and interactive entertainment services to users.

By cooperating with a set-top box developing term, we have also investigated the possibility of using media-processor based MPEG-4 decoder in set-top box related applications. A prototype of low-cost PCI I/F MPEG-4 decoder card has been realized. Fig. 10 shows a picture of the completed MPEG-4 decoder card. The card can support for bit rates from 16 Kbps to 10 Mbps and resolutions up to 4CIF (720 x 576) at 30 fps. The corresponding applications include the broadband set-top box accessing, broadband video streaming and high quality surveillance.

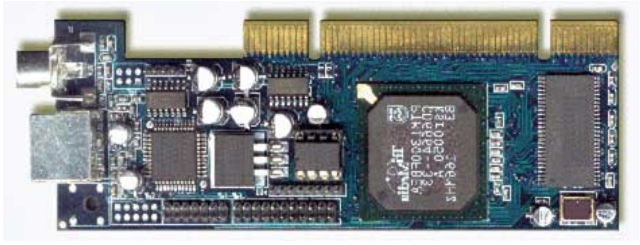


Fig. 10: The picture of the completed PCI MPEG-4 decoder card.

VII. CONCLUSION AND FUTURE WORK

In this paper, by implementing a realistic MPEG-4 simple profile decoder, over a specific media-processor, we have investigated a series of issues related to the use of a media-processor for satisfying multimedia compression and processing requirements. Further, we compared the instruction sets provided by the general-purpose processor, which shows the incompleteness and the further requirements in VLIW and SIMD instruction sets. The comparisons and experiment results can provide a hint towards the solution of designing a single MPEG-4 chip. To cooperate with a set-top box

developing term, a powerful MPEG-4 decoder PCI card for Set-top box Streaming Applications has been realized.

More tools, defined in the MPEG-4 core and advance simple profiles, such as 1/2 quantization, interlace, global motion compensation, and 1/4 pixel motion compensation, will also be investigated and implemented in the near future.

ACKNOWLEDGMENT

The author would like to thank SetaBox Technology Corp. for supporting this work. This work is also partially funded by the Ministry of Education (MOE, Taiwan) Program for Promoting Academic Excellence of Universities under the grant number 89E-FA06-2-4-8, and the National Science Council (NSC, Taiwan) under the grant number NSC90-2622-E-002 –008.

REFERENCES

[1] ISO/IEC International Standard 14496, Information Technology – Generic Coding of Audio-Visual Objects – Part 1: System, Part 2: Visual, Part 3: Audio, Part 5: Reference software, Part 6: DMIF. 2000.

[2] Le, T.M., Nita, A., Giernalczyk, E., Denny Wong, Tuan Ho, "A low-power and scalable MPEG-4 solution for wireless communications," in IEEE Int. Conf. Consumer Electronics, 2001.

[3] Fermo, A., Sicuranza, G.L., Pahor, V., "Hardware-oriented region based algorithm for low power motion estimation," in IEEE Proceedings of the 2nd International Symposium on Image and Signal Processing and Analysis, 2001.

[4] Ramkishor, K., Gunashree, V., "Real time implementation of MPEG-4 video decoder on ARM7TDMI," in IEEE Proceedings of International Symposium on Intelligent Multimedia, Video and Speech Processing, 2001.

[5] Burleson, W., Jain, P., Venkatraman, S., "Dynamically parameterized architectures for power-aware video coding: motion estimation and DCT," in IEEE Proceedings of the 2nd International Workshop on Digital and Computational Video, 2001.

[6] "Special Issue on Multimedia Implementation," in IEEE Transaction on Circuits and System for Video Technology, August 2002.

[7] Philips TriMedia™ Software Development Environment Version 2.1.

[8] Touradj Ebrahimi and Caspar Horne, "MPEG-4 Natural Video Coding - An overview," http://cselt.it/mpeg.

[9] Fernando Perira, "MPEG-4: Why, What, How and When," http://cselt.it/mpeg.

[10] MPEG, "MPEG-4 Overview," *Doc. ISO/MPEG N2995,* Melbourne MPEG Meeting, October 1999.

[11] ISO/IEC 14496-2:2001, "*Coding of Audio-Visual Objects - Part 2: Visual*", 2nd Edition, 2001

[12] Yi-Shin Tung, Chia-Chiang Ho, Ja-Ling Wu, "The MMX-based IDCT and MC Algorithms for Real-Time Pure Software MPEG Decoding," ICMCS'99, Florence Italy.

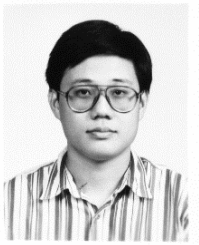
[13] Trimedia· Exchange of technical information and questions concerning the Philips TriMedia video/audio/telecomms DSP processor. http://groups.yahoo.com/group/trimedia/

[14] Y. Arai, T.Agui, and M.Nakajima, "A Fast DCT-SQ Scheme for Images", Trans. Of the IEICE, E71(11):1095, Nov 1988.

[15] William B. Pennebaker and Joan L. Mitchell, "JPEG : STILL IMAGE DATA COMPRESSION STANDARD", VAN NOSTRAND REINHOLD, New York.

Jin-Hau Kuo received the B.S. degree in information and computer engineering from Tatung University, Taipei, Taiwan, in 1997. He is currently a Ph.D. student in the department of computer science and information engineering, National Taiwan University. His current research interests include data compression, digital image processing, content-based retrieval, code optimization, and multimedia systems.

Chia-Chiang Ho received the B.S. degree in computer science and information engineering from National Taiwan University, Taipei, Taiwan, in 1996. He is currently a Ph.D. student in the Dept. of Computer Science and Information Engineering, National Taiwan University. His current research interests include image and video coding, multimedia systems, scalable codecs, and error resilient.



Jim Shiu received the B.S. degree in information and computer engineering from Tatung University, Taipei, Taiwan, in 1987, and the M.S. and Ph.D degree in computer science and information engineering from National Taiwan University, Taipei, Taiwan, in 1989 and 1993, respectively. respectively. His current research interests include data compression, digital image processing, code optimization, and multimedia systems.

Kan-Li Huang received the B.S. degree in computer science and information engineering from National Taiwan University, Taipei, Taiwan, in 1998. He is currently a master student in the Dept. of Computer Science and Information Engineering, National Taiwan University. His current research interests include image and video coding, multimedia systems, scalable codecs, video transcoding and optimization techniques for video compression algorithms.

Ja-Ling Wu received the B.S. degree in electronic engineering from TamKang University, Tamshoei, Taiwan, R.O.C. in 1979, and the M.S. and Ph.D degree in electrical engineering from Tatung Institue of Technology, Taipei, Taiwan, in 1981 and 1986, respectively. From 1986 to 1987, he was an Associate professor of the Electrical Engineering Department at Tatung Institute of Technology, Taipei, Taiwan. Since 1987, he has been with the Department of Computer Science and Information Engineering, National Taiwan University, where he is presently a Professor. He was also the Head of the Department of Information Engineering, National Chi Nan University, Puli, Taiwan, from Aug. 1996 to July 1998. Prof. Wu was the recipient of the 1989 Outstanding Youth Medal of the Republic of China, the Outstanding Research Award sponsored by the National Science Council, from 1987 to 1992, and the Excellent Research Award from NSC, 1999. Prof. Wu has published more than 200 technique and conference papers. His research interests include algorithm design for DSP, data compression, digital watermarking techniques and multimedia systems.

分項計畫三：MPEG-4/7三維虛擬實境內容處理技術 (III)

Content Manipulation for MPEG-4/7 3D Virtual Environment (III)

成果報告

計畫編號：NSC91-2622-E-002 –002

全程計畫：民國 90 年 08 月 01 日至民國 93 年 07 月 31 日

本年度計畫：民國 92 年 08 月 01 日至民國 93 年 07 月 31 日

計畫主持人　：歐陽明　台灣大學資訊工程系教授
　　陳炳宇　台灣大學資訊管理系助理教授
共同主持人　：黃肇雄　台灣大學資訊工程系教授
　　陳文進　台灣大學資訊工程系教授
　　吳家麟　台灣大學資訊工程系教授
　　周承復　台灣大學資訊工程系助理教授

一、摘要

本分項計畫分為兩個系統：三維 MPEG-4 動畫場景建構工作台(3D Workbench for MPEG-4 Scene and Animation Construction)和三維物件搜尋器(Query by Example in 3D)，分別支援及實證【媒體內容工程：MPEG-4/7 相關技術之研發】總計畫中有關【互動式場景編輯器】與【搜尋多媒體物件】兩項技術。本分項計畫利用所開發出之技術，支援【媒體內容工程:MPEG-4/7 相關技術之研發】總計畫中，【互動資訊媒體服務系統(Interactive Information Media Service System)】與【家庭媒體中心(Home Media Center)】兩個應用系統的技術部分。

在【互動式場景編輯器】的技術部分，主要目的在支援總計畫中所提出的應用系統【互動資訊媒體服務系統(Interactive Information Media Service System)】，使得互動資訊媒體服務系統可以結合三維場景與三維物件，產生更吸引人的視覺效果。在三維 MPEG-4 動畫場景建構工作台的部分，必須要有方便使用的人機介面，可以快速地建立豐富的場景與動畫，使得分項計畫一中的【MPEG-4 場景編輯器】，可以輕易的加入想要的三維模型。在【搜尋多媒體物件】的技術部分，主要目的在支援總計畫中所提出的應用系統－【家庭媒體中心(Home Media Center)】。期計能在【家庭媒體中心】中，不只可以快速找到需要的影像、聲音或影片，也能夠迅速找到所想要的三維模型。在面對越來越多與豐富的三維模型資料庫，建立適當的管理系統是必要的，因此，本分項計畫利用 MPEG-7 及符合其精神之三維模型之述詞(Descriptor)與描述結構(Descriptor Structure)作為特徵(Feature)，建立自動或半自動化的三維模型搜尋分類系統。

This project have developed two techniques: (1) the 3D workbench for MPEG-4 scene and animation construction and (2) query by example in 3D, which support the proposed systems: Interactive Information Media Service System and Home Media Center in the collaborated NSC project titled with Content Engineering: Research on MPEG-4/7 Multimedia Technologies.

In order to quickly create abundant 3D scenes and animation, it is important to have a convenient user interface, such as force-feedback space arm and 3D tracker with an immersive display environment. When using force-feedback space tracker, collision detection and mechanics in 3D scene must be taken into account. Therefore, the techniques of 3D Workbench for MPEG-4 Scene and Animation Construction include: collision detection, multi-resolution space partition, 3D painting on mesh surface, and force-feedback scene editing.

The proposed technique Query by Example in 3D is to fulfill the application Home Media Center, proposed in the collaborated project, so that user can search and retrieve not only images, audio and video, but also 3D models. Nowadays, more and more 3D virtual reality contents are created, so how to manage and storage them becomes a crucial issue. Therefore, we develop an automatic or semi-automatic searching and indexing system for 3D model retrieval. This system use MPEG-7 techniques.

Keywords: MPEG-4, Virtual Sculptor, 3D Modeling, Extended Marching Cube, Tangent-Plane Force Model, MPEG-7, 3D Model Retrieval, Light Field Descriptor, Orthogonal Visual Hull.

二、計畫緣由與目的

長久以來，MPEG(Moving Picture Experts Group)組織所訂定的MPEG標準國際影音壓縮標準(如MPEG-1，MPEG-2) 在市場上一直有成功的表現而帶動整個影音消費市場的成長。MPEG-4和MPEG-7就在這樣的基礎上出現。越來越多的聲音影像多媒體資料以數位的方式呈現，人們似乎也想要使用這些資料，但是，在使用之前有一個問題需要克服，就是如何找到想要的資料。在此同時，可能被使用的多媒體資料量一直增加，使得找尋資料的這個動作變得困難。在1996年10月，MPEG組織開始了一個新的工作，試圖找出一套描述影音多媒體資料的解決方案，於是便有了MPEG-7(Multimedia Content Description Interface)，且於2001年9月MPEG-7定案。配合之前的MPEG-4(Multimedia Synthetic/Natural Hybrid Coding)國際規格，成為更完整的多媒體解決方案。

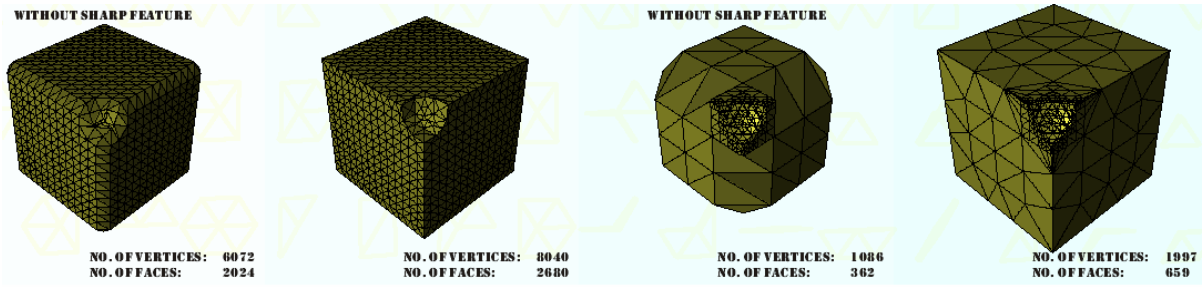
在三維MPEG-4動畫場景工作台裡，三維模型建立技術主要的目的在於提供MPEG-4場景編輯器更多的三維模型資料，當作三維物件(3D Objects)來組合建立場景(MPEG-4 Scene)。在計畫中的使用者介面，透過已發展出初步的三維模型雕塑系統，結合三維定位器，將可以用來幫助三維場景的建構與描述。三維MPEG-4動畫場景工作台將結合三維定位器(3D Tracker)、力回饋空間定位器(Force Feedback Space Tracker)、和半球形顯示器(Spherical Immersive Display)，透過立體顯示模式與力回饋的方式，提供一個虛擬實境的浸入式環境讓人融入其中。

在MPEG-7方面，關於管理搜尋三維物件場景，目前內容產業常遇到的問題是，累積的多媒體資料數量太過龐大，而必須要花費大量的人力和時間來整理，而即便經過整理，搜尋檢索時仍然不易找到想要的資料。以三維物件為例，除了自己建立模型之外，網路上也有非常多可以取得的三維模型。當三維模型愈來愈多的同時，如何整理、查詢和取得使用者想要的資料，變得十分重要。MPEG-7被定位為多媒體內容查詢(Multimedia Content Retrieval)的重要編碼標準，有很多述詞(Descriptor)，描述結構(Description Scheme)和Query的描述。利用MPEG-7來處理三維資料和多媒體資料庫，並與提供使用者取得所需要的定資料，對內容產業將有相當的貢獻。

三、研究方法與成果

【三維 MPEG-4 動畫場景建構工作台】

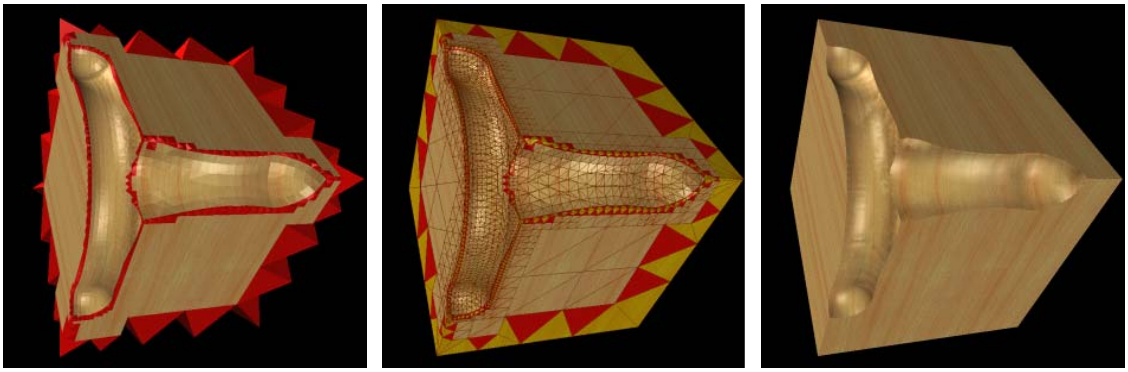
在前兩年所發展的成果中，在這個三維模型雕塑系統的實作上是使 Marching Cube 技術當成建構三維模型的基本演算法。雖然目前的系統提供了保留尖銳特徵的功能(Sharp Feature Preserving)，但由於 Marching Cube 仍然相當受限於格子切割的解析度，越精細的切割就需耗費呈三次方成長的記憶體需求。為了提高三維模型雕塑系統的精細度以及顧及系統的即時性，我們在今年進一步改良，以八元分割樹為基礎(Octree-Based)的做多階層式空間分割，針對雕塑物體的精細部分再做切割(Subdivision)以期達到即時性和雕塑品的精確性。



圖一：

在三維 MPEG-4 動畫場景建構工作台方面，三維模型雕塑系統多階層的空間分割已發展完成，如圖一與圖二所示。並套用更為精確的力回饋力量計算模型，稱為正切平面力模型(Tangent-Plane Force Model)。這套系統令使用者能以更直覺而方便的方式來建立動畫場景中的物件。使用者可以利用系統中各種不同的形狀的雕刻刀，如球形，立方體等工具，來進行三維模型的雕塑。此系統結合了三維空間定位器和力回饋裝置， 讓使用者能精準的控制雕刻刀的三維位置(Position)與方位(Orientation)。透過此人機界面，使用者也可以任意地操控三維模型雕塑系統中的物體，做任意的旋轉(Rotation)、平移(Translation)等動作。這套系統提供了兩種雕塑模式：雕刻模式(Carving Mode)與填充(Fill Mode)，亦即雕與塑。使用者可以根據自己的需求來以切割的方式產生雕塑品，或者以填充的方式補出雕塑品來。我們的雕刻系統也提供對物體簡單的著色功能，使用者可透過調整雕刻刀的大小，來控制著色區域的大小，也因此可以創造出色彩多樣化的物件。簡單的成品如圖四與圖五所示。

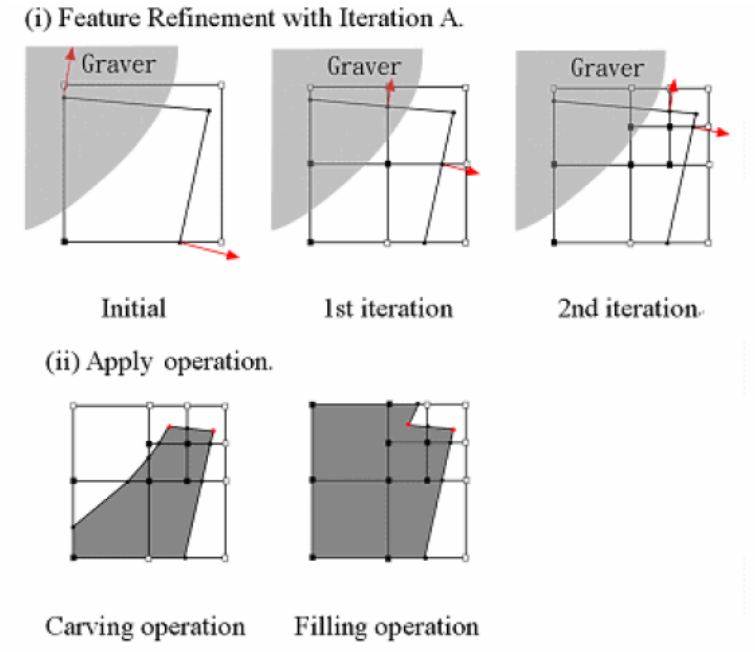
要將特徵邊角保留演算法(Feature-Preserving Algorithm)套用在多重解析度的雕塑品上，就必須在邊緣翻轉(Edge Flipping)這個部驟做修正，此套系統已發展出初步的多重邊緣翻轉演算法，如圖三所示。圖三(a)顯示多重解析度下的邊角擷取，圖三(b)顯示的是套用多重邊緣翻轉演算法之後的網格模型，圖三(c)顯示最後重建完成的雕刻面。



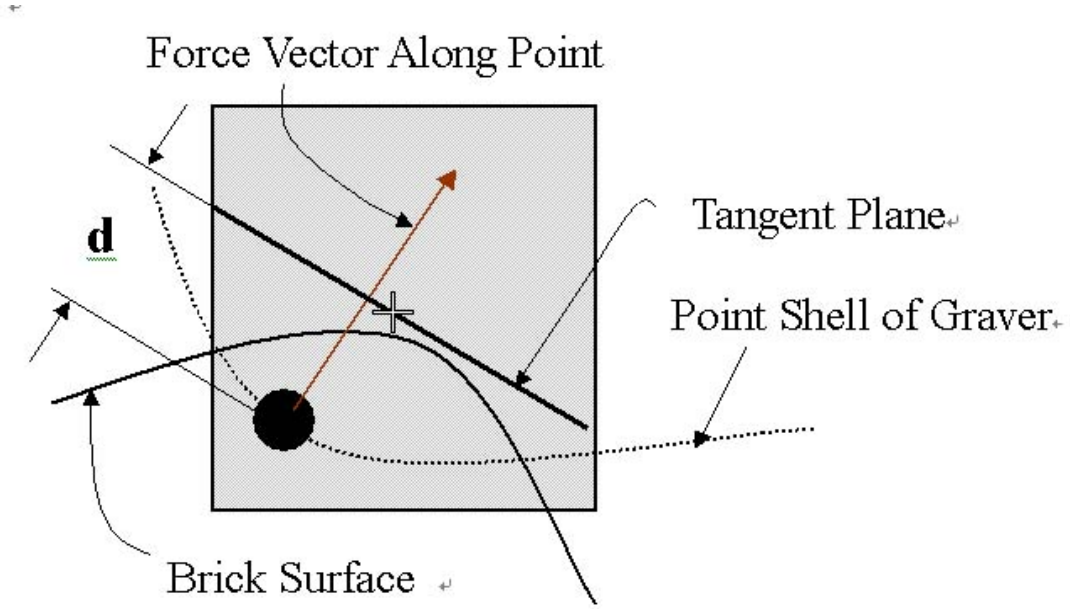
(a) (b) (c)

圖二：特徵邊角保留演算法在多重解析度下的特徵邊角重建過程與最後結果。

動畫場景建構平台除了提供原有的 Extended Marching Cube (EMC)來保留雕塑物品的特徵邊緣之外，也引進了多重解析度的概念。利用雕刻面的高斯曲率(Gaussian Curvature)來作為多接層的空間分割的條件。當雕刻工具的曲率值越大時，就以越精細的解析度來表示。此概念使得本系統可以做出比單一解析度系統更加精細的模型，如圖四與圖五所示，這樣的改進使得此系統得以更接近實用階段。力回饋方面的力計算模型，我們採用了 ACM SIGGRAPH 1999 的一篇論文 Six Degree-of-Freedom (DOF) Haptic Rendering Using Voxel Sampling，並將其中的正切平面力模型套用在多重解析度八元樹上，以計算出更加精確的碰撞力量。



圖三：多重解析度下的特徵邊角重建演算法範例。

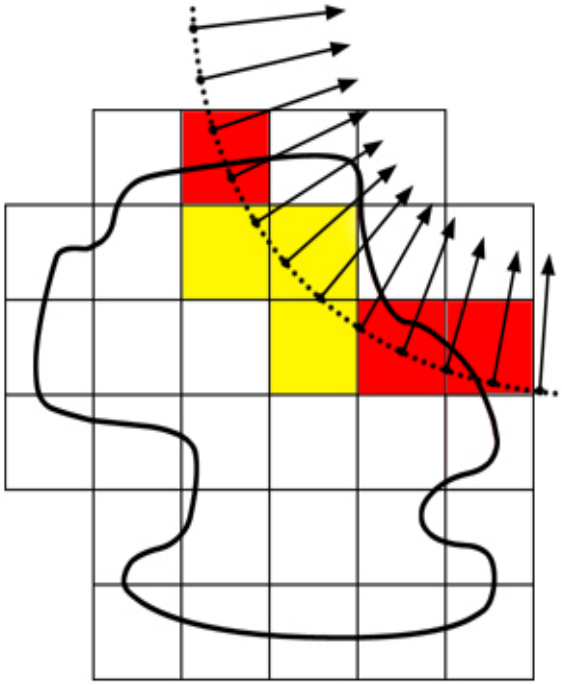


圖四：正切平面力模型。

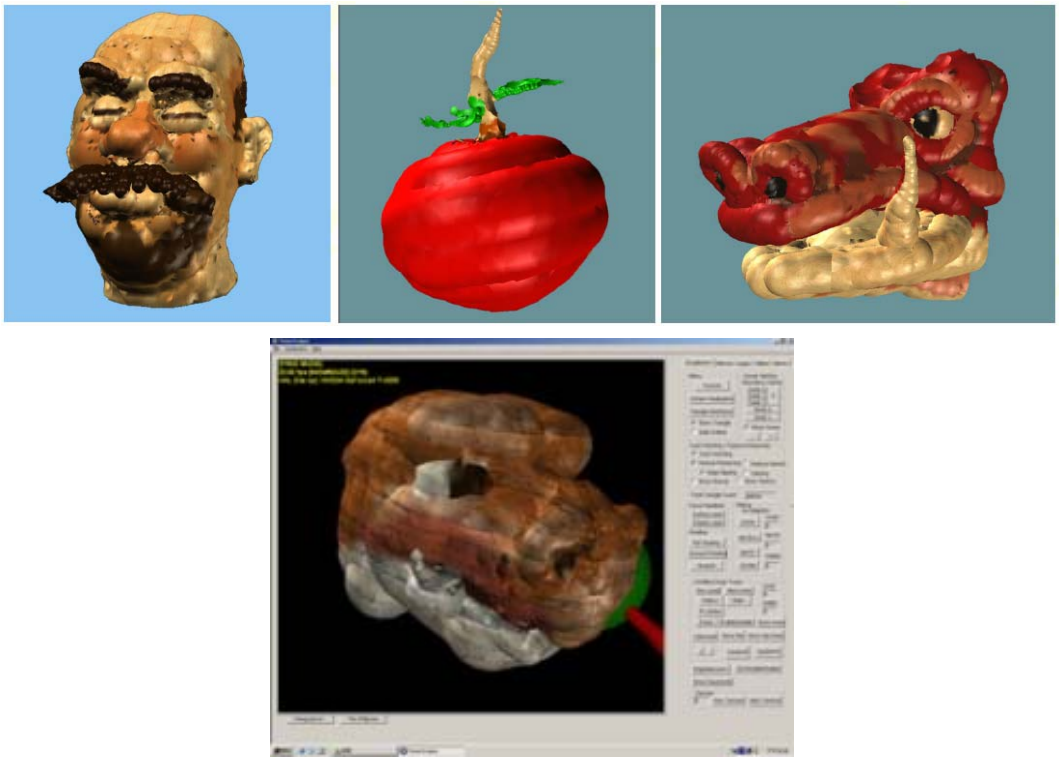
關於力回饋系統，我們使用正切平面力模型 (Tangent-Plane Force Model)，進一步解釋如下：在我們的雕刻系統中，我們用PHANToM DESKTOP[1]這個力回饋裝置做為我們的虛擬雕刻刀，並且採用了正切平面力模型的概念來模擬當雕刻刀與積木接觸時的碰撞力。我們首先在雕刻刀的表面去取樣點(Voxelization)，並將積木格子化以產生Voxmap；當雕刻刀表面的點和積木上的格子相交的時候(如圖四)，我們就計算出一個穿透距離d，並將這個距離值套用在Spring-Damper系統上(F = kd - vb)。

我們以 2D 的圖示來說明如何在雕刻時做即時運算，我們將積木做格子化之後(如圖五)，我們將這些格子分類，主要是分成表面層(Surface Layer)及裡層(Inner Layer)，當我們的雕刻刀與積木接觸的時候，首先我們先判斷雕刻刀上的點是與哪一層的格子相交，若是與表面層的相交，我們會用先前提到的正切平面力模型來計算穿透距離 d，若是與裡層的格子相交，我們則會以這個格子的對角線長度做為穿透距離，這樣做的原因是希望以較大的力來防止雕刻刀穿透過積木。最後我們以這些距離的總合來做為 Spring-Damper 系統中的 d 值。

[1] PHANToM Desktop 是 Sensable 公司所出品的力回饋裝置。http://www.sensable.com/



圖五：雕刻刀上的點與表面層相交的格子以紅色表示，與裡層相交的則以黃色表示。紅色的格子我們以正切-平面 力模型來穿透距離 d，而黃色的格子則以格子的對角線來做為穿透距離。



圖六：一個恐龍頭部模型的雕塑過程。

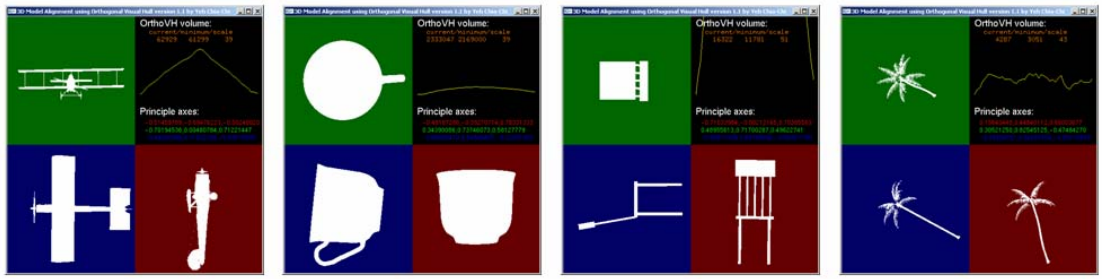
【三維物件搜尋器】

本年度的工作以先前的成果為基礎，在搜尋方式、速度以及正確率方面皆所進展。於今年年初由美國普林斯頓大學(Princeton University)所做的統合性研究中指出，去年度本實驗室的三維比對系統 LightField Descriptors (LFD)與全世界一些有名的方法比較後，在正確率上高居第一名。在提高線上比對速度方面，對每一個三維物件，新的 Characteristic Views Descriptor (CVD)採用了比較有代表性的六張二維投影，這六張投影的取得是先透過所提出的 Orthogonal Visual Hull 技術將三維物件盡量擺正後（見圖七），再分別於 x、y、z 軸的兩個方向上作投影。相較去年度所提出的 LFD 使用的 100 張投影，張數上的減少大大地加快了線上的比對速度，加上是使用比較有代表性的六張，使得比對正確率不會因為使用比較少的資訊而變太差。

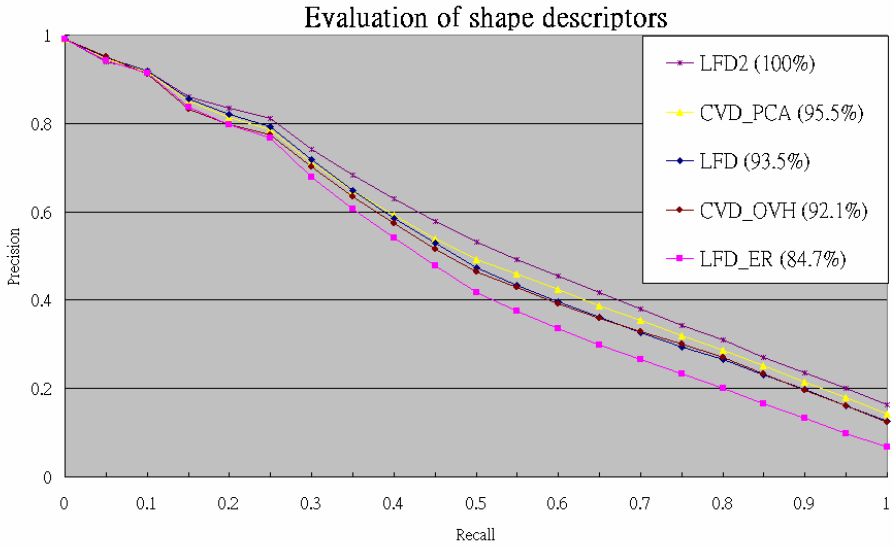
本年度把三維物件的深度值加入考量，使得所使用的述詞較去年多一，即 Fourier Descriptor(輪廓)、Zermike Moments Descriptor(形狀)、Circularity、Eccentricity 及 Generic Fourier Descriptor(深度值，見公式)。

$$PF(\rho,\theta)=\sum_{r}\sum_{i} f(r,\theta i)\exp[j2\pi(\frac{r}{R}\rho+\frac{2\pi j}{T}\theta)]$$

於圖八，原來的系統為圖中的 LFD，其比對率定為 93.5%。我們首先只把深度值加入原系統後（LFD2），比對正確率為原來的 106%。接著採用求得到的六張投影來取代原系統的 100 張（CVD_OVH），比對率為原來系統的 98%。接著由表一可得知，在儲存空間上，原有的系統需要 4700 bytes，新系統只需要 654 bytes。線上比對速度由原來的 3.233 毫秒變為 0.072 毫秒，增快了約 50 倍。



圖七：經由 Orthogonal Visual Hull 擺正後的三維物件，由 x、y、z 軸上的正投影範例。



圖八：這兩年度三維模型搜尋系統的比對準確度比較圖表。

	Storage Size (bytes)	Ratio	Off-line Generate Time (s)	Ratio	On-line Compare Time (ms)	Ratio
LFD	4,700	100%	3.25	100%	3.233	100%
LFD_ER	4,700	100%	3.25	100%	0.693	21.4%
LFD2	11,000	234%	None	None	9.988	308.9%
CVD_OVH	654	13.9%	17.31	532.6%	0.072	2.2%
CVD_PCA	654	13.9%	None	None	0.072	2.2%

表一：三維模型搜尋系統的時間與空間效能比。

四、總結本計畫第三年的成果摘要

第三年所需發展之技術要點如下：

1. 定義與推算符合 MPEG-7 精神之三維模型述詞與描述結構之加強與改進

除了將第一年所發展的述詞與描述結構加以改進擴充外，也必須由實驗找出其他適用於三維模型(x,y,z)與四維虛擬 3D 動畫(x,y,z,t)之述詞與描述結構。或者，找出適當的方法將二維影像(x,y)的述詞與描述結構擴充，使其適用於三維物件模型的搜尋、比對。倘若有合適的述詞與描述結構可提案送 MPEG-7 標準制定審查委員會，以列入國際標準。

2. 研發並實作利用多重解析度比對方法之三維模型搜尋引擎

利用前述符合 MPEG-7 精神之述詞與描述結構，發展三維模型搜尋引擎。在比對三維模型之述詞時，利用多重解析度(Multi-resolution)的比對方法，在較粗糙解析度時先事先淘汰(early jump out)一些相差很多的三維模型，然後剩下的三維模型再套用同樣地技巧，一層層地以較精細的解析度來比較與淘汰，最後再列出搜尋得到的結果，因此可以增加三維模型比對與搜尋的速度。

3. 研發並實作三維模型之分類與索引技術

為了加速三維模型搜尋的速度，有必要先對資料庫中的三維模型進行分類與索引。資料庫內的三維模型根據所發展出來的各種述詞事先分類或索引後，將使得三維模型搜尋引擎更有效率。另外，藉由分類技術，可將一些在三維模型中已知的關鍵詞，套用到其他形狀相似的三維模型上。

4. 不同三維模型述詞權重之調整與選擇

每種述詞有其特性與優缺點，很難只單用某一述詞而得到使用者所想要且滿意的結果。如果提供使用者自由挑選及調整不同三維述詞權重的組合，便可以讓使用者更容易地得到想要的三維模型。另一方面，如果事先研究並調配出三維模型述詞間合適的權重，結合不同述詞間的優點與特性，讓使用者也可以更方便地得到他所想要的三維模型。

五、結論與討論

有關三維模型搜尋器(Query by Example in 3D)：

預期工作項目	實際工作成果	說明
繼續發展以範例與關鍵字為搜尋索引之三維模型搜尋技術	已發展已關鍵字、顏色、形狀等多重輸入的三維模型搜尋技術	符合
且更進一步研發三維模型的不同性質述詞	研究出新的述詞 Orthogonal Visual Hull 並在國際會議發表	符合
將三維 MPEG-4 動畫場景工作台與三維模型搜尋器加以整合	可以將動畫場景工作台得到的三維 MPEG-4 模型，輸入三維模型搜尋器中。同時三維模型搜尋器找到的模型，也可由工作台開啟。	符合

有關三維 MPEG-4 動畫場景建構工作台(3D Workbench for MPEG-4 Scene and Animation Construction)：

預期工作項目	實際工作成果	說明
增強 MPEG-4 三維虛擬場景之力回饋反應	使用了新的力學模型與計算方式，使得使用者在操作的時候得到更佳的觸感	符合
三維模型後處理之技術	獨立實作出模型中軸自動產生、動畫三維模型中軸驅動套用、具調節三維動畫模型。	符合
傳統模型的多邊形資料(Polygonal Data)與工作站所用的體積格資料(Volume Data)做轉換用的介面需要定義與開發	已實作將 Maya OBJ 檔案與體積格資料轉換的程式	符合
將三維 MPEG-4 動畫場景工作台改寫成為三維建模動畫軟體的外掛程式(Plug-ins)，如 Maya 軟體	單獨實作出 Maya plug-ins 可做出三維模型中軸自動產生後處理。另一方面將 Maya 場景物件與動畫場景建構工作台做互相轉換。	符合

六、相關論文完成

1. Chien-Chang Ho, Yan-Hong Lu, Hung-Te Lin, Shuen-Huei Guan, Sheng-Yao Cho, Rung-Huei Liang, Bing-Yu Chen, and Ming Ouhyoung, “Feature Refinement Strategy for Extended Marching Cubes: Handling on Dynamic Nature of Real-time Sculpting Application”, **International Conference on Multimedia and Expo 2004**, Taipei, Taiwan, May 2004.

2. Shuen-Huei Guan, Ming-Kei Hsieh, Chia-Chi Yeh and Bing-Yu Chen, “Enhanced 3D Model Retrieval System through Characteristic Views using Orthogonal Visual Hull”, **SIGGRAPH 2004** (poster section), Los Angeles, USA, Aug. 2004.

3. Ding-Yun Chen, Xiao-Pei Tian, Yu-Te Shen and Ming Ouhyoung, “On Visual Similarity Based 3D Model Retrieval”, **Eurographics 2003**, Granada, Spain, Sep. 2003 (also in Computer Graphics Forum, Vol. 22, No. 3).

4. Wan-Chun Ma, Fu-Che Wu, Ming Ouhyoung, “Skeleton Extraction of 3D Objects with Radial Basis Functions”, **Shape Modeling International 2003**, Seoul, Korea, May 2003.

5. Pin-Chou Liu, Fu-Che Wu, Wan-Chun Ma, Rung-Huei Liang, Ming Ouhyoung, “Automatic Animation Skeleton Construction Using Repulsive Force Field” **Pacific Graphics 2003** (poster section, accepted), Canmore, Alberta, Canada, Oct. 2003.

6. Fu-Che Wu, Wan-Chun Ma, Ping-Chou Liou, Rung-Huei Laing, Ming Ouhyoung, “Skeleton Extraction of 3D Objects with Visible Repulsive Force”, currently submitted.

7. Ding-Yun Chen, “On Visual Similarity Based 3D Model Retrieval”, Ph.D. thesis, Dept. of Computer Science and Information Engineering.

8. Xiao-Pei Tian, “A 3D Model Retrieval System Based on MPEG-7 3DSSD”, Master thesis, Dept. of Computer Science and Information Engineering.

9. Shuen-Huei Guan, “A Feature Preserving Multiresolution Volumetric Sculpting System”, Master thesis, Dept. of Computer Science and Information Engineering.

10. Shung-Yao Cho, “Constructing Scalable 3D Animated Model by Deformation Sensitive Simplification”, Master thesis, Dept. of Computer Science and Information Engineering.

七、參考文獻

其餘參考文獻請參照分項計畫三計畫書參考文獻。

八、相關論文成果 (見後頁)

EUROGRAPHICS 2003 / P. Brunet and D. Fellner
(Guest Editors)

Volume 22 (*2003*), *Number 3*

On Visual Similarity Based 3D Model Retrieval

Ding-Yun Chen, Xiao-Pei Tian, Yu-Te Shen and Ming Ouhyoung

Department of Computer Science and Information Engineering,
National Taiwan University, Taipei, Taiwan
{dynamic, babylon, edwards}@cmlab.csie.ntu.edu.tw, ming@csie.ntu.edu.tw

Abstract

A large number of 3D models are created and available on the Web, since more and more 3D modelling and digitizing tools are developed for ever increasing applications. The techniques for content-based 3D model retrieval then become necessary. In this paper, a visual similarity-based 3D model retrieval system is proposed. This approach measures the similarity among 3D models by visual similarity, and the main idea is that if two 3D models are similar, they also look similar from all viewing angles. Therefore, one hundred orthogonal projections of an object, excluding symmetry, are encoded both by Zernike moments and Fourier descriptors as features for later retrieval. The visual similarity-based approach is robust against similarity transformation, noise, model degeneracy etc., and provides 42%, 94% and 25% better performance (precision-recall evaluation diagram) than three other competing approaches: (1)the spherical harmonics approach developed by Funkhouser et al., (2)the MPEG-7 Shape 3D descriptors, and (3)the MPEG-7 Multiple View Descriptor. The proposed system is on the Web for practical trial use (http://3d.csie.ntu.edu.tw), and the database contains more than 10,000 publicly available 3D models collected from WWW pages. Furthermore, a user friendly interface is provided to retrieve 3D models by drawing 2D shapes. The retrieval is fast enough on a server with Pentium IV 2.4GHz CPU, and it takes about 2 seconds and 0.1 seconds for querying directly by a 3D model and by hand drawn 2D shapes, respectively.

Categories and Subject Descriptors (according to ACM CCS): H.3.1 [Information Storage and Retrieval]: Indexing Methods

1. Introduction

Recently, the development of 3D modeling and digitizing technologies has made the model generating process much easier. Also, through the Internet, users can download a large number of free 3D models from all over the world. This leads to the necessities of a 3D model retrieval system. Although text-based search engines are ubiquitous today, multimedia data, such as 3D models, usually lacks meaningful description for automatic matching. The MPEG group aims to create an MPEG-7 international standard, also known as "Multimedia Content Description Interface", for the description of multimedia data [11]. However, little description is about 3D models. The need of developing efficient techniques for content-based 3D model retrieval is increasing.

To search 3D models that are visually similar to a queried model is the most intuitive way. However, most methods concentrate on the similarity of geometric distributions rather than directly searching for visually similar models. The geometric-based approach is feasible since much appearance for an object is controlled by its geometry. In this paper, however, we present a novel approach that matches 3D models using their visual similarities, which are measured with image differences in light fields. We take this approach to better fit the goal of comparing models that appear to be similar to a human observer. The concept of the visual similarity-based approach is similar to that of the image-driven simplification, proposed by Lindstrom and Turk [14].

The geometry-based approach is broadly classified into two categories: shape-based and topology-based matching. The shape-based approach uses the distribution of vertices or polygons to judge the similarity between 3D models [1, 2, 4, 5, 6, 7]. The challenge of the shape-based approach is how to define shape descriptors, which need to be sensitive, unique, stable, efficient, and robust against similarity transformations of various kinds of 3D models. The topology-based approach utilizes topological structures of 3D models

to measure the similarity between them [3]. The difficulties of the topology-based approach include automatic topology extraction from all types of 3D models, and the discrimination between topologies from different categories. Each of the two approaches has its inherent merits and demerits. For example, a topology-based approach leads to high similarity between two identical 3D models with different gestures, whereas a shape-based approach cannot. On the other hand, a shape-based approach results in high similarity between 3D models with different connections among parts, whereas a topology-based approach cannot. For instance, both a finger and the shoulder of a human model are parts of a human body. The topologies are quite different whether the finger does or does not connect to a human body, but the shapes are similar.

Most previous works of 3D model retrieval focused on defining suitable features for the matching process [1~13], and were based on either statistical properties, such as global shape histograms, or the skeletal structures of 3D models. Osada et al. [2] proposed and analyzed a method for computing shape signatures of arbitrary 3D polygonal models. The key idea is to represent the signature of a 3D model as a shape distribution, which is a histogram created from the distance between two random points on a surface for measuring global geometric properties. The approach is simple, fast and robust, and could be applied as a pre-classifier in a complete 3D model retrieval system.

Funkhouser et al. [1] proposed a practical web-based search engine that supports queries based on 3D sketches, 2D sketches, 3D models, and/or text keywords. For 3D shape queries, a new matching algorithm that uses spherical harmonics to compute similarities is developed. It does not require repair of model degeneracy or alignment of orientations. In their system, a multimodal query is applied to increase the retrieval performance by combining features such as text and 3D shapes. It is also fast enough to retrieve from a repository of 20,000 models in less than one second.

Hilaga et al. [3] proposed a technique in which the similarity between polyhedral models is accurately and automatically calculated by comparing the skeletal and topological structure. The skeletal and topological structure decomposes a 3D model to a one-dimensional graph structure. The graph is invariant to similarity transformations, robust against simplification and deformation caused by changing posture of an articulated object, etc. In their experimental results, the average search time from 230 3D models is about 12 seconds with a Pentium II 400MHz processor. Another 3D model retrieval system [10] , having 445 models in the database, is extended from the work of Hilaga et al., and takes about 12 seconds on a server with Pentium IV 2.4 GHz processor.

In this paper, a novel visual similarity-based approach for 3D model retrieval is proposed, and the system is also available on the web for practical trial use. The proposed approach is robust against similarity transformations, noise and model degeneracy, etc. There are more than 10,000 3D models in our database, and a user-friendly interface is provided for 3D model retrieval by drawing 2D shapes, which are taken as one or more projection views.

In general, a retrieval system contains off-line feature extraction and on-line retrieval processes. We introduce the *LightField Descriptor* to represent 3D models, which is detailed in Section 2, as well as the feature extraction. In Section 3, comparing 3D models is represented for the on-line retrieval process. The experimental results and the performance evaluations are shown in Section 4. Section 5 concludes the write up.

2. Feature Extraction for Representing 3D Models

The proposed descriptor used for comparing the similarity among 3D models is extracted from 4D light fields, which are representations of a 3D object. The phrase light field describes the radiometric properties of light in a space and was coined by Gershun [23]. A light field (or plenoptic function) is traditionally used in image-based rendering and is defined as a five dimensional function that represents the radiance at a given 3D point in a given direction [24, 25]. For a 3D model, the representation is the same along a ray, so the dimension of the light field around an object can be reduced to 4D [25, 14]. Each 4D light field of a 3D model is represented by a collection of 2D images, which are rendered from a 2D array of cameras. The camera positions of one light field can be put either on a flat surface [25] or on a sphere [14] in the 3D world. The light field representation has not only been used in image-based rendering, but also in image-driven simplification by Lindstrom and Turk [14], whose approach uses images to decide which portions of a model to simplify.

In this paper, we extract features from the light fields rendered from cameras on a sphere. The main idea of using the approach to get the similarity between two models is introduced in Section 2.1. To reduce size of the features and speed up the matching process, the cameras of the light fields are distributed uniformly and positioned on vertices of a regular dodecahedron. We name the descriptor *LightField Descriptor*, and describe it in Section 2.2. In Section 2.3, we show that one 3D model is represented by a set of *LightField Descriptors* in order to improve the robustness against rotations while comparing between two models. One that is also important is the image metric used, and is detailed in Section 2.4. Finally, the flow chart of extracting the *LightField Descriptors* for a 3D model is summarized in Section 2.5.

2.1. Measuring similarity between two 3D models

The main idea comes from the following statement, "If two 3D models are similar, they also look similar from all viewing angles." Accordingly, the similarity between two 3D models can be measured by summing up the similarity from all corresponding images of a light field. However, what

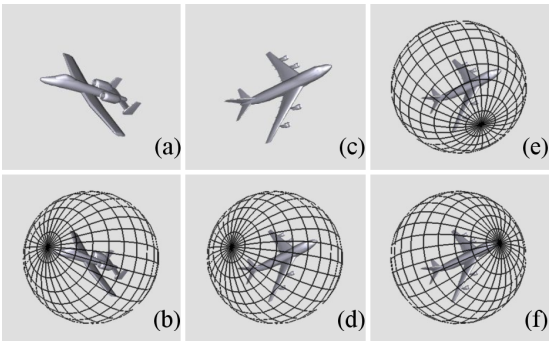


Figure 1: *The main idea measuring similarity between two 3D models*

must be considered is the transformation, including translation, rotation and scaling. The translation and the scaling problems are discussed in Section 2.5 and ignored by our image metric described in Section 2.4. As for rotation, the key to this problem is visual similarity. The camera system surrounding each model is rotated until the highest overall similarity (cross-correlation) between the two models from all viewing angles is reached. Take Figure 1 as an example, where (a) and (c) are two different airplanes with inconsistent rotations. First, for the airplane in Figure 1 (a), we place the cameras of a light field on a sphere, as shown in Figure 1 (b), where cameras are put on the intersection points of the sphere. Then, cameras of this light field can be applied, at the same positions, to the airplane in Figure 1 (c), as shown in Figure 1 (d). By summing up the similarities of all pairs of corresponding images in Figure 1 (b) and (d), the overall similarity between the two 3D models is obtained. Next, the camera system in Figure 1 (d) can be rotated to a different orientation, such as Figure 1 (e), which leads to another similarity value between the two models. After evaluating similarity values, the correct corresponding orientation, in which the two models look most similar from all corresponding viewing angles, can be found, such as Figure 1 (f). The similarity between the two models is defined as summing up the similarity from all corresponding images between Figure 1 (b) and (f).

However, the computation will be very complicated and impractical to a 3D model retrieval system using current processors. Therefore, the camera positions of a light field are distributed uniformly on vertices of a regular dodecahedron, such that reduced camera positions are used for approximation.

2.2. A *LightField Descriptor* for 3D models

To reduce the retrieval time and the size of features, the light field cameras can be put on 20 vertices of a regular dodecahedron. That is, there are 20 different views, which are distributed uniformly, over a 3D model. The 20 views can

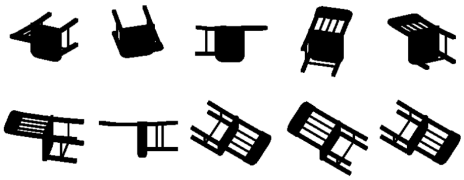


Figure 2: *A typical example of the 10 silhouettes for a 3D model*

roughly represent the shape of a 3D model, as been applied similarly in previous works. Huber and Hubert [27] proposed an automatic registration algorithm, which is able to reconstruct real world objects from 15 to 20 various viewpoints from a laser scanner. Lindstrom and Turk [14] also employ the 20 views in comparing 3D models for image-driven simplification.

Since their applications are different from the retrieval system, the requirements of rendering image and the image metric used are also different. First, lighting is different while rendering images of an object. We turn all lights off, so that the rendered images will be silhouettes only, which enhance the efficiency and the robustness of image metric. Second, orthogonal projection is applied in order to speed up the retrieval process and reduce the size of features. Therefore, ten different silhouettes are produced for a 3D model, since the silhouettes projected from two opposite vertices on the dodecahedron are identical. Figure 2 shows a typical example of the 10 silhouettes of a 3D model. In our implementation, the rendered image size is 256 by 256 pixels. Consequently, the rendering process can filter out high-frequency noise of 3D models, and also make our approach reliable from degeneracy of meshes, such as those missing, wrongly-oriented, intersecting, disjoint and overlapping polygons.

Since the cameras are placed on the vertices of a fixed regular dodecahedron, we need to rotate the camera system 60 times (to be explained below), so that the cameras can be switched onto different vertices, while measuring the similarity between descriptors of two 3D models. The dissimilarity, D_A, between two 3D models is defined as:

$$D_A = \min_i \sum_{k=1}^{10} d(I_{1k}, I_{2k}), \qquad i = 1..60 \qquad (1)$$

where d denotes the dissimilarity between two images, defined in Section 3.1, and i denotes different rotations between camera positions of two 3D models. For a regular dodecahedron, each of the 20 vertices is connected by 3 edges, which results in 60 different rotations for one camera system (mirror mapping is not available). I_{1k} and I_{2k} are corresponding images under i-th rotation.

Here is a typical example to explain our approach. There are two 3D models, a *pig* and a *cow*, in Figure 3 (a), both

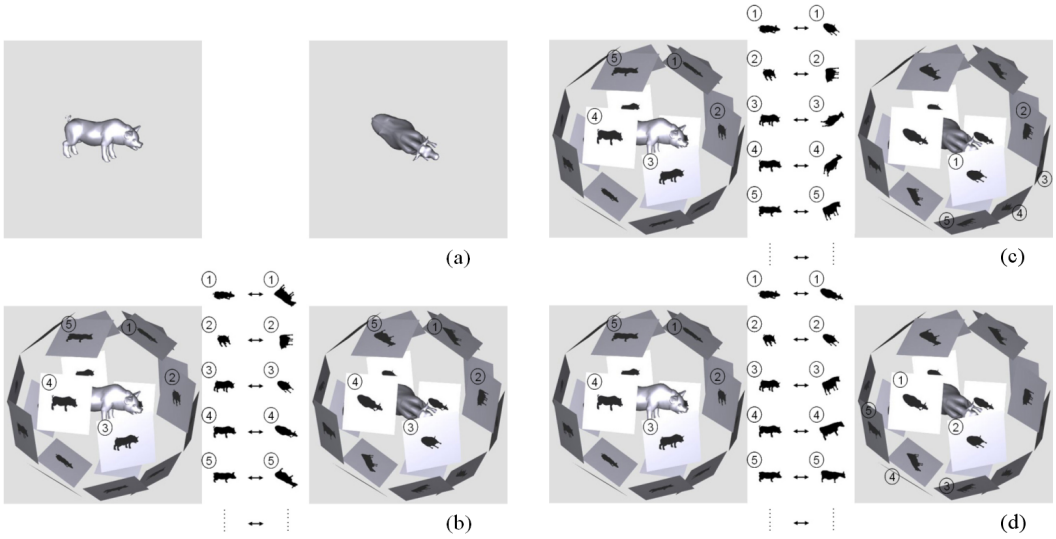


Figure 3: *Comparing LightField Descriptors between two 3D models*

rotated randomly. First, 20 images are rendered from vertices of a dodecahedron for both the 3D model. As shown in Figure 3 (b), we compare all the corresponding 2D images from the same viewing angles, such as, the order 1∼5 between *pig* and *cow* model. Thus we get a similarity value under this rotation of camera system. Then, we map the order 1∼5 differently as in Figure 3 (d), and get another similarity value. After repeating this process, we find a rotation of camera positions with the best similarity (cross-correlation being highest), as shown in Figure 3 (d). Therefore, the similarity between the two models is the summation of the similarities among all the corresponding images.

Consequently, the *LightField Descriptor* is defined as the basis representation of a 3D model, and is defined as features of 10 images rendered from vertices of dodecahedron over a hemisphere. A *LightField Descriptor* somehow eliminates the rotation problem, but this is not exact enough. Therefore, a set of light fields is applied to improve the robustness.

2.3. A set of *LightField Descriptors* for a 3D model

To be robust against rotations among 3D models, a set of *LightField Descriptors* is applied to each 3D model. If there are N *LightField Descriptors*, which are created from different camera system orientations for both 3D models, there are $(N \times (N-1)+1) \times 60$ different rotations between the two models. Therefore, the dissimilarity, D_B, between two 3D models is then defined as:

$$D_B = \min D_A(L_j, L_k), \qquad j, k = 1..N \tag{2}$$

where D_A is defined in Equation (1), and L_j and L_k are light field descriptors of two models, respectively.

The relationship of the N light fields needs to be carefully set to ensure that all the cameras are distributed uniformly and able to cover different viewing angles to solve the rotation problem effectively. The approach of generating the evenly distributing camera positions of the N light fields comes from the idea in relaxation of random points proposed by Turk [15]. The process can be pre-processed, and then all 3D models use the same distributed light fields to generate corresponding descriptors. In our implementation, we set $N = 10$, as shown in Figure 4, that is, the similarity between two 3D models is obtained from the best one of 5,460 different rotations. Therefore, the average maximum error of rotation angle between two 3D models is about 3.4 degree in longitude and latitude. That is:

$$\frac{180^\circ}{x} \times \frac{360^\circ}{x} = 5460 \;\Rightarrow\; x \cong 3.4^\circ \tag{3}$$

which is small enough for our 3D model retrieval system according to our experimental results. Of course, the number

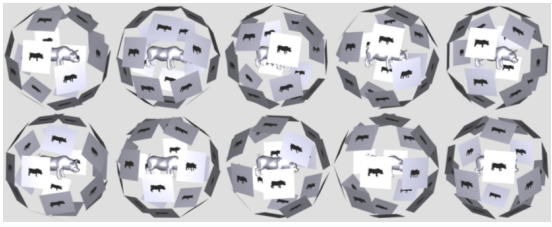


Figure 4: *A set of LightField Descriptors for a 3D model*

N can be bigger than 10, and we will evaluate in the future the saturation effect when *N* becomes bigger.

2.4. Image metric

An image metric is a function measuring the distance between two images. Recently, Content-Based Image Retrieval (CBIR) has become a popular research, and different image metrics have been proposed [19~22]. Many approaches of image metrics are robust against transformations such as translation, rotation, scaling, and image distortion.

One of the important features of images is the shape descriptor, which can be broadly classified into region-based and contour-based descriptor. The use of a combination of different shape descriptors has been proposed recently in order to improve the retrieval performance [21, 22]. In this paper, we adopt an integrated approach proposed by Zhang and Lu [21], which combines a region shape descriptor (Zernike moments descriptor) and a contour shape descriptor (Fourier descriptor). The Zernike moment descriptor is also used in MPEG-7, which is named *RegionShape* descriptor [11]. Different shape signatures have been used to derive Fourier descriptor, and the retrieval using Fourier Descriptors derived from the centroid distance has significantly higher performance than those of the other methods. These were also compared by Zhang and Lu [20]. The centroid distance function is expressed by the distance to boundary points from the centroid of the shape. The boundary points of a shape are extracted through a contour tracing algorithm, proposed by Pavlidis [31]. Figure 5 shows a typical example of the centroid distance. Figure 5 (a) shows a 2D shape rendered from a viewpoint of a 3D model, and the contour tracing result is shown in Figure 5 (b). Figure 5 (c) shows the centroid distance of (a).

Sometimes, however, a 3D model might be rendered into several separated 2D shapes, as shown in Figure 6 (a). When this situation occurs, the following two stages are applied to connect. First, we apply Erosion operation [32] from one to several times to connect the separated parts, as shown in Figure 6 (b). Second, a thinning algorithm [32] is applied in order to connect the separated parts, as shown in Figure 6 (c). Note that the pixels of rendered 2D shape cannot be removed during the thinning algorithm. The separated parts are then connected, and the high-frequency noise will be filtered out by the Fourier descriptor. But if there are still separated parts after running the Erosion operation for several times, a bounding box algorithm will replace the first stage above.

There are 35 coefficients for Zernike moment descriptor and 10 coefficients for Fourier descriptor. Each coefficient is quantized to 8 bits in order to reduce the size of descriptors and accelerate the retrieval process. Consequently, the approach is robust against translation, rotation, scaling and image distortion, and is very efficient.

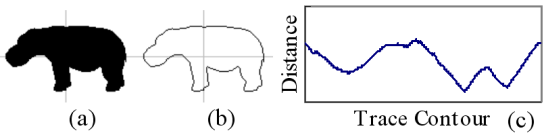


Figure 5: *A typical example of the centroid distance for a 2D shape*

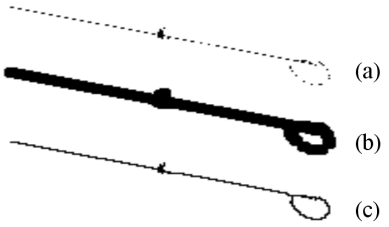


Figure 6: *Connection of different parts of 2D shapes*

2.5. Steps of extracting *LightField Descriptors* for a 3D model

The steps of extracting the *LightField Descriptors* for a 3D model are summarized in the following.

(1) Translation and scaling are applied first to ensure that 3D model is entirely contained in rendered images. The input 3D model is translated from the center of the model to the origin of world coordinate system. The axis is then scaled such that the maximum length is 1.

(2) Render images from the camera positions of light fields, as described in Section 2.3.

(3) For a *LightField Descriptor*, 10 images are represented for 20 viewpoints, and are in a pre-defined order for storage. For a 3D model, 10 descriptors are created, so that totally 100 images should be rendered.

(4) Descriptors for a 3D model are extracted from the 100 images, as in Section 2.4.

3. Retrieval of 3D Models Using *LightField Descriptors*

In the off-line process mentioned in last section, the *LightField Descriptors* of each 3D model in the database are calculated and stored for 3D model retrieval. This section details the on-line retrieving process, which compares the descriptors of the queried one with all the other 3D models in the database. Comparing the *LightField Descriptors* within two models is described in Section 3.1. Those who are greatly dissimilar to the queried model will be rejected early in the process, detailed in Section 3.2, which accelerates the retrieval with a large database. Practically, when a user wants to retrieve 3D models, he/she can upload a 3D model as a query key. However, early experiences of Funkhouser et al. [1] suggest that even a simple gesture interface, such as Teddy

system [26], is still too hard for novice and casual users to learn quickly. They proposed that drawing 2D shapes with a painting program to retrieve 3D models is intuitive for users. In this paper, a user-friendly drawing interface for 3D model retrieval is also provided. The approach of comparing 2D shapes with 3D models is described in Section 3.3.

3.1. Similarity between *LightField Descriptors* of two 3D models

The retrieving process can be referred as calculating the similarity one by one between the queried one and each of the models in the database and showing those similar to the queried one. The similarity between two models is defined as summing up the similarity from all the corresponding images, as described in Section 2.3. The comparison of two descriptors is as Equation (2). When comparing the dissimilarity, d, of corresponding images, we use simple $L1$ distance to measure:

$$d(J,K) = \sum_i |C_{1i} - C_{2i}| \tag{4}$$

where C_1 and C_2 denote coefficients of two images, and i denotes the index of their coefficients. There are 45 coefficients for each image, each quantized to 8 bits. To simplify the computation, a table is created and stored for the value of $L1$ distance from 0 to 255. Thus, a table-look-up method is used to speed up the retrieval process.

3.2. Retrieval of 3D models from database with large number of models

For a 3D model retrieval system, a database with a large number of models should be considered. For example, there are over 10,000 3D models in our database. To efficiently retrieve 3D models from an enormous database, an iterative early-jump-out method is applied in our 3D model retrieval system. First, when comparing the queried model with all the others, only parts of the images and coefficients are used. This can remove almost half of the models. The threshold of removing models is set as the mean of the similarity rather than the median, since the calculation of the mean is simpler. Then, compare the queried model to the remainder models using more images and coefficients. Repeat the above steps in several times. All iterations are detailed as follows.

(1) In the initial stage, all 3D models in the database are compared with the queried one. Two *LightField Descriptors* of the queried model are compared with ten of those in the database. Three images of each light field are compared, and each image is compared using 8 coefficients of Zernike moment. Each coefficient is quantized to 4 bits.

(2) In the second stage, five *LightField Descriptors* of the queried model are compared to ten of the others in the database. Five images of each light field are compared, and each image is compared using 16 coefficients of Zernike moment. Each coefficient is quantized to 4 bits.

(3) Thirdly, seven *LightField Descriptors* of queried model are compared with ten of the others in the database. The other is the same as the second stage, while another five images of each light field are compared.

(4) The fourth stage is the same as full comparison, but only the Zernike moment coefficients, quantized to 4 bits, are used. In addition, the top 16 of the 5,460 rotations are recorded between the queried one and others.

(5) Each coefficient of Zernike moment is quantized to 8 bits, and the retrieval uses the top 16 rotations, recorded from the 4^{th} stage, rather than 5,460 rotations.

(6) In the last stage, each coefficient of Fourier descriptor is added to the retrieval.

The approach speeds up the retrieval process by early rejection of non-relevant models. The query performance and robustness of each step will be evaluated in the future.

3.3. A user friendly interface to retrieve 3D models from 2D shapes

Creating a queried model for retrieval is not easy and fast for general users. Thus, a user-friendly interface, a painting program, is provided in our system. Furthermore, users can again utilize the retrieved model to find more specific 3D models, since 2D shapes carry less information than a 3D model. To sum up, our 3D model retrieval system is easy for a novice user.

Recognizing 3D objects from single 2D shape is an interesting and difficult problem, and has been researched long time ago. Dudani et al. [16] identified aircrafts with moment invariants derived from the boundary of 2D shapes. They captured 2D images of 3D objects from every 5 degrees of azimuth and roll angle. 3,000 images for 6 aircrafts are used for comparison with an input 2D shape. A 3D aircraft recognition algorithm of Wallace and Wintz [17] used 143 projections to represent an aircraft over the hemisphere, and performed the recognition using normalized Fourier descriptors of the 2D shape boundary. Cyr and Kimia [18] proposed 3D object recognition by generating "equivalent view" from different positions on the equator for 65 3D objects. They recognized an unknown 2D shape by comparing all views of 3D objects using shock matching, which takes about 5 minutes. Recently, Funkhouser et al. [1] proposed a search engine for 3D models, which also provides a 2D drawing interface for 3D model retrieval. The boundary contours are rendered from 13 viewpoints for each 3D model, and then additional 13 shape descriptors are created. In our 3D model retrieval system, it is intuitive and direct to compare 2D shapes with 3D models, since the descriptors for 3D models are composed of features of 100 2D shapes over the hemisphere, as

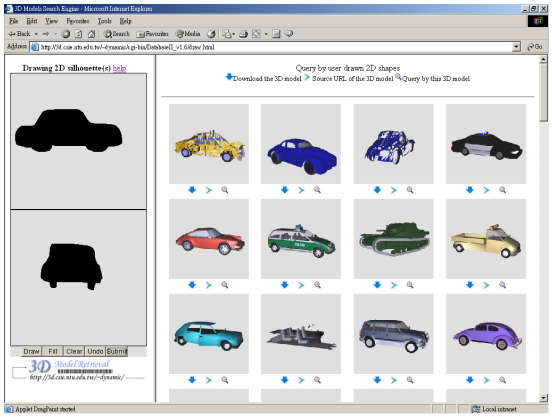


Figure 7: *Retrieval results from user drawn 2D shapes*

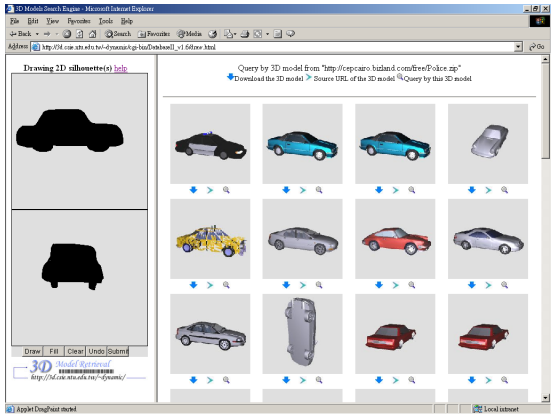


Figure 8: *Retrieval results from interactively searching of selecting a 3D model from Figure 7*

described in Section 2.3. The image metric we used is defined in Section 2.4.

4. Experimental Results

In Section 4.1, the proposed 3D model retrieval system is demonstrated. The performance and robustness of the approach are evaluated in Section 4.2 and 4.3, respectively.

4.1. The proposed 3D model retrieval system

The 3D model retrieval system is on the following web site for practical trial use: *http://3d.csie.ntu.edu.tw*. There are 10,911 3D models in our database now, all free downloaded via the Internet. Users can query with a 3D model or drawing 2D shapes, and then search interactively and iteratively for more specific 3D models using the first retrieved results. Models are available for downloading from the hyperlink of their original downloaded path listed in the retrieval results. Figure 7 shows a typical example of a query with 2D drawing shapes, and Figure 8 shows the interactive search by selecting a 3D model from Figure 7.

The system consists of off-line feature extraction in preprocessing and on-line retrieval processes. In the off-line process, the features are extracted in a PC with a Pentium III 800MHz CPU and GeForce2 MX video card. On the average, each 3D model with 7,540 polygons takes 6.1 seconds to extract features, detailed in Table 1. Furthermore, the average time of rendering and feature extraction for a 2D shape takes 0.06 seconds. Extracting features are suitable for both 3D model and 2D shape matching. No extra effort should be done for 2D shapes. In the on-line process, the retrieval is done in a PC with two Pentium IV 2.4GHz CPUs. Only one CPU is used for the query at one time, and the retrieval takes 2 and 0.1 seconds with a 3D model and two 2D shapes as the query keys, respectively.

4.2. Performance Evaluation

Traditionally, the diagram of "Precision" vs "Recall" is a common way of evaluating performance in documental and visual information retrieval. Recall measures the ability of the system to retrieve all models that are relevant. Precision measures that the ability of the system to retrieve only models that are relevant. They are defined as:

$$Recall = \frac{relevant\ \ correctly\ \ retrieved}{all\ \ relevant}$$

$$Precise = \frac{relevant\ \ correctly\ \ retrieved}{all\ \ retrieved}$$

In general, the recall and precision diagram requires a ground truth database to assess the relevance of models with a set of significant queries. Test sets are usually large, but only a small fraction of the relevant models are included [30]. Therefore, a test database with 1,833 3D models is used for evaluation. The test database contains free 3D models from 3DCafe [34], downloaded in Dec. 2001, but removes several models with failed formats in decoding. One student independent of this research, regarded as a human evaluator, classified the models according to functional similarities. The test database was clustered into 47 classes including 549 3D

	Average	Standard Deviation	Minimum	Maximum
Vertex	4941.6	13582.8	4	262882
Polygon	7540.7	21003.8	2	519813
Time	6.11 sec	4.38 sec	2.32 sec	48.93 sec

Table 1: *Vertex and polygon number of the 10,911 3D models and the feature extraction time from a PC with a Pentium III 800 MHz CPU*

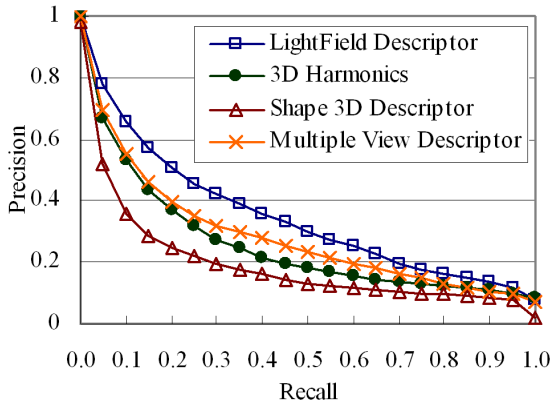


Figure 9: *Performance evaluation of our approach,* Light-Field Descriptor*, and those of others.*

models mainly for vehicle and household items (such as categories of airplane, car, chair, table, etc.), and all the other 1,284 models are classified as "miscellaneous".

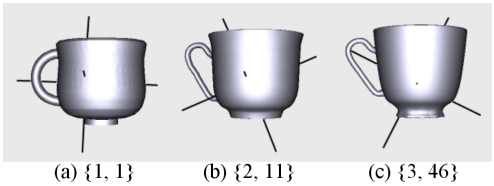
To compare the performance with others systems, three major previous works are implemented as follows:

(1) 3D Harmonics: This approach is proposed by Funkhouser et al. [1], and outperforms many other approaches, such as Moments [12], Extended Gaussian Images [8], Shape Histograms [9] and D2 Shape Distributions [2], which are evaluated in their paper. The source code of SpharmonicKit 2.5 [35], also used in their implementation, is used for computing the spherical harmonics.

(2) Shape 3D Descriptor: The approach is used in MPEG-7 international standard [11], and represents a 3D model with curvature histograms.

(3) Multiple View Descriptor: This method aligns 3D objects with Principal Component Analysis (PCA) [33], and then compares images from the primary, secondary and tertiary viewing directions of principal axes. Descriptor of the viewing directions is also recorded in MPEG-7 international standard [11], but does not limit the usage of image metrics. To get better performance, integration with different image metrics described in Section 2.4 are used. Furthermore, for calculating PCA correctly from vertices, each 3D model is resampled first to ensure that vertices are distributed evenly on the surface.

Figure 9 shows the comparison of the retrieval performance of our approach, *LightField Descriptors*, with those of the others. Each curve plots the graph of "recall and precision" averaged over all 549 classified models in the test database. Obviously, *LightField Descriptor* performs better than the others. The precision values are 42%, 94% and 25% higher than those of 3D Harmonics, Shape 3D Descriptor and Multiple View Descriptor, respectively, after comparing and averaging over all the "recall" axis.



(a) {1, 1} (b) {2, 11} (c) {3, 46}

Figure 10: *Three similar cups with their principal axes, orienting the models in different directions. Retrieval results of querying are done by the model in (a). The first number in bracket shows the queried number by our method, and the second number shows the Multiple View Descriptor.*

However, in our implementation of 3D Harmonics, the precision is not as good as that indicated in the original paper [1], shown in Table 2. Evaluating by different test database is one possible reason, and another one may lie in a small amount of different details between our implementation and original paper, even if we try to implement the same as the original paper. The test database used in the original paper is also purchased by us, and will be evaluated in the future. As for PCA applied to Multiple View Descriptor, Funkhouser et al. [1] found that principal axes are not good while aligning orientations of different models within the same class, and also demonstrated this problem using 3 mugs. Retrieval with similar examples in our test database is shown in Figure 10. Clearly, our approach works well against this particular problem of PCA.

4.3. Robustness evaluation

All the classified 3D models in the test database are applied to the following evaluation in order to assess the robustness. Each transformed 3D model is then used for queries from the test database. The average recall and precision of all 549 classified models are used for the evaluation. The robustness is evaluated by the following transformation:

(1) Similarity transformation: For each 3D model, seven random numbers are applied to x-, y-, and z-axis rotations (from 0 to 360 degree), x-, y- and z-axis translations (-10∼+10 times of the length of the model's bounding box), and scaling (a factor of -10∼+10).

(2) Noise: Each vertex of 3D model is applied three ran-

Recall	0.2	0.3	0.4	0.5	0.6	0.7
Our approach	0.51	0.42	0.36	0.30	0.25	0.20
3D Harmonics	0.37	0.27	0.22	0.18	0.16	0.14
3D Harmonics with different test database [1]	0.41	0.33	0.26	0.20	0.17	0.14

Table 2: *Precision of 3D Harmonics in the original paper for comparison.*

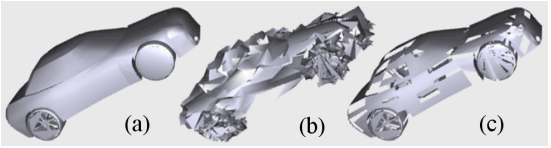


Figure 11: *Robustness evaluation of (b) noise and (c) decimation from (a) original 3D model*

dom number to x-, y- and z-axis translation (-3%∼+3% times of the length of the model's bounding box). Figure 11 (b) shows a typical example of the effect.

(3) Decimation: For each 3D model, randomly select 20% polygons to be deleted. Figure 11 (c) shows a typical example of the effect.

Experimental result of the robustness evaluation is shown in Figure 12. Clearly, our approach is robust against similarity transformation, noise and decimation.

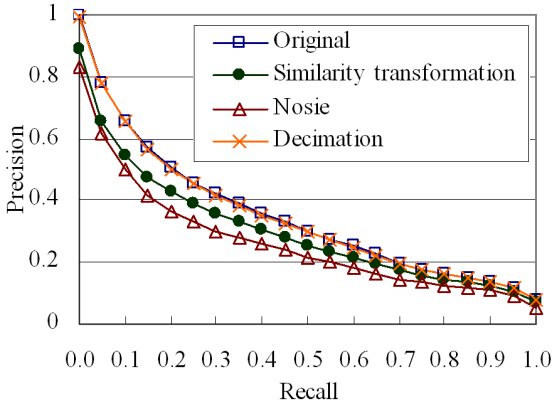


Figure 12: *Robustness evaluation of similarity transformation, noise and decimation*

5. Conclusion and Future Works

In this paper, a 3D model retrieval system is proposed based on visual similarity. The new metric based on a set of *Light-Field Descriptors* is proposed for matching among 3D models. The visual similarity-based approach is robust against translation, rotation, scaling, noise, decimation and model degeneracy etc. A practical retrieval system that includes more than 10,000 3D models is available on the web for expert and novice users, and the retrieval can be done in less than 2 seconds on a server with Pentium IV 2.4 GHz CPU. A friendly user interface is also provided to query by drawing 2D shapes. The experimental results demonstrate that our approach outperforms 3D Harmonics, MPEG-7 Shape 3D Descriptor and Multiple View Descriptor.

In future work, several investigations are described as follows. First, other image metric for 2D shapes matching may be evaluated and included to improve the performance. In addition, the image metric for color and texture [11] can also be included to retrieval 3D model using more visual features. Second, different approaches ("cocktail" approach) can be combined to improve the overall performance. Third, the mechanism of training data or active learning [12, 13] may be used to adjust the weighting among different features. Finally, partial matching from several objects takes a long time to compute in general, and is also an important and difficult research direction in the future work [28, 29].

Acknowledgements

We would like to thank Miss Wan-Chi Luo, a graduate student, to help us manually classify the 1833 3DCafe objects into 47 classes plus another class "miscellaneous". We also appreciate Jeng-Sheng Yeh to set up the web server for our 3D model retrieval system. This project is partially funded by the National Science Council (NSC, Taiwan) under the grant number NSC91-2622-E-002-040 and by the Ministry of Education under the grant number 89-E-FA06-2-4-8.

References

1. T. Funkhouser, P. Min, M. Kazhdan, J. Chen, A. Halderman, D. Dobkin and D. Jacobs, "A Search Engine for 3D Models", *ACM Transactions on Graphics*, **22**(1):83-105, Jan. 2003.

2. R. Osada, T. Funkhouser, B. Chazelle and D. Dobkin, "Shape Distributions", *ACM Transactions on Graphics*, **21**(4):807-832, Oct. 2002.

3. M. Hilaga, Y. Shinagawa, T. Kohmura and T. L. Kunii, "Topology Matching for Fully Automatic Similarity Estimation of 3D Shapes", *Proc. of ACM SIGGRAPH*, 203-212, Los Angeles, USA, Aug. 2001.

4. E. Paquet, M. Rioux, A. Murching, T. Naveen and A. Tabatabai, "Description of Shape Information for 2-D and 3-D Objects", *Signal Processing: Image Communication*, **16**:103-122, Sept. 2000.

5. R. Ohbuchi, T. Otagiri, M. Ibato and T. Takei, "Shape-Similarity Search of Three-Dimensional Models Using Parameterized Statistics", *Proc. of 10th Pacific Graphics*, 265-273, Beijing, China, Oct. 2002.

6. D. V. Vranic and D. Saupe, "Description of 3D-Shape using a Complex Function on the Sphere", *Proc. of IEEE International Conference on Multimedia and Expo (ICME)*, 177-180, Lausanne, Switzerland, Aug. 2002.

7. I. Kolonias, D. Tzovaras, S. Malassiotis and M. G. Strintzis, "Fast Content-Based Search of VRML Models based on Shape Descriptions", *Proc. of Interna-*

tional Conference on Image Processing (ICIP), 133-136, Thessaloniki, Greece, Oct. 2001.

8. B. K.P. Horn, "Extended Gaussian Images", *Proceedings of the IEEE*, **72**(12):1671-1686, 1984.
9. M. Ankerst, G. Kastenmuller, H.-P. Kriegel and T. Seidl, "3D Shape Histograms for Similarity Search and Classification in Spatial Databases", *Proc. of 6th International Symposium on Advances in Spatial Databases (SSD)*, Hong Kong, China, 207-228, 1999.
10. D.-Y. Chen and M. Ouhyoung, "A 3D Object Retrieval System Based on Multi-Resolution Reeb Graph", *Proc. of Computer Graphics Workshop*, 16, Tainan, Taiwan, June 2002.
11. S. Jeannin, L. Cieplinski, J. R. Ohm and M. Kim, *MPEG-7 Visual part of eXperimentation Model Version 7.0, ISO/IEC JTC1/SC29/WG11/N3521*, Beijing, China, July 2000.
12. M. Elad, A. Tal, and S. Ar. "Content Based Retrieval of VRML Objects - An Iterative and Interactive Approach", *Proc. of 6th Eurographics Workshop on Multimedia*, 97-108, Manchester UK, Sept. 2001.
13. C. Zhang and T. Chen, "An Active Learning Framework for Content-Based Information Retrieval", *IEEE Transactions on Multimedia Special Issue on Multimedia Database*, **4**(2):260-268, June 2002.
14. P. Lindstrom and G. Turk, "Image-Driven Simplification ", *ACM Transactions on Graphics*, **19**(3):204-241, July 2000.
15. G. Turk, "Generating Textures on Arbitrary Surfaces Using Reaction-Diffusion", *Computer Graphics (Proc. of ACM SIGGRAPH)*, **25**(4):289-298, July 1991.
16. S. A. Dudani, K. J. Breeding and R. B. McGhee, "Aircraft Identification by Moment Invariants", *IEEE Transactions on Computers*, **C-26**(1):39-46, Jan. 1977.
17. T. P. Wallace and P. A. Wintz, "An Efficient Three-Dimensional Aircraft Recognition Algorithm Using Normalized Fourier Descriptors", *Computer Graphics and Image Processing*, **13**:99-126, 1980.
18. C. M. Cyr, B. B. Kimia, "3D Object Recognition Using Shape Similarity-Based Aspect Graph", *Proc. of International Conference on Computer Vision (ICCV)*, 254-261, Vancouver, Canada, July 2001.
19. C. E. Jacobs, A. Finkelstein, D. H. Salesin, "Fast Multiresolution Image Querying", *Proc. of ACM SIGGRAPH*, 277-286, Los Angeles, USA, Aug. 1995.
20. D. S. Zhang and G. Lu. "A comparative Study of Fourier Descriptors for Shape Representation and Retrieval". *Proc. of 5th Asian Conference on Computer Vision (ACCV)*, 652-657, Melbourne, Australia, Jan. 2002.
21. D. S. Zhang and G. Lu. "An Integrated Approach to Shape Based Image Retrieval". *Proc. of 5th Asian Conference on Computer Vision (ACCV)*, 652-657, Melbourne, Australia, Jan. 2002.
22. D. Heesch and S. Ruger, "Combining Features for Content-Based Sketch Retrieval - A Comparative Evaluation of Retrieval Performance", *Proc. of 24th BCS-IRSG European Colloquium on IR Research Glasgow (LNCS 2291)*, UK, Mar. 2002
23. A. Gershun, "The Light Field", Moscow, 1936. Translated by P. Moon and G. Timoshenko in *Journal of Mathematics and Physics*, **18**:51-151, MIT, 1939.
24. L. McMillan and G. Bishop, "Plenoptic Modeling: An Image-Based Rendering System", *Proc. of ACM SIGGRAPH*, 39-46, Los Angeles, USA, Aug. 1995.
25. M. Levoy and P. Hanrahan, "Light Field Rendering", *Proc. of ACM SIGGRAPH*, 31-42, New Orleans, USA, Aug. 1996.
26. T. Igarashi, S. Matsuoka and H. Tanaka, "Teddy: A Sketching Interface for 3D Freeform Design", *Proc. of ACM SIGGRAPH*, 409-416, Los Angeles, USA, Aug. 1999.
27. F. Huber and M. Hebert, "Fully Automatic Registration of Multiple 3D Data Sets", *Proc. of IEEE Workshop on Computer Vision Beyond the Visible Spectrum: Methods and Applications (CVBVS)*, Kauai, Hawaii, USA, Dec. 2001.
28. A. E. Johnson and M. Hebert, "Using Spin Images for Efficient Object Recognition in Cluttered 3D Scenes", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **21**(5):433-449, May 1999.
29. S. M. Yamany and A. A. Farag, "Surfacing Signatures: An Orientation Independent Free-Form Surface Representation Scheme for the Purpose of Objects Registration and Matching", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **24**(8):1105-1120, Aug. 2002.
30. A. E. Bimbo, *Visual Information Retrieval*, Morgan Kaufmann Publishers, Inc., 1999.
31. T. Pavlidis, *Algorithms for Graphics and Image Processing*, Computer Science Press, 1982.
32. R. M. Haralick and L. G. Shapiro, *Computer and Robot Vision*, Addison-Wesley Pub. Co., 1992.
33. I. T. Jolliffe, *Principal Component Analysis*, 2nd edition, Springer, 2002.
34. 3DCAFE, http://www.3dcafe.com
35. SpharmonicKit 2.5: Fast spherical transforms. http://www.cs.dartmouth.edu/∼geelong/sphere/

A 3D Model Alignment and Retrieval System

Ding-Yun Chen and Ming Ouhyoung
Department of Computer Science and Information Engineering,
National Taiwan University, Taipei, Taiwan
dynamic@cmlab.csie.ntu.edu.tw, ming@csie.ntu.edu.tw

Abstract

Techniques for 3D model alignment and retrieval are proposed in this paper. Since the techniques of 3D modeling and digitizing tools are in great demand, the expectations of 3D models alignment and retrieval are increasingly. We propose an algorithm for 3D model alignment, which gets the affine transformation between two 3D models. The main idea of our 3D alignment algorithm in rotation is to search the similarity of projected 2D shapes from each viewing angle between two models. Then, we apply the technique to match two 3D models after recovering the affine transformation. Our database has more than 10,000 3D models, and a query interface of drawing is easy to use to retrieval 3D models from 2D shapes.

1. Introduction

The problem of 3D objects recognition, retrieval, clustering and classification is a traditional research topic during previous decades in computer vision, computer graphic, mechanical engineering, medical imaging and molecular biology. The research topic is more and more important since the techniques of 3D modeling and digitizing tools are in great demand. Many tools of digitized and constructed 3D objects are getting more and more popular, for example, 3D model acquisition systems [19], 3D model capturing systems [22], 3D freeform design systems [17] and sculpting systems [18]. Therefore, 3D objects can be digitized and modeled easier, faster and less expensive. A large number of free 3D models can be accessed all over the world via the Internet, such as in [15]. Although text-based search engines are ubiquitous today, multimedia data such as 3D models lack meaningful and semantic description for automatic matching. The MPEG group aims to create an MPEG-7 international standard, also known as "Multimedia Content Description Interface", for the description of the multimedia data, including image, video, audio, 2D shapes and 3D objects [16]. However, there is currently only one descriptor for 3D model. This has highlighted the need for developing efficient techniques of content-based retrieval for 3D model.

The problem of 3D model retrieval can be stated as follows: given a 3D model, the retrieval system compares it with all other 3D models from the database, and shows ranked similar models. In short, the problem is to determine the similarity between two given 3D models. The most important issue is to extract suitable features for matching. The feature should represent the characteristics of different 3D models, and should be invariant to affine transformation (translation, rotation and scaling), and robust against most 3D model processing, such as re-meshing, subdivision, simplification, noise and deformation. The second important issue is to define a meaningful distance metric, which should be efficient.

Most previous works of 3D model retrieval focus on finding a good feature for matching [1~14]. Most of those are based on either statistical properties, such as global shape histograms, or the skeletal structure of 3D model. Zhang and Chen [2] proposed an algorithm to efficiently calculate features, such as volume, moments and Fourier transformation coefficients. They also applied active learning algorithm in 3D model retrieval using volume-surface ratio, aspect ratio, moment invariants and Fourier transformation coefficients [3, 4]. The features are normalized to be within range (-1, 1).

Osada et al. [6, 7] propose and analyze a method for computing shape signatures for arbitrary 3D polygonal models. The key idea is to represent the signature of an object as a shape distribution sampled from a shape function measuring global geometric properties of an object. The best one of their experimental results

is using the D2 shape function, which measures the distance between two random points on a surface. Then, the entire shape distribution is scaled based on the mean in order to deal with the scaling problem. Finally, they examine that the PDF L1 norm performed the best for comparing shape distributions. In their experimental results, the top matches are mostly from the same class as the query object and that shape distributions provide a useful signature for shape-based retrieval of 3D models. They also compare D2 shape distribution method against surface moments, and find the D2 shape distributions outperform moments for classification of 3D models. Their approach is simple, fast, and is invariant under affine transformation, and is not sensitive to most 3D model processing and degeneracies.

Thomas Funkhouser et al. [8] describe a web-based search engine system that supports queries based on 3D sketches, 2D sketches, 3D models, and/or text keywords. For the shape-based queries, they have developed a new matching algorithm that uses spherical harmonics to compute discriminating similarity measures without requiring repair of model degeneracies or alignment of orientations. Their matching methods are fast enough to match a single query to a repository of 20,000 models in under a second.

Hilaga et al. [1] propose a technique in which similarity between polyhedral models is accurately and automatically calculated by comparing the skeletal and topological structure. Therefore, their algorithm can handle the global and local properties simultaneously. The skeletal and topological structure decomposes to a one-dimensional graph structure. The graph is invariant to affine transformation and robust against most 3D model processing and deformation caused by changing posture of an articulated object. Their search key is a multi-resolution structure of the graph, so that the comparison can simply and fast. Their experiments made use of 230 different polyhedral meshes. In their experimental results, the average search time is about 12 seconds in a PC with a Pentium II 400MHz processor. A 3D model retrieval system [13] in our previous work is extended the work of Hilaga et al.

In general, features of 3D models should be invariant to affine transformations, since each 3D model has its own coordinate axis for different use. In contrast, we propose an algorithm to recover translation, scaling and rotation between two 3D models, and then extend the technique to measure the similarity. Furthermore, the function of 3D model alignment can not only be used in 3D model retrieval, but also in many other applications, such as mesh watermarking, 3D model morphing, 3D animation, and so on.

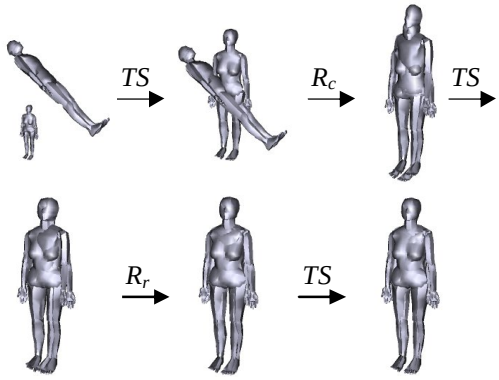


Figure 1 The order of 3D models alignment.

The basic concept of our approach is that when two 3D models are similar, they will also look similar from their corresponding viewing angles. Therefore, the main idea of our 3D alignment algorithm in rotation is to render 2D silhouettes from each viewing angle of two models, and get rotation which has minimum error summing from all viewing angles using 2D shape matching algorithm. Our approach of 3D model retrieval takes the minimum error as the similarity between two 3D models. The remaindered part of this paper is organized as follows. In Chapter 2, we propose an algorithm to do 3D model alignment. We detail rotation alignment in Chapter 3. The experimental results of 3D model alignment are represented in Chapter 4. 3D model retrieval is proposed in Chapter 5. Finally, the paper is summarized and concluded in Chapter 6.

2. Flow of 3D Model Alignment

The order of 3D model alignment in our approach is as follows: $TS \rightarrow R_c \rightarrow TS \rightarrow R_r \rightarrow TS$. Where *TS* denotes the translation and scaling alignment from one 3D model to another; R_c and R_r denote the coarser and refined rotation alignment, respectively (detailed in next chapter). All *TS* apply the same operator, that is, translate to the same origin and scale to the same size between two models. The purpose of first two *TS* is to let the two models be roughly in similar position and close to the same size, which will make it easier to get the correct rotation R_c and R_r. Once the correct rotation is recovered, the last *TS* will be easier to get the correct translation and scaling. Figure 1 shows an example of the five steps.

The approach of translation and scaling is very simple. The translation $T=(T_x,T_y,T_z)$ assigns the middle point of the whole model to be the

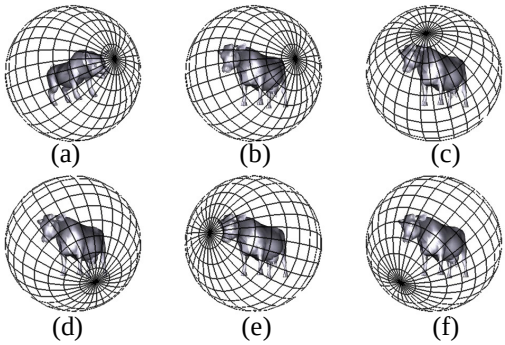


Figure 2 A typical example to show our algorithm.

new origin. The scaling is isotropic, and normalizes according to the maximum distance from x, y and z axes of the whole model. That is,

$$T_i = \frac{MaxCoor_i + MinCoor_i}{2}, i = x, y, z, \quad (1)$$

$$S = \frac{1}{\max_{i=x,y,z}(MaxCoor_i - MinCoor_i)}, \quad (2)$$

where the $MaxCoor_i$ and $MinCoor_i$ are the coordinate of vertex, which has maximum and minimum value in *i*-axis, respectively.

An intuitional thought of recovering the rotation from two models is to rotate model to all possible viewing angles, and get the rotation that has minimum error from all viewing angles. We take Figure 2 as a typical example to explain the idea. There are two models, *pig* and *cow*, with different rotations, and both models have been applied coarser translating and scaling (*TS*) alignment. To recover the rotation from model *cow* to model *pig*, a set of cameras surrounding a model to render 2D shapes from each viewing angle. Those cameras are put on the surface of a sphere and scatter viewing angles all over the sphere. Figure 2 (a) shows a set of cameras surrounding the model *pig*, where each intersection point indicates a camera position. Then, apply the camera set to the model *cow*, as shown in Figure 2 (b), and then get the difference between two 2D shapes for each camera pair (A camera pair is the corresponding viewing angle between the two models). Therefore, we define the error between two 3D models, when applying some rotated camera set, as summing the difference from all camera pairs. The rotation of a camera set, which has minimum error, also indicates the rotation between the two 3D models. The minimum error can also be used for judging the similarity of the two 3D models, and is defined as:

$$\min_i \sum_j ShapeDiff(j), \quad (3)$$

where *ShapeDiff* denotes the difference between two 2D shapes; *i* denotes different rotation of the camera set, and *j*-th camera pair between the two models. Figure 2 (b)~(f) show several examples of different rotation of the camera set, and we can suppose that the camera set in Figure 2 (e) will get the minimum error.

When rotating the camera set to a new orientation, all 2D shapes should be re-rendered, and the similarity matching between each camera pair also has to be calculated again. This will cause large amount of calculation, and is time consuming. Therefore, we put the camera set in the vertices of a regular convex polyhedron, so that the number of rendering and matching will be greatly reduced.

However, there are only five regular convex polyhedrons, which are named as Platonic bodies, and were known to the ancient Greeks. The fact can also be proved using Euler's theorem. The five regular convex polyhedrons are tetrahedron, hexahedron or cube, octahedron, dodecahedron, and icosahedron. We take vertices of dodecahedron, which has the maximum amount of vertices, as the position of the camera set. There are 20 scattering viewing angles for each 3D model. Huber and Hubert [15] proposed an automatic registration algorithm, which is able to reconstruct real world objects from 15 to 20 various viewpoints of laser scanner. Therefore, the 20 viewing angles can roughly represent the shape of a 3D model. The 20 2D shapes are the bases of each 3D model for alignment, and contain enough information from various viewing angles for a 3D model. Figure 3 show that we can reduce the number of rendering and matching by using the dodecahedron. Figure 3 Figure 3 (a) shows a camera set for model *pig*, and then the same camera set is applied to model *cow*, as shown in Figure 3 (b). That is, the indices of camera set are all the same between Figure 3 (a) and (b). In addition, we can rotate the dodecahedron resulting in the camera set is at the same position. For instance, rotate edge (1,2) from Figure 3 (b) to edge (1,3) and (1,4), which show in Figure 3 (c) and (d), respectively. Since a dodecahedron has 20 vertices and each vertex connects 3 edges, there are 60 kinds of different rotation, which share the same 20 camera positions. Table 1 shows the number of rendering and 2D shapes

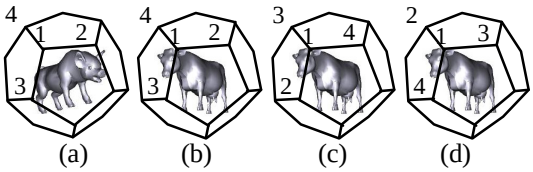


Figure 3 We can reduce the number of calculation from rendering and 2D shapes matching by using the dodecahedron.

Table 1 Number of rendering and 2D shapes matching for 60 different rotations with and without using dodecahedron.

For testing 60 kinds of different rotation	Number of rendering	Number of 2D shapes matching
Without using dodecahedron	20 + 20 × 60 = 1220	20 × 60 = 1200
Using dodecahedron	20 + 20 × 1 = 40	20 × 20 = 400
Ratio	30.5	3

Table 2 Number of rendering and matching 2D shapes are calculated by mapping m to n different dodecahedrons.

(m-n)	1-1	1-10	1-20	1-40	10-10
Number of rendering ($20\times m+20\times n$)	40	220	420	820	400
Number of 2D shape matching($20\times m\times 20\times n$)	400	4000	8000	16000	40000
Number of different rotations ($60\times m\times n$)	60	600	1200	2400	6000
Vertices of scattering dodecahedrons					

matching for 60 different rotations with and without using dodecahedron. Consequently, amount of rendering and matching can be reduced.

There are 60 different rotations for searching by using one camera set of both models. However, it's usually not enough to recover rotation from the best one of the 60 candidates for the coarser rotation alignment, R_c. The coarser rotation alignment should provide a good initial, so that the refined rotation alignment, R_r, can easily get the best result from the local estimation. Therefore, we can use more camera sets for each 3D model. When adding one camera set, 60 different rotations will be searched. Furthermore, we can also apply more camera sets to both 3D models. That is, when applying m and n camera set to both models, respectively, there will be $60\times m\times n$ kinds of different rotation for searching. Table 2 shows the number of rendering and 2D shape matching by using different number of camera sets. In our implementation, we set m=10 and n=10 (10-10), that is, we take the best one from 6000 different rotations as the coarser rotation alignment, R_c. The algorithm of rotation alignment, R_c and R_r, will detail in next chapter. In addition, those camera sets are pre-processed for scattering over the whole rotation space, so that any possible rotation can close to one of them.

In the end of this chapter, we detail the flow of aligning one model to another. The operations of aligning model B to model A is shown as follows:

$A' = A \cdot T_A \cdot S_A$

$B' = B \cdot T_B \cdot S_B$

$R_c = RotateCoarse(B', A')$

$B'' = B' \cdot R_c \cdot T_B \cdot S_B$

$R_r = RotateRefine(B'', A')$

$A' \sim B'' \cdot R_r \cdot T_B \cdot S_B$

$A \cdot T_A \cdot S_A \sim B' \cdot R_c \cdot T_B \cdot S_B \cdot R_r \cdot T_B \cdot S_B$

$A \sim B \cdot T_B \cdot S_B \cdot R_c \cdot T_B \cdot S_B \cdot R_r \cdot T_B \cdot S_B \cdot S_A^{-1} \cdot T_A^{-1}$

where *RotateCoarse* and *RotateRefine* recover the coarser and refined rotation, respectively, and detail in next chapter.

3. 3D model Alignment in Rotation

This chapter details our approach of aligning rotation between two models. The rotation alignment has two levels: coarser (R_c) and refined (R_r) alignment. The coarser alignment gets the approximate rotation from all possible orientation between two models. The refined alignment is to adjust the rotation from the above result to more accurate one. After doing first *TS* (described in previous chapter), the position and size of both models are approximate, but not exactly the same, so our approach of 2D shape matching should be invariant to translation and scaling.

The rotation between two models is aligned by matching 2D shapes from camera sets of both models. Figure 4 shows the flow of rotation alignment. First, 2D shapes should be rendered from camera sets for both models. For each 2D shape, its feature can be extracted and saved for matching later. The operation of rendering and feature extraction do 40 times, if using 1-1 camera set. Then, 2D shapes of each camera pair have to be matched, and the operation does 400 times if using 1-1 camera set. Next, get minimum error from different rotations, as defined in Eqn. (3). Finally, once two camera sets of both models are determined, the rotation matrix of two models can be obtained by the rotation of the two camera sets. The main flows of coarser and refined rotation alignment are the

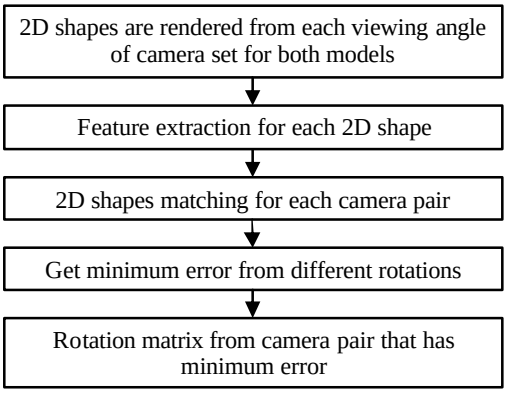


Figure 4 The flow of rotation alignment

Figure 5 A typical example of 2D silhouettes from a camera set of dodecahedron.

same.

The character of 2D shape matching depends on which algorithm is used. We use OpenGL to render 2D silhouettes by putting camera to vertex of dodecahedron and facing to origin. The size of 2D silhouettes is 256 by 256 pixels in our implementation. Since 3D models should be translated and scaled, *TS*, before rotation alignment, it's easier to make sure that whole 3D models will be rendered into a 2D silhouette, that is, no clipping happen. Figure 5 shows a typical example of 2D silhouettes from a camera set of dodecahedron. In our implementation, we render to screen by perspective projection, and then use *glReadPixels*() to get 2D silhouettes.

To measure the similarity between two shapes, we use region-based shape descriptor of the MPEG-7 [16] to match. The matching algorithm can be invariant to translation, scaling and rotation in 2D shapes, and allowable of minor non-rigid deformations. The region-based shape descriptor makes use of all pixels constituting the shape, so that it can describe complex shape including holes and several disjoint regions. The descriptor utilizes a set of ART (Angular Radial Transform) coefficients to describe the shape. The ART is a 2D complex transform defined on a unit disk in polar coordinates. Twelve angular and three radial functions are used, and 35 ART coefficients of 2D shapes are used for matching.

There are several notices for using the 2D shape matching to our approach. In general, in order to invariant to translation in pure 2D case, the center of mass in 2D shape should be aligned to the center of the unit disk. However, our final alignment is in 3D case, so it's no reason to align the center for each 2D shape. Since translating 3D model to origin has applied before rotation alignment, we use center of rendered 2D shape as the center of the unit disk. Furthermore, in order to invariant to scaling, linear interpolation is applied to align between rendered 2D shapes from each viewing aspect and the unit disk. The same as translation, each 2D shapes in a camera set should have the same scaling.

After feature extraction for each 2D shape, shape matching for each shape from two models is calculated. Amount of feature extraction is the same as rendering, but the amount of shape matching is much more. In general, the computation of matching is much less than that of feature extraction in order to speedy retrieval from a large database, since the feature can be previously calculated and saved to database. The region-based shape matching algorithm use simple L1 distance to measure similarity:

$$ShapeDiff((A,B)) = \sum_i |ArtM_A[i] - ArtM_B[i]| \quad (4)$$

where *ArtM* is the ART coefficients, *ShapeDiff* is the same in the Eqn. (3); *A* and *B* are two 2D shapes for matching; *i* is index of ART coefficients. Therefore, the 2D shape matching is speedy.

Next, get minimum error from different rotations of dodecahedron, as defined in Eqn. (3). The error between different rotations is defined as summing differences from all corresponding 2D shapes pair. However, there is a little difference in this stage between coarser and refined rotation alignment. In coarser rotation alignment, we use 10-10 camera sets, that is, searching the best one from 6000 different rotations. In refined rotation alignment, we use iterative approach to close the best solution. We start from 10° and step half for each iterative until less than 1°. In each iterative, we adjust rotation of one axis and fix that of another two axes in the order of X, Y and Z axis, respectively. When adjusting rotation of one axis, we rotate the camera set to the direction that has less error, until no improvement. Therefore, we can align the rotation with error less than 1°.

Finally, rotation matrix between dodecahedron pair that has minimum error, should be calculated. The rotation matrix (R_c or R_r) is then applied to one model to align rotation to another. The problem of solving the rotation matrix can be considered as aligning an edge between two dodecahedrons, because all edges are aligned if one edge is aligned. We utilize the function of coordinate conversion between the Cartesian and an arbitrary coordinate system to obtain the rotation matrix. We use Figure 6 to explain our approach. Suppose that Figure 6 (a) and (c) are the dodecahedron pair of model A and B, respectively. The rotation matrix aligns edge (1,2) in Figure 6 (c) to that in (a). The vector $\overrightarrow{o1}$ and $\overrightarrow{o2}$ can form a unique coordinate frame, defined as follows:

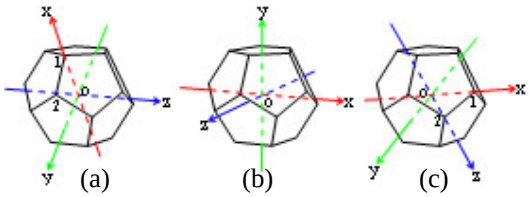


Figure 6 Rotation matrix is calculated between two camera sets.

$$\begin{cases} \vec{x} = \overrightarrow{o1} \\ \vec{y} = \vec{z} \times \vec{x} \\ \vec{z} = \overrightarrow{o1} \times \overrightarrow{o2} \end{cases} \quad (5)$$

where "×" denotes cross produce. The notation F_A and F_B denote the coordinate system of model A and B, respectively, and F_C denote the Cartesian coordinate system. Therefore, the rotation matrix is the coordinate conversion from F_B to F_A, that is, F_{BA}. However, 3D models are in Cartesian coordinate system, F_C, so we cannot apply F_{BA} to model B directly. Model B should be converted to F_B coordinate system first and back to Cartesian coordinate system after applying F_{BA}. The rotation matrix is defined as:

$$F_{CB} \cdot F_{BA} \cdot F_{BC}$$
$$= F_{CB} \cdot F_{AC} \cdot F_{CB} \cdot F_{BC}$$
$$= F_{CB} \cdot F_{AC}$$

The F_{BA} can be obtained by $F_{AC} \cdot F_{CB}$, and $F_{CB} \cdot F_{BC}$ can be eliminated, so the rotation matrix from model B to model A is $F_{CB} \cdot F_{AC}$.

4. Experimental Results of 3D Alignment

In order to experiment with the 3D alignment algorithm, we use 445 models, downloaded from [20] and [21], for initial testing. The alignment algorithm should work well at least using the same models. So those models are randomly rotated, translated and scaled by another program, and then using our 3D alignment algorithm to test. Figure 7 shows the results, and almost all of them work well. Each model has five pictures. Picture 1 is the original model, and picture 2 is the destination model, which randomly translate, scale and rotate from original model. To clearly look the relation of the two models, picture 3 put original and destination model together. So we can see the difference of translation, rotation and scaling between two models. Picture 4 and 5 are the results of coarser and refined alignment between two models, respectively.

In the 445 models, there are 5274.4 vertices and 10233.8 triangles in average. The average execution time for coarser and refined alignments are 14.5 and 23.5 seconds, respectively, in a PC with Pentium III 800MHz CPU.

In addition, we also test our algorithm by using different models. Those models are also randomly rotated, translated and scaled by another program first, and then using our 3D alignment algorithm to test. Figure 8 shows the experiment results. All experiment results are available in the web pages: http://www.cmlab.csie.ntu.edu.tw/~dynamic/3DRetrieval/index.html.

5. 3D Model Retrieval

We apply the technique of 3D model alignment to perform 3D model retrieval. In order to reduce the retrieval time, we move 3D model alignment up to coarser rotation alignment stage. That is, the search key of 3D models is the ART coefficients from 2D shapes of each camera set. We take the minimum error between two models as the similar measurement (defined in Eqn. (3)). All models are randomly rotated, translated and scaled by another program first, and then using our 3D model retrieval to test. The retrieval time is about 4 seconds in a PC with Pentium III 800MHz CPU. Figure 9 shows several experimental results of 3D model retrieval. In addition, we also test a database which has more than 10,000 3D models, and the results are shown in Figure 10 and 11. Furthermore, a query interface can be used to retrieval 3D models from 2D shapes, as shown in Figure 12. The retrieval from user drawing is very fast, about 1~2 seconds for searching from more than 10,000 3D models. The demo can also be found in http://www.cmlab.csie.ntu.edu.tw/~dynamic/3DRetrieval/index.html.

6. Conclusion and Future Works

The paper presents an algorithm of 3D model alignment based on a set of 2D shapes, which are projected from a 3D position, and then applies the technique to 3D model retrieval. The goal of 3D model retrieval is to recover coarser affine transformations first, and is robust against affine transformation and most 3D model processing. In the future, other 2D shape matching algorithms can be applied to improve the 3D model alignment algorithm. In addition, other attributes, such as color and texture, can be introduced for 3D model retrieval.

Acknowledgement

We would like to thank Jeng-Sheng Yeh for providing the web server in our experiments. This project is partially funded by the National Science Council (NSC, Taiwan) under the grant number NSC90-2622-E-002-008.

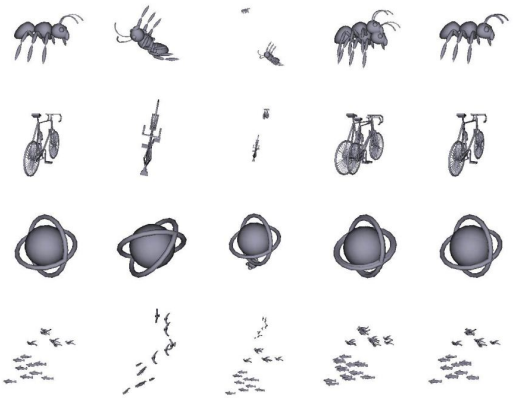


Figure 7 Results of experiments by using the same model among different affine transformations.

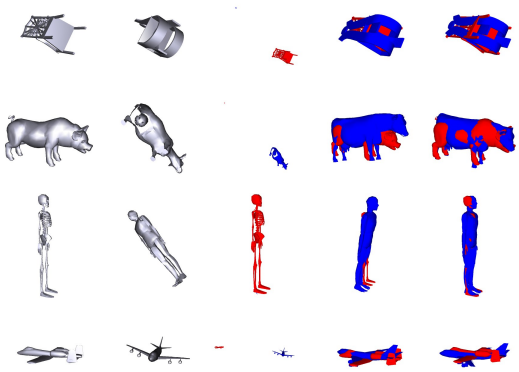


Figure 8 Results of experiments in aligning model before and after randomly rotating, translating and scaling.

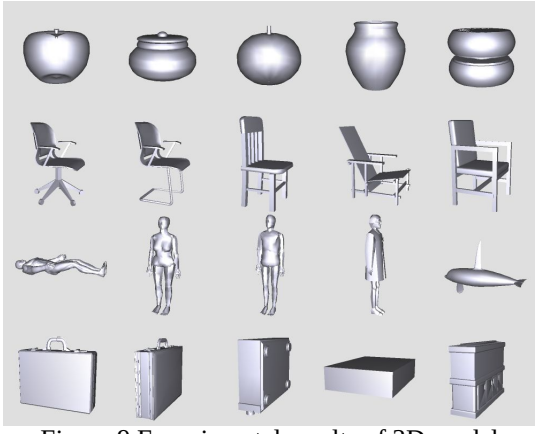


Figure 9 Experimental results of 3D model retrieval. The first one of each row is the target to be queried. The top 5 similar models are ranked from left to right

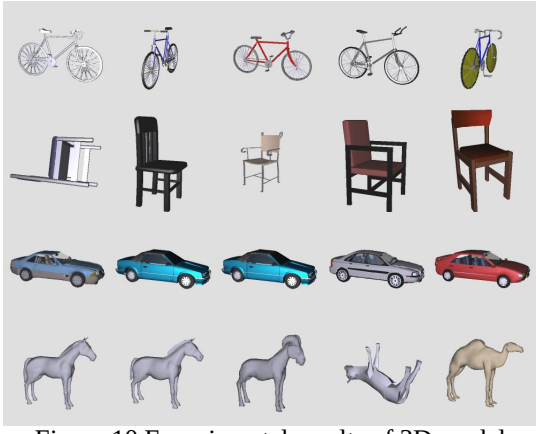


Figure 10 Experimental results of 3D model retrieval using a database which has more than 10,000 3D models. The first one of each row is the target to be queried. The top 5 similar models are ranked from left to right

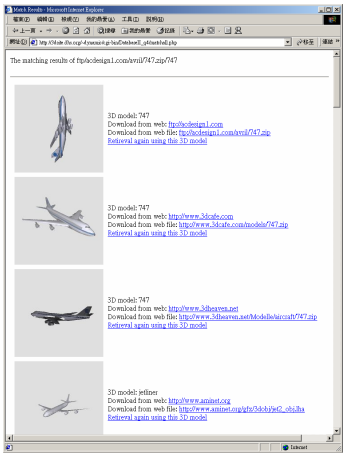


Figure 11 A retrieval result using a database which has more than 10,000 3D models. Top one is the input 3D models, and the similar models are ranked from top to down.

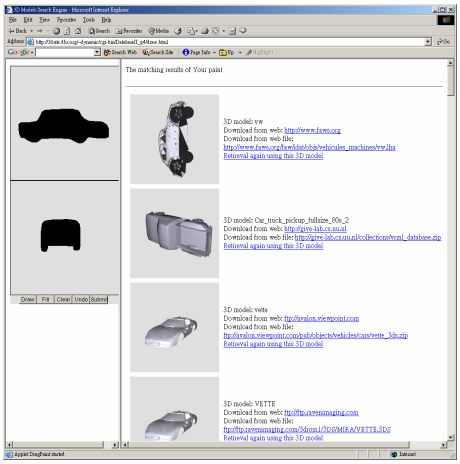


Figure 12 A query interface of user drawing can be used to retrieval 3D models from 2D shapes.

References

[1] Masaki Hilaga, Yoshihisa Shinagawa, Taku Kohmura and Tosiyasu L. Kunii, "Topology Matching for Fully Automatic Similarity Estimation of 3D Shapes", *Proceedings of ACM SIGGRAPH*, pp. 203-212, Los Angeles, USA, Aug. 2001.

[2] Cha Zhang and Tsuhan Chen, "Efficient Feature Extraction for 2D/3D Objects in Mesh Representation", *Proceedings of IEEE International Conference on Image Processing (ICIP)*, Thessaloniki, pp. 935-938, Greece, Oct. 2001.

[3] Cha Zhang and Tsuhan Chen, "Indexing and retrieval of 3D models aided by active learning", *Proceedings of ACM International Conference on Multimedia*, pp. 615-616, Ottawa, Canada, Oct. 2001.

[4] Cha Zhang and Tsuhan Chen, "An Active Learning Framework for Content-Based Information Retrieval", *IEEE Transactions on Multimedia*, Vol. 4, No. 2, June 2002.

[5] Ilias Kolonias, Dimitrios Tzovaras, Stratos Malassiotis and Michael G. Strintzis, "Fast Content-Based Search of VRML Models Based on Shape Descriptors", *Proceedings of IEEE International Conference on Image Processing (ICIP)*, pp. 133-136, Thessaloniki, Thessaloniki, Greece, Oct. 2001.

[6] Robert Osada, Thomas Funkhouser, Bernard Chazelle and David Dobkin "Matching 3D Models with Shape Distributions", *Shape Modeling International*, pp. 154-166, Genova, Italy, May 2001.

[7] Robert Osada, Thomas Funkhouser, Bernard Chazelle and David Dobkin "Shape Distributions", *ACM Transactions on Graphics*, Vol. 21, No. 4, pp. 807-832, Oct. 2002.

[8] Thomas Funkhouser, Patrick Min, Michael Kazhdan, Joyce Chen, Alex Halderman, David Dobkin, and David Jacobs, "A Search Engine for 3D Models" to appear in *ACM Transactions on Graphics*, Jan. 2003.

[9] Christopher M. Cyr and Benjamin B. Kimia, "3D Object Recognition Using Shape Similiarity-Based Aspect Graph", *Proceedings of International Conference on Computer Vision (ICCV)*, pp. 254-261, 2001.

[10] Michael Elad, Ayellet Tal and Sigal Ar, "Content Based Retrieval of VRML Objects – A Iterative and Interactive Approach", *Proceedings of 6th Eurographics Workshop on Multimedia*, Manchester UK, Sept. 2001

[11] Eric Paquet and Marc Rioux, "Content-Based Access of VRML Libraries", *Lecture Notes in Computer Sciences*, Vol. 1464, pp. 20-32, 1998.

[12] Eric Paquet, Marc Rioux, Anil Murching, Thumpudi Naveen and Ali Tabatabai. "Description of shape information for 2-D and 3-D objects", *Signal Processing: Image Communication*, Vol. 16, pp. 103-122, Sept. 2000.

[13] Ding-Yun Chen and Ming Ouhyoung, "A 3D Object Retrieval System Based on Multi-Resolution Reeb Graph", *Proceedings of Computer Graphics Workshop*, pp. 16, Tainan, Taiwan, June 2002.

[14] Ryutarou Ohbuchi, Tomo Otagiri, Masatoshi Ibato and Tsuyoshi Takei, "Shape-Similarity Search of Three-Dimensional Models Using Parameterized Statistics", *Proceedings of 10th Pacific Graphics*, pp. 265-273, Beijing China, Oct. 2002.

[15] Daniel F. Huber and Martial Hebert, "Fully Automatic Registration of Multiple 3D Data Sets", *Proceedings of IEEE Workshop on Computer Vision Beyond the Visible Spectrum: Methods and Applications* (*CVBVS*), Kauai, Hawaii, USA, Dec. 2001.

[16] Sylvie Jeannin, Leszek Cieplinski, Jens Rainer Ohm and Munchurl Kim, *MPEG-7 Visual part of eXperimentation Model Version 7.0, ISO/IEC JTC1/SC29/WG11/N3521*, Beijing, China, July 2000.

[17] Takeo Igarashi, Satoshi Matsuoka and Hidehiko Tanaka, "Teddy: A Sketching Interface for 3D Freeform Design", *proceedings of ACM SIGGRAPH*, pp. 409-416, Los Angeles, USA, Aug. 1999.

[18] Guo-Luen Perng, Wei-Teh Wang and Ming Ouhyoung, "A Real-time 3D Virtual Sculpting Tool Based on Marching Cubes", *Proceedings of International Conference on Artificial Reality and Tele-existence (ICAT)*, pp. 64-72, Tokyo, Japan, Dec. 2001.

[19] Szymon Rusinkiewicz, Olaf Hall-Holt and Marc Levoy, "Real-Time 3D Model Acquisition", *proceedings of ACM SIGGRAPH*, pp. 438-446, San Antonio, USA, July 2002.

[20] http://www.3dcafe.com

[21] http://deep.sitenest.net

[22] http://www.3dfamily.com/products/digibox/main.htm

3D Model Search Engine Based on Lightfield Descriptors

Yu-Te Shen, Ding-Yun Chen, Xiao-Pei Tian and Ming Ouhyoung

Department of Computer Science and Information Engineering,
National Taiwan University, Taipei, Taiwan
{edwards, dynamic, babylon}@cmlab.csie.ntu.edu.tw, ming@cmlab.csie.ntu.edu.tw

Abstract

*With the development of modern 3D modelling and digitizing tools, more and more models have been created recently, which leads to the necessity of the technique of 3D model retrieval system. Our 3D model search engine has been designed with the goal to meet entertaining, industrial, commercial, medical, and educational needs. The system is available on the Web (*http://3d.csie.ntu.edu.tw*) with a database containing over 10,000 models free downloaded through the Internet. Users can query 3D models by text, drawing 2D shapes using a friendly painting interface provided, or selecting one 3D model as a query key interactively. The features representing a 3D object, namely, the Lightfield Descriptors, are extracted from 2D images, which are rendered from cameras positioned on the vertices of a regular dodecahedron. The Lightfield Descriptors of each model are used to compare similarities among each other, and the retrieval process takes only 2 and 0.1 seconds with a 3D model and 2D shapes, respectively. During four months, the system has been widely used for querying over three thousand times from 418 IP addresses in at least 27 countries.*

Categories and Subject Descriptors (according to ACM CCS): I.3.3 [Computer Graphics]: Line and Curve Generation

1. Introduction

1.1. Motivation and applications

An information retrieval system is a system that is capable of storage, retrieval, and maintenance of information. Original definitions focused on "documents" for information retrieval rather than multimedia integrated information [26]. More and more multimedia information has become available from sources all over the world, and may be represented in various forms, including still pictures, graphics, 3D models, audio, speech, and video. Nevertheless, the value of information often depends on how easy it can be found, retrieved, accessed, filtered and managed [15]. As we can see, the transition between the second and the third millennium abounds with new ways to produce, offer, filter, search, and manage digitized multimedia information, the trend is getting clear: in the next few years, users will be confronted with a large number of contents provided by multiple sources that efficient and accurate access to these boundless contents seems unimaginable today [15]. Therefore, the need of multimedia information retrieval has increased.

Recently, the development of 3D modelling and digitizing technologies has made the model generating process much easier. The 3D modelling tools, including 3ds Max [29], MAYA [30], AutoCAD [31], 3D freeform design systems [14] and sculpting systems [19], [20], facilitate users to create 3D models directly on the computers. The 3D digitizing tools, on the other hand, digitize 3D objects from the real world, and include 3D scanner machines [32], registration from range data [16], [17], automatic modelling from multi-view video [18], etc. Since it is obvious that the 3D modelling and digitizing techniques will be developed and improved in the future, more and more 3D models, accordingly, will be created easier, faster and less expensive. The need of developing efficient techniques for content-based 3D model retrieval is also increasing.

The technique of 3D model retrieval can be applied to many practical applications, and is introduced with the following scenario. Suppose that a user wants to create a digital content, for example, a slide show for presentation or a 3D computer game. He/She needs a number of 3D models, but it is impractical to create each one from the scratch. Hence,

he/she connects to the Internet and visits a 3D model search engine. The search engine is based on content-based 3D model retrieval, and provides a friendly interface for users to draw 2D shapes for query. Soon, he/she downloads plenty of 3D models retrieved from the search engine, and slightly modifies them before use. Subsequently, he/she needs to design a 3D trademark for a project. In the existing Trademark Laws of many countries, 3D shapes are allowed to form elements of a trademark. Therefore, during designing a 3D trademark, he/she uses the technique of 3D model retrieval to ensure that no other similar shapes are registered. A 3D trademark plays an important symbol for a project, and may be conceived and designed for a long time. Therefore, watermark was proposed to be embedded in the 3D trademark to protect the intellectual property [21]. One day, he/she finds the same 3D trademark spread on the Internet for other use, and doubts that the trademark might be infringed by other companies. Therefore, he/she searches similar 3D models from thousands examples via the Internet using the 3D model retrieval for pre-filtering, and then applies watermark technique to find the infringed one.

At a later time, he/she wants to buy a chair with a specific shape using electronic commerce (E-commerce) via the Internet, and then visits a shopping center, which includes many shopping sites. When he/she queries by the keyword "chair", it's very difficult to find what he/she wants since too many models are retrieved from the text-based search engine. Fortunately, the shopping center also provides content-based 3D model retrieval system. With the help of the content-based retrieval system, he/she found the chair in a shopping site soon. Then, he/she visits an art sculpture museum on a web site, which contains many sculptures digitized by a 3D scanner, such as David by Michelangelo [17]. He/She wants to review a sculpture, but he/she forgets or never knows the name of that. Therefore, he/she finds the sculpture using the 3D model retrieval technique, and learns more about it. Since he/she is interested about the specific shape of the work, he/she uses the 3D model retrieval system to get more sculptures, which share similar 3D shapes, and investigates among them.

Main applications described above include 3D model search engines, verification of 3D trademarks, pre-filtering for 3D watermarks, and user interfaces for E-commerce and sculpture museum. Other possible applications of the 3D model retrieval are 3D object recognition, multimedia editing, education, digital libraries, functional labelling in 3D medical images, and molecular biology, etc. In this paper, one of the kernel applications, a 3D model search engine, is proposed in Section 5.

1.2. Objectives and challenges

The general objective of an information retrieval system is to minimize the overhead of a user locating needed information. Overhead can be expressed as the time a user spends in all of the steps leading to reading an item containing the needed information (e.g., query generation, query execution, and scanning results of query to select items to read, reading non-relevant items) [26]. To minimize the overhead, one of the most important issues is to increase the retrieval precision when designing a 3D model retrieval system. Therefore, the purpose of 3D model retrieval is to search relevant 3D models efficiently and correctly by querying from a database. One straightforward way is matching the input 3D model to each one in the database, and ranking by matching similarity. The key point is the way to define the similarity between two 3D models according to human perception. In general, representative features are extracted for each 3D model, and then matched among these models. Therefore, the problem of 3D model retrieval should be focussed on how to define features for representing 3D models and distance metrics for matching the features.

Several characteristics of matching 3D models make differences for other related problems. For example, recognizing objects from 2D images or range images is a traditional problem in computer vision [22], [23]. The difficulty of the problem lies in different projections, illuminations, shadows, reflections, segmentations, clutters, occlusions and partial matching. Most of these problems don't exist in 3D model retrieval. However, 3D model is in higher dimension than images, and usually representing irregularly sampled points. Another related problem is in molecular biology [24]. One of the problem focuses on partial matching in order to find the common active site among proteins. Note that there is no scaling problem in molecular biology, since binding distance between oxygen (O) and hydrogen (H) is a constant value.

In 3D model retrieval, there are many challenges when designing representative features, including extracting and matching among the features. The main challenges are listed in the following.

(1) Automation: Feature extraction should be automatic for dealing with large number of 3D models, which are collected from the Internet. In addition, when querying by uploading a 3D model, the features of the 3D model should also be automatically extracted.

(2) Efficiency: Both feature extraction and matching should be efficient, especially in feature matching. Therefore, the size of features should be small for quick comparison.

(3) Scope: Feature extraction and matching should work well in various kinds of 3D models.

(4) Robustness: The features should be robust against geometric processing, such as similarity transformation (translation, rotation and scaling), connectivity changes (remeshing, sub-division and simplification), model degeneracy (missing, wrongly oriented, intersecting, disjoint and

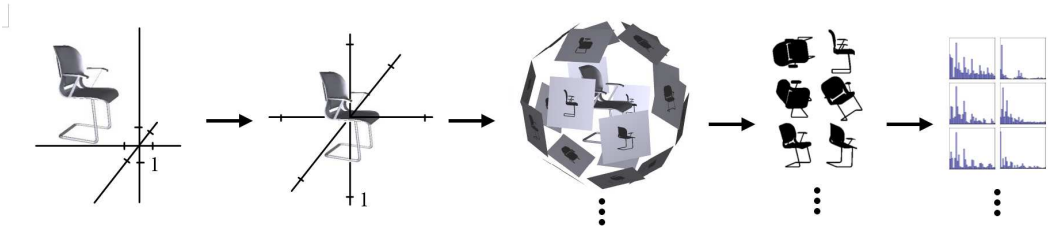


Figure 1: *Steps of extracting the LightField Descriptors for a 3D model*

overlapping polygons), random noise, smoothing, deformation, and posture changing, etc.

(5) Discrimination: The features should be sensitive to preserve important distinctions among 3D models. In addition, by ensuring that similar 3D models will have similar features, small changes in 3D models should lead to small changes in features.

2. Previous works

The technique of 3D model retrieval and matching is an important research topic in many research organizations from all over the world, such as Princeton [2], [3], CMU [12], Berkeley, Stanford, Brown and Texas University in USA, Tokyo [4] and Yamanashi [6] University in Japan, National Research Council [5] in Canada, University of Konstanz [7] in Germany, Aristotle University [8] in Greece, and HP Laboratories [11] in Israel, etc. However, since the software and hardware environment for 3D models comes to maturity just recently, the research topic is a new challenge for all researchers. There are many experimental systems available, however, up to now, only two 3D model search engine systems are relatively "complete" on the Web for usage all over the world: one is built by Funkhouser et al. in Priceton University and the other one is developed by us. The search engine built by Funkhouser et al. is published in *ACM Trans. on Graphics* in Jan, 2003 [2], and our approach will be published in *EUROGRAPHICS 2003* that is also published in *Computer Graphics Forum* in Sep, 2003 [1]. Different methods of 3D model retrieval are used for each system: the former is geometry-based approach and the latter is image-based. According to the experimental results [1], the retrieval precision (precision-recall diagram) of our approach is 42% higher than that of method proposed by Funkhouser et al. That is, the retrieval results of our approach will be closer to human perception than that of method proposed by Funkhouser et al.

Previous works of 3D model retrieval can be broadly classified into two categories: geometry-based and image-based approaches. Geometry-based approach matches 3D models according to geometric distribution, and image-based approach does according to the similarity of rendered projection. One advantage of geometry-based approach is in the use 3D characteristics, such as topology structure [4], [9] and curvature of a patch [10]. However, higher dimension and irregular sampled points make the analysis more difficult. Moreover, invisible polygons will damage matching results in geometry-based approach. For example, 3D models created by modelling tools usually possess invisible polygons especially in articulation, but 3D models digitized by scanning tools do not. On the other hand, one advantage of image-based approach is easiness of processing since images are in lower dimension and in regular sampled points. Besides, unlike still images, the rendered images are well segmented and easily matched. Nevertheless, image-based approach might lose 3D information and misplace the corresponding rendered images between two 3D models.

3. Feature extraction for 3D model retrieval

The steps of extracting the *Lightfield Descriptors* [1] for a 3D model are shown in Figure 1, and detailed in the following.

(1) Since each model has its own coordinated system, translation and scaling are applied first in order to ensure that a model is entirely contained in each rendered image. The input 3D model is translated from the center of the model to the origin of the world coordinate system, and then scales the axis of maximum length to be 1. The translation $T = (T_x, T_y, T_z)$ assigns the middle point of the whole model to be the new origin:

$$T_i = \frac{MaxCoor_i + MinCoor_i}{2}, \qquad i = x, y, z \qquad (1)$$

where the $MaxCoor_i$ and $MinCoor_i$ are the maximum and minimum coordinate value of i axis, respectively. The scaling is isotropic, and normalizes according to the maximum distance from x, y and z axes of the whole model:

$$S = \frac{1}{\min_{i=x,y,z}(MaxCoor_i - MinCoor_i)} \qquad (2)$$

Although the stages cannot get the exact translation and scaling between two 3D models, the image metric of our approach is robust against translation and scaling.

(2) Render images from the camera positions of the light

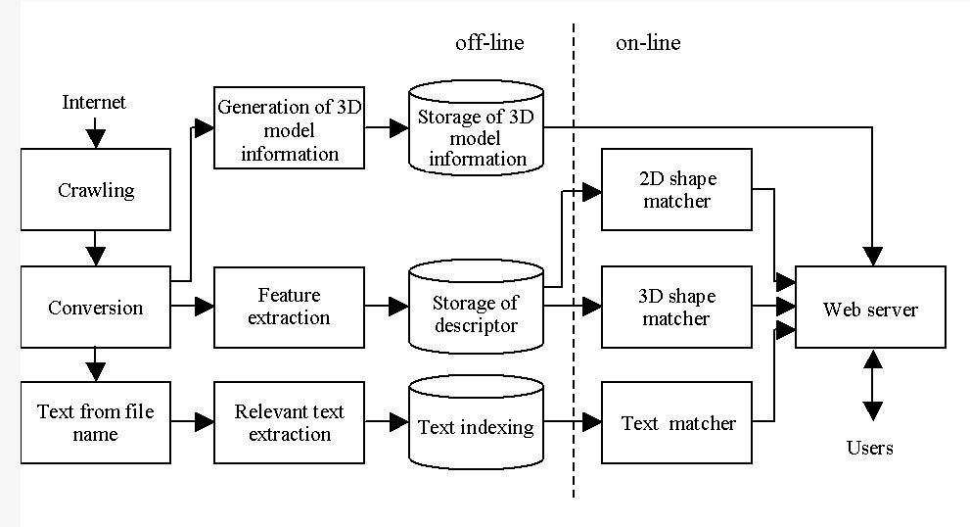


Figure 2: *Overflow of the search engine system*

fields, which is on the surface of a larger sphere. There are 10 light fields for each 3D model, and the camera positions of each light field are set at the 20 vertices of a regular dodecahedron. The camera at each viewpoint is directed towards the center of the sphere, and the up-vectors of cameras are placed uniformly. If two 3D models are of different orientations, their proper corresponding viewing images will have different rotational angles. It doesn't matter, since the image metric of our approach is also robust against rotations.

(3) We use an orthogonal projection in order to reduce the size of descriptors. Therefore, in a descriptor of light field, there are 10 images represented from 20 viewpoints. For a 3D model, 10 descriptors of light fields are created, so there are totally 100 images that should be rendered and extracted for features.

(4) Extract Zernike moment and Fourier Descriptor [13] from each image. Descriptors for a 3D model are those features from the 100 images.

4. 3D Model Search Engine Overview

The search engine system for 3D models consists of off-line pre-process and on-line retrieval process, as shown in Figure 2.In the off-line pre-process, a crawling is executed for downloading 3D models from the Internet first. File conversion is then applied for getting raw data of 3D models, including un-compression and graphics file format conversion. Now, there are 10911 3D models in our database after the conversion. For content-based retrieval, features of 3D models are extracted and stored in about 6 seconds on the average using a PC with Pentium III 800 CPU. For text-based retrieval, relevant texts are extracted according to file name of 3D models using WordNet [25] and indexed by RainBow [27] toolkit. In addition, thumbnails and other related information of 3D models are generated and stored for user browsing later.

In the on-line retrieval process, keyword search is available for text-based retrieval by using RainBow toolkit. For content-based retrieval, a user friendly interface of drawing 2D shapes is also provided to retrieve 3D models easily even for a novice. Querying by a retrieved 3D model interactively and iteratively allows user to get 3D models more similar and specific. The retrieval is down in a PC with two Pentium IV 2.4GHz CPUs. Only one CPU is used for the query at one time, and the retrieval takes 2 and 0.1 seconds with a 3D model and two 2D shapes as the queried keys, respectively. Several implementation details are listed in the following:

(1) Crawling: The 3D models collection is built by crawling from the Internet. However, it's not easy to collect 3D models from the Web since there are many different file formats for 3D models, and, what is worse, they are usually compressed in different compressing formats. The crawling process focuses on several web sites containing high quality 3D models. The crawling is executed in UNIX using command "wget" in shell. The downloaded files are saved in folder named after the downloaded path in order to avoid overlapping among files with the same name. For example, a file "xxx.zip" downloaded from "http://www.3dcafe.com/models/xxx.zip" will be saved to "http/www.3dcafe.com/models/xxx.zip".

(2) Conversion: As mentioned above, many 3D models are compressed and in different graphic file formats. In this conversion stage, three steps are applied in a PC with Windows 2000. First, in order to avoid repetition of the same file name after un-compression in a fold, for each downloaded

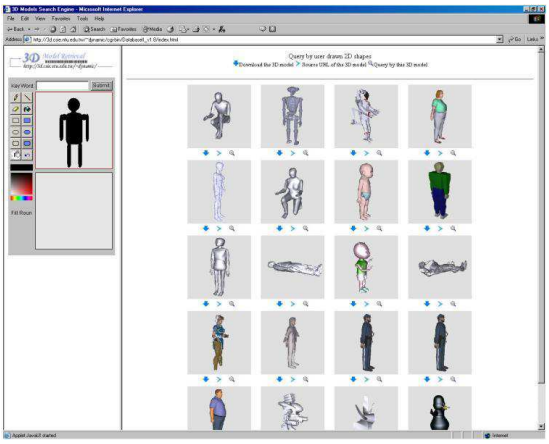


Figure 3: *Retrieval results from user drawn 2D shapes*

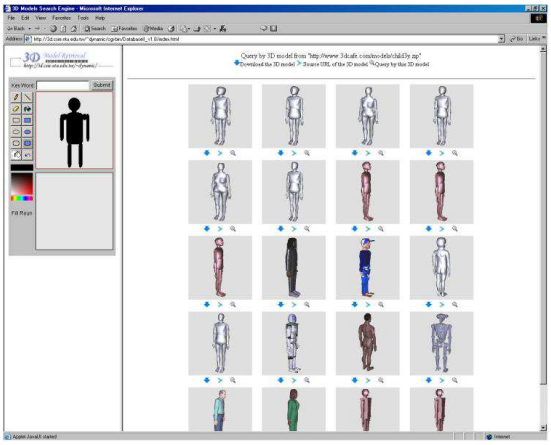


Figure 4: *Retrieval results from selecting a 3D model*

file, a folder is created according to the file name and then the file is moved to the folder. At the same time, the directory is also recorded where the model comes form. Second, PowerArchiver Command Line (PACL) [34] is adopted in our un-compression tools in command line, and can deal with many compression formats, such as zip, rar, lha, etc.

Third, Deep Exploration 2.0 [33]is used for converting graphics file formats since the software can convert batch from a folder, and can process many different graphics file formats, such as 3ds, wrl, obj, max, etc.

Thus, the file name of each 3D models is appended a number to avoid repetition and copied to a folder, and then copy the converted file format back after removing the added number. The file format of Wavefront OBJ [30] file (*.obj) is adopted in our implementation, and the material is save as another file (*.mtl). As a result, for example, a 3D model "yyy.3ds" is un-compressed from "xxx.zip", and the converted file will be put in "http/www.3dcafe.com/models/xxx.zip/yyy.obj". Note that, the conversion may be failed in un-compression or graphic file format conversion. Therefore, valid 3D models are checked and listed after measuring in several criteria, such as polygon number and vertex number.

(3) Feature extraction and storage: The approach of feature extraction for content-based retrieval is proposed in Section 3. The features are extracted in a PC with a Pentium III 800MHz CPU and GeForce2 MX video card in windows 2000. On the average, each 3D model with 7,540 polygons takes 5.7 seconds to extract features, and the average time of rendering and extracting a 2D shape takes about 0.06 seconds. Extracting features of 3D models is also suitable for both 3D model and 2D shape matching. No extra effort should be done for 2D shapes. When saving the features, coefficients of one image metric for all models are saved in a file. For each 3D model, the *LightField Descriptors* are saved sequentially, and are in a pre-defined order. For each *LightField Descriptor*, rendered images are also in a pre-defined order for storage. For each image metric, coefficients are saved in a pre-defined order if more than one coefficient is in the image metric. For each coefficient, only 8 bits are saved. For retrieval from database with large number of models, each coefficient of Zernike moment is also saved for 4 bits in another file. As a result, the features are 37.5MB and 18.7MB for Zernike moment, 10.4MB for Fourier Descriptor and 1.0MB for Circularity, whereas the raw data of all 10911 3D models is about 9GB.

5. Experimental results

Figure 3 shows a typical example of querying by user drawn 2D shapes of "human" model. Many "human" 3D models are retrieved. Figure 4 shows the interactive search by selecting a "human" 3D model from Figure 3. As we can see, after the iterative querying, retrieval results will return more human models with similar shape.

Our search engine with 10911 3D models has been publicly available on the Web since Jan. 2003. Up to now, the search engine has been improved for many times. The usage of the search engine is only listed on querying by 2D shapes and 3D models from "Find Similar" button or browsing. From Jan. 4 2003 to May 20 2003, the search engine served 3,475 queries (not counting queries from our lab by removing IP address 140.112.29.xxx), including 2144 query by 2D shapes and 1331 query by 3D model. Those queries come from 418 unique hosts, and in at least 27 different countries (Domain name can only be found in 309 IP addresses). The countries include Australia (.au), Belgium (.be), Brazil (.br), Canada (.ca), Switzerland (.ch), Czech Republic (.cz), Germany (.de), France (.fr), Greece (.gr), Hong Kong (.hk), Croatia/Hrvatska (.hr), Israel (.il), Italy (.it), Japan (.jp), Republic of Korea (.kr), Malta (.mt), Mex-

ico (.mx), The Netherlands (.nl), Poland (.pl), Portugal (.pt), Singapore (.sg), Turkey (.tr), Taiwan (.tw), Ukraine (.ua), United Kingdom (.uk), Yugoslavia (.yu), and USA (.edu). The educational institutions in USA (.edu) include UMN, Princeton, TAMU, CMU, Washington, Purdue, Dratmouth, Arizona and MIT, etc. Other domain names include .com, .gov, .mil and .net.

6. Conclusion and future works

In this paper, the work for creating a 3D model search engine is introduced. *LightField Descriptors* are extracted for matching similarity among 3D models. Users can query 3D model by text, 2D shape and 3D shape. Other possible applications and design issues are also described in this paper.

Several future works are listed in the following. First, automatic crawling 3D models in all format is an important step for a 3D model search engine. Then, query interface also need to be improved in the future. Next, browsing is also an important issue for a content-based search engine [28]. Therefore, better browsing tools can be built automatically for a 3D model search engine.

References

1. D.-Y. Chen, X.-P. T., Y.-T. Shen and Ming Ouhyoung, "On Visual Similarity Based 3D Model Retrieval", to appear in *Computer Graphics Forum (EUROGRAPHICS '03)* , **22**(3), Sept. 2003. 3
2. T. Funkhouser, P. Min, M. Kazhdan, J. Chen, A. Halderman, D. Dobkin and D. Jacobs, "A Search Engine for 3D Models", *ACM Transactions on Graphics*, **22**(1):83-105, Jan. 2003. 3
3. R. Osada, T. Funkhouser, B. Chazelle and D. Dobkin, "Shape Distributions", *ACM Transactions on Graphics*, **21**(4):807-832, Oct. 2002. 3
4. M. Hilaga, Y. Shinagawa, T. Kohmura and T. L. Kunii, "Topology Matching for Fully Automatic Similarity Estimation of 3D Shapes", *Proc. of ACM SIGGRAPH*, pp. 203-212, Los Angeles, USA, Aug. 2001. 3
5. E. Paquet, M. Rioux, A. Murching, T. Naveen and A. Tabatabai, "Description of Shape Information for 2-D and 3-D Objects", *Signal Processing: Image Communication*, **16**:103-122, Sept. 2000. 3
6. R. Ohbuchi, T. Otagiri, M. Ibato and T. Takei, "Shape-Similarity Search of Three-Dimensional Models Using Parameterized Statistics", *Proc. of 10th Pacific Graphics*, pp. 265-273, Beijing, China, Oct. 2002. 3
7. D. V. Vranic and D. Saupe, "Description of 3D-Shape using a Complex Function on the Sphere", *Proc. of IEEE International Conference on Multimedia and Expo (ICME)*, pp. 177-180, Lausanne, Switzerland, Aug. 2002. 3
8. I. Kolonias, D. Tzovaras, S. Malassiotis and M. G. Strintzis, "Fast Content-Based Search of VRML Models based on Shape Descriptions", *Proc. of International Conference on Image Processing (ICIP)*, pp. 133-136, Thessaloniki, Greece, Oct. 2001. 3
9. D.-Y. Chen and M. Ouhyoung, "A 3D Object Retrieval System Based on Multi-Resolution Reeb Graph", *Proc. of Computer Graphics Workshop*, pp. 16, Tainan, Taiwan, June 2002. 3
10. L. Cieplinski, M. Kim, J.-R. Ohm, M. Pickering and A. Yamada, "Text of ISO/IEC 15938-3/FCD Information Technology - Multimedia Content Description Interface - Part 3 Visual", *ISO/IEC JTC1/SC29/WG11/N4062*, Singapore, Mar. 2001. 3
11. M. Elad, A. Tal, and S. Ar. "Content Based Retrieval of VRML Objects - An Iterative and Interactive Approach", *Proc. of 6th Eurographics Workshop on Multimedia*, pp. 97-108, Manchester UK, Sept. 2001. 3
12. C. Zhang and T. Chen, "An Active Learning Framework for Content-Based Information Retrieval", *IEEE Transactions on Multimedia Special Issue on Multimedia Database*, **4**(2):260-268, June 2002. 3
13. D. S. Zhang and G. Lu. "An Integrated Approach to Shape Based Image Retrieval". *Proc. of 5th Asian Conference on Computer Vision (ACCV)*, pp. 652-657, Melbourne, Australia, Jan. 2002. 4
14. T. Igarashi, S. Matsuoka and H. Tanaka, "Teddy: A Sketching Interface for 3D Freeform Design", *Proc. of ACM SIGGRAPH*, pp. 409-416, Los Angeles, USA, Aug. 1999. 1
15. J. M. Martinez, MPEG-7 Overview (Version 8.0), *ISO/IEC JTC1/SC29/WG11/N4980*, July 2002. 1
16. Paul J. Besl and Neil D. McKay, "A Method for Registration of 3-D Shapes", *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, **14**(2): 239-256, 1992. 1
17. M. Levoy, K. Pulli, B. Curless, S. Rusinkiewicz, D. Koller, L. Pereira, M. Ginzton , S. Anderson, J. Davis, J. Ginsberg, J. Shade and D. Fulk, "The Digital Michelangelo Project: 3D Scanning of Large Statues", *Proc. of ACM SIGGRAPH*, pp. 131-144, New Orleans, USA, July 2000. 1, 2
18. F.-C. Wu, M. C.-C. Ho and M. Ouhyoung, "Automatic Modeling of Animated Faces from Multiview Video", *Proc. of EUROGRAPHICS* **20**(3):211-217, Manchester, UK, Sept. 2001. 1
19. G.-L. Perng, W.-T. Wang and M. Ouhyoung, "A Real-time 3D Virtual Sculpting Tool Based on Marching Cubes", *Proc. of International Conference on Artificial Reality and Tele-existence (ICAT)*, pp. 64-72, Tokyo, Japan, Dec. 2001. 1

20. T. A. Galyean and J. F. Hughes, "Sculpting: an interactive volumetric modeling technique", *Proc. of ACM SIGGRAPH*, pp. 267-274, July 1991. 1

21. E. Praun, H. Hoppe and A. Finkelstein, "Robust Mesh Watermarking", *Proc. of ACM SIGGRAPH*, pp. 49-56, Los Angeles, USA, Aug. 1999. 2

22. F. Arman and J. K. Aggarwal, "Model-Based Object Recognition in Dense-Range Images - A Review", *ACM Computing surveys*, **25**(1):5-43, 1993. 2

23. P. L. Besl and R. C. Jain, "Three-Dimensional Object Recognition", *ACM Computing surveys*, **17**(1):75-145, Mar. 1985. 2

24. A. P. Singh and D. L. Brutlag, "Protein Structure Alignment: A Comparison of Methods", submitted, 2001. 2

25. G. A. MILLER, "WordNet: A lexical database for English", *Communications of the ACM*, **38**(11):39-41, 1995. http://www.cogsci.princeton.edu/ wn/ 4

26. G. J. Kowalski and M. T. Maybury, *Information Storage and Retrieval Systems: Theory and Implementation*, 2nd Edition, MA: Kluwer Academic, 2002. 1, 2

27. A. Mccallum, Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering, 1996. http://www.cs.cmu.edu/ mccallum/bow/. 4

28. M. Hearst, A. Elliott, J. English, R. Sinha, K. Swearingen and K.-P. Yee, "Finding the Flow in Web Site Search", *Communications of the ACM*, **45**(9):42-49, Sept. 2002. 6

29. 3ds Max, http://www.discreet.com/ 1

30. Maya, http://www.aliaswavefront.com/ 1, 5

31. AutoCAD, http://www.autodesk.com/ 1

32. DigiBox, http://www.3dfamily.com/ 1

33. Deep Exploration 2.0, http://www.righthemisphere.com/ 5

34. PowerArchiver Command Line (PACL), http://www.powerarchiver.com/ 5

Enhanced 3D Model Retrieval System through Characteristic Views using Orthogonal Visual Hull

Shuen-Huei Guan* Ming-Kei Hsieh* Chia-Chi Yeh* Bing-Yu Chen†

National Taiwan University

1 Introduction

The introduction of the Princeton Shape Benchmark database (PSB) [Shilane et al. 2004] has given rise to extensive research into ways to accurately perform 3D matching. Among the extant methods proposed, the LightField Descriptor (LFD) [Chen et al. 2003] based method currently is the most accurate, but it suffers from a deficiency in on-line matching speed. The contribution of this research is twofold. First, fewer characteristic views are used instead of huge images to boost the on-line retrieval time. Second, depth information is employed to keep the retrieval accuracy.

2 Description

In Chen et al.'s system [Chen et al. 2003], several images, taken from different views around, are used to represent a 3D model, which are then transformed into descriptors. Thereafter, the matching of any two 3D models is achieved by matching two groups of 2D images by those descriptors. 100 images for each model are required, represented by 4 descriptors, which are Fourier descriptor for contour region, Zernike moment descriptor for shape region, circularity and eccentricity.

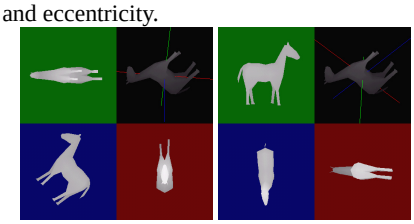


Figure 1: Different pose estimations by PCA (left) and OVH (right).

Our proposal to improve on-line speed is based on reducing the number of images required for each 3D model. Some other research made use of three principal axes by Principal Component Analysis (PCA), applied in pose estimation phase. Although PCA has nice meaning in statistics, it does not fit most cases in 3D model retrieval and its resulting views do not fit user's input of 2D sketches, as illustrated in Figure 1.

Hence, we propose Orthogonal Visual Hull (OVH) to establish pose estimation. Visual hull is an algorithm to reconstruct a 3D model from several views. Given a 3D model and images taken around, since visual hull does not capture concavities, the volume of the reconstructed model is always greater. In other words, the smaller volume we get, the closer approximation it fits the original model. There are two criteria to determine the number and mutual relation of views: 1) The views can be transformed into characteristic axes without obstruction and 2) the visual hull by these views is very close to the original model. Therefore, we chose three orthogonal views because there are three characteristic axes in pose estimation, and relatively orthogonal views carve much space in most cases. Then the next task is to get the views forming the least volume. The global minimum is extracted by the heuristic: (i) find local extrema from widely varying starting orientation of the 3D model by an iterative stochastic method or downhill simplex algorithm as Figure 2, (ii) pick the most extreme of these. The method is slow, but off-line and we are exploring to use faster algorithms, like conjugate gradient descent. Sometimes, the extreme is not unique. In our experiment of randomly picking one extrema to 3D model retrieval system [Chen et al. 2003], the resulting retrieval accuracy is nearly the same.

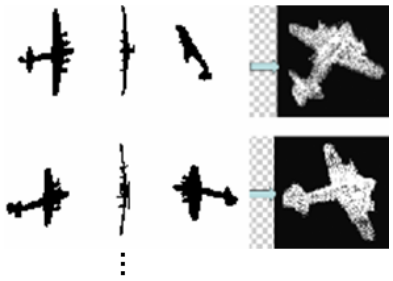


Figure 2: Heuristically pick the characteristic views that form the least volume.

Using fewer images raises the possibility of losing information necessary for 3D matching. By employing the Generic Fourier Descriptor (GFD) [Zhang and Lu 2002], we are able to compensate the information loss because it supplies additional depth information. Therefore, for each image, its contour, shape and depth information are all covered. In order to make use of GFD, the system takes 2 images for each characteristic axes. Consequently, the total number of images taken from each 3D model is reduced from 100 to 6. The number of pair-wise image matching is reduced from 5460 to 24. And the retrieval accuracy is nearly the same to the original system.

In summary, this research enhances the extant best matching method [Chen et al. 2003] by two means: 1) improving the on-line speed 93 times faster at the cost of 5 times deceleration of the off-line process, and 2) keeping the retrieval accuracy. The proposed OVH is also a brand new idea for pose estimation. Additionally, the shapes of the resulting characteristic views generated by OVH more closely fit human visual perception.

	Storage Size (bytes)	Off-line Generate Time (s)	On-line Match Time (ms)
LFD	4,700	3.25	1.300
LFD with OVH	2,592	15.72	0.014

References

CHEN, D.-Y., TIAN, X.-P., SHEN, Y.-T., AND OUHYOUNG, M. 2003. On visual similarity based 3d model retrieval. *Conputer Graphics Forum 22*, 3, 223–233. (Proc. EG 2003).

SHILANE, P., MIN, P., KAZHDAN, M., AND FUNKHOUSER, T. 2004. The princeton shape benchmark. In *Proc. SMI 2004*.

ZHANG, D., AND LU, G. 2002. Generic fourier descriptor for shape-based image retrieval. In *Proc. ICME 2002*, vol. 1, 425–428.

*e-mail:{drake, lionkid, id5}@cmlab.csie.ntu.edu.tw

†e-mail:robin@ntu.edu.tw

分項計畫四：MPEG-4/7保證服務品質之媒體內容傳輸技術 (III)

MPEG-4/7 QoS-Supported Content Delivery Technology (III)

成果報告

計畫編號：NSC91-2622-E-002 –002

全程計畫：民國 90 年 08 月 01 日至民國 93 年 07 月 31 日

本年度計畫：民國 92 年 08 月 01 日至民國 93 年 07 月 31 日

計畫主持人　：黃肇雄　台灣大學資訊工程系教授
　　周承復　台灣大學資訊工程系助理教授
共同主持人　：吳家麟　台灣大學資訊工程系教授
　　陳文進　台灣大學資訊工程系教授
　　歐陽明　台灣大學資訊工程系教授
　　陳炳宇　台灣大學資訊管理系助理教授

一、 摘要

本分項計畫『MPEG-4/7 保證服務品質之媒體內容傳輸技術』，劃分為兩項技術的研發：(1)【智慧型頻寬調節傳輸模組】(Intelligent Bandwidth Adapted Transmission module)與(2)【MPEG-4/7 保證服務品質的代理伺服器】(MPEG-4/7 QoS-Supported Proxy Server) 。所發展出來的技術，將擔任總計畫中"互動資訊媒體服務系統（Interactive Information Media Service System）"封裝傳輸的角色及提供網路頻寬的管理和服務品質的保證。

在【智慧型頻寬調節傳輸模組】的技術部分： 為了能適應網際網路上快速變動的可用頻寬，客戶端不斷的將接收情況反映給伺服器端。伺服器端將根據所接收而來的參數，使用適當的方法，如速率調節（Rate shaping）等等，以調整傳輸頻寬與模式，使客戶端能有效率的利用網路資源，達到最佳的服務品質。

在【MPEG-4/7 保證服務品質的代理伺服器】的技術部分：代理伺服器(Proxy Server)將先分析媒體內容(Content)中 MPEG-7 定義的述詞(Description)和描述結構(Description Scheme) 網路相關參數及詢問客戶端、原伺服器(Origin Server)所需資源來建立適當的連線，以保證傳輸品質，並輔以暫存媒體內容(Content)的片頭資料，可有效縮短初始網路延遲(Network Initial Delay)、網路延遲差異(Network Jitter)。如此，藉由了解媒體的特徵，有效的利用有限的資源，以滿足不同媒體間對傳輸品質的個別要求。在底層串流傳輸的協定方面，我們將採用工業界標準，即時串流協定 RTSP(Real Time Streaming Protocol) 為控制協定，以 RTP(Real-time Transport Protocol)/RTCP(Real-Time Control Protocol)來傳送即時性資料。本計畫的目的在於研究未來互動電視的傳輸格式及設計新一代代理伺服器(Proxy Server)來保證服務品質，落實網路多媒體應用的可行性。

The project , 『MPEG-4/7 QoS-supported content delivery technology』, will develop two technologies : (1)【MPEG-2 Transport Stream Model for MPEG-4 Data】and (2) 【MPEG-4/7 QoS-Supported Proxy Server】. Technologies developed will support the transport stream layer of interactive TV proposed by the collaborated project and both manage the using of bandwidth and guarante the quality of service .

The proposed【Intelligent Bandwidth Adapted Transmission module】technology : In order to adapt the dynamic available bandwidth on Internet, Clients continueously report the receicing status to the Server. After receiving the parameter, the Server will adjust the transmission bandwidth and mode by suitable schemes, e.q., rate shaping, FEC coding to make Clients use network bandwidth efficiently and get the highest quality.

The proposed 【MPEG-4/7 QoS-Supported Proxy Server】technology : The proxy server will extract the network-related parameters from Ds and DSs defined in MPEG-7 and ask the origin server and the client for what is the request of resource to create proper connections for achieving the quality of servive. And, caching the prefix data of content, proxy server can minilize the network initial delay and network gitter. So, by understanding the properties of content and using the limited resource more efficiently, proxy server can achieve the individual request of network transport of different media applications. In the part of transport protocol, we apply the standard protocols of industry. Using RTSP (Real Time Streaming Protocol) for the control protocol and RTP(Real-time Transport Protocol)/RTCP(Real Time Control Protocol) for transfering real time data.The goal of this project is to develop the transport stream layer of the interactive TV and to design the new generation proxy server to guarantee the quality of service and realize the network media application.

二、 計畫緣由與目的

網際網路的逐漸普及，不僅意味著使用者人數的快速膨脹，並且連帶的使得其上的服務變得難以預測與管理。然而，網際網路上的傳輸需求，無論是品質與規模，卻是與日遽增。使用者對於傳統的網路服務，如電子郵件與網頁瀏覽，已經漸漸的不能滿足，而有更高品質的要求。特別是各種多媒體的應用，提供了更為即時，更為生動的影音享受，已是目前網際網路服務的主流之一。這些多媒體應用大多以串流模式(streaming mode)提供服務，以減少使用者的等待時間。目前常見的多媒体內容，有聲音、影片、圖片等，而以影片佔傳輸量的大部分。

雖然高速網路已大為普及，但在網際網路並未提供服務品質保証(QoS Guarantee)、無線網路的頻寬仍低及多媒体應用的使用人數與日俱增的情況下，「網路資源不足」仍是串流服務常遇到的問題。最原始的串流機制只是不斷地傳輸多媒體檔案，一旦遇到網路壅塞，使用都所觀看影片的品質便大幅滑落。因此設計一套既能高效率地使用網路資源，又能同時提高串流應用之服務品質的傳輸機制，便成為研究者們努力的目標。

要能真正落實網路互動多媒體的應用，除了制定相關的國際標準外，亦須有完善的網路基礎建設。光纖骨幹及ADSL已將我們領進寬頻的世界，然而網路頻寬的拓寬僅能增加傳輸的資料量，並無法縮短初始網路延遲(Network Initial Delay)，且寬頻並不意味著頻寬可無限使用，亦需藉由頻寬的控管來保證服務品質。有鑑於此，本計畫基於「對客戶端作出立即性反應」及「保證服務品質」的需求，設計了新一代代理伺服器(Proxy Server)與傳輸模組。

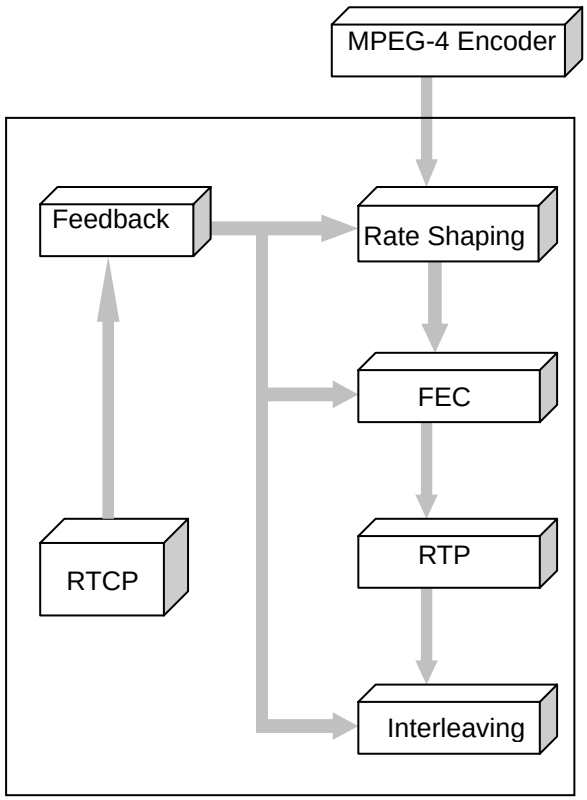
在本計畫的前兩年中，我們已經完成了【智慧型頻寬調節傳輸模組】與【MPEG-4/7保證服務品質的代理伺服器】的初步架構。本年度計畫針對其上的功能作更進一步的改良與加強。在【智慧型頻寬調節傳輸模組】方面，我們在速率調節（Rate shaping）模組中加強了「選擇性丟棄訊框」（Selective Frame Discard，SDF）功能；至於【MPEG-4/7保證服務品質的代理伺服器】，我們則加強了代理伺服器間的協同工作功能；包含了暫存決策器（caching policyer）、轉向器(Redirector)與多源傳輸（multi-source transmission）模組。

三、 研究方法與成果

【智慧型頻寬調節傳輸模組】

RTP提供了傳送具有即時性資料的功能，但是它並沒有想像中的去做資源預約 (resource reservation)或保證服務品質(QoS) 的功能，它的運作方式就是在要送出的資料前加上RTP的封包標頭 (packet header)，如圖九，利用這些標頭它可以提供許多有用的資訊，例如： payload type identification、sequence numbering、timestamping和delivery monitoring，再加上RTCP的回饋 (feedback) 功能，我們就可以更清楚的了解網路狀況，以達到即時性傳送的要求。

蒐集RTCP所回饋的資訊後，必須根據目前的網路與使用者狀況，做出相對應的動作，已獲取最佳的網路使用效能。目前這部分並沒有公定的標準，因此我們在此設計一套系統，對RTCP所提供的回饋訊息做出適當的決策，如圖一所示。



圖一、保證服務品質模組架構圖

為了節省儲存及傳輸的成本，影片大多會先經過壓縮處理。然而，經過壓縮的影片，尤其是不定位元率(Variable Bit Rate)的影片，都會有明顯的激增資料量(Burst)。資料量激增的原因主要來自於編碼(Encode)的機制和影片內容的特性：動作激烈或場景變換頻繁的影片，其激增的資料量較為明顯。影片資料量

的激增將使串流機制的設計更為複雜。在經由資源有限的網路傳輸影片時，最原始的方法就是不理會網路資源的限制，傳送每一個訊框(Frame)。但在網路資源不足時，封包遺失是在所難免的。然而每個訊框對於使用者觀看的影片服務品質的重要性不同，若是為了傳送重要性較低的訊框而造成重要的訊框被丟棄，不啻為一種網路資源的浪費。此外，使用者端的播放程式也可能因為訊框到達的時間太晚，來不及在該訊框的播放時間播出，而將之丟棄。這也造成網路資源的浪費。

為了改善上述問題，選擇性丟棄訊框(Selective Frame Discard)的概念被提出，簡稱為 SFD。其基本想法是由串流伺服器考量網路資源、使用者端緩衝區大小的限制、影片內容特性及使用者對服務品質的要求，選擇性地丟棄對於服務品質較不重要的訊框。SFD 的優點有 ：

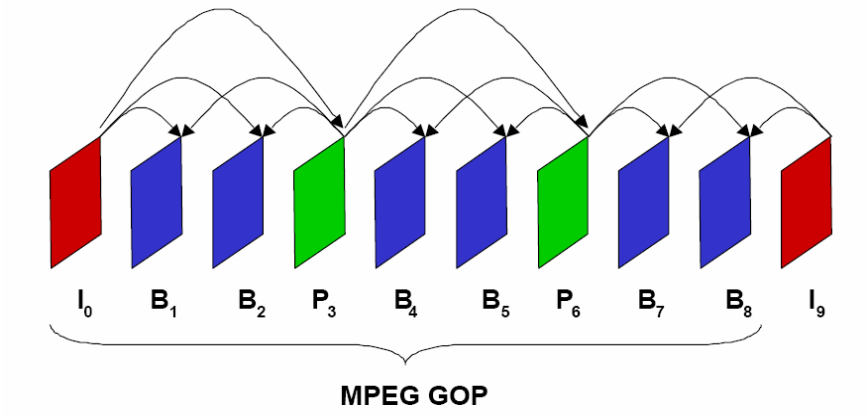
1．藉由將網路頻寬和使用者端緩衝區大小的限制入考量，串流伺服器可選擇性地先丟棄部分訊框，降低網路上的資料量，以減少之後其他訊框被丟棄的可能性，因而提高網路資源的使用效率。
2．選擇較為重要的訊框傳送，以同樣的頻寬可達到較高的影片服務品質。
3．相較於位在網路層的封包丟棄機制 (Packet Discard)，位在應用層的ＳＦＤ可利用對多媒體內容的了解，如編碼方式或訊框間的依賴關係 (Frame Dependency)，做出更恰當的決定。

以下我們將詳細介紹 SFD 模組的運作原理。

一、MPEG

MPEG 為一被廣泛使用的影片壓縮標準，它定義了三種訊框類型：I、P 和 B。在壓縮時，Ｉ訊框是獨立地被編碼，如同一張圖片。Ｐ訊框則以前一張Ｉ或Ｐ訊框為參考訊框(Reference Frame)，只儲存和參考訊框的差異，因此必須在參考訊框存在時才能被正確地解碼。Ｂ訊框則以之前和之後的Ｉ、Ｐ訊框為參考訊框。Ｉ訊框帶有最大量的的資料，Ｐ次之，Ｂ最少。

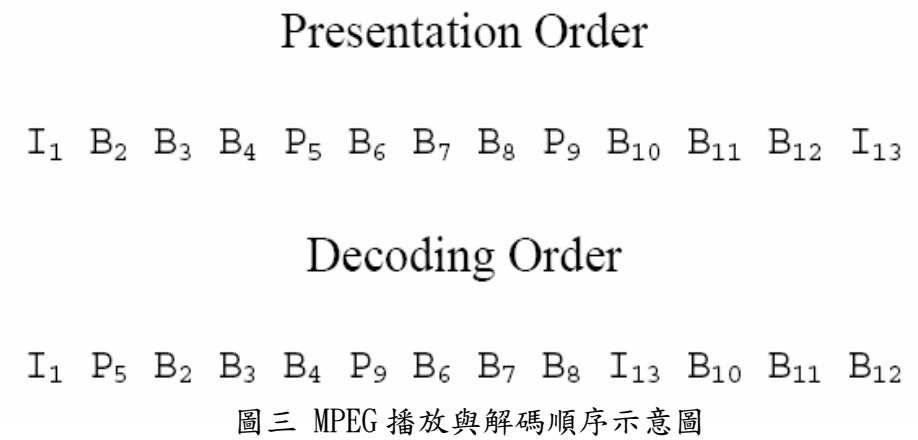
MPEG定義GOP（Group Of Picture）為一系列連續的訊框，其第一張訊框



圖二、 訊框間依賴關係示意圖

類型為I，包含到下一張I訊框的前一張訊框，換句話說，一個GOP包含了一張I訊框及下一張I訊框之前的所有P、B訊框。如果訊框A必須在訊框B的資料正確無誤的情況下，才能被正確地解碼，則稱訊框A依賴(Depend)於訊框B。圖二的箭頭指向依賴的訊框，例如B_1依賴I_0和P_3。

由於訊框間的依賴關係，訊框的解碼順序和播放順序是不同的，如圖三。因為每個訊框必須在依賴的訊框完整地被解出後，才能被正確地解碼，例如圖三的B_2、B_3和B_4，都依賴於I_1和P_5，因此這解碼的順序是先解I_1和P_5，才能夠解出B_2、B_3和B_4。



圖三 MPEG 播放與解碼順序示意圖

二、基本概念

影片的原理是利用人眼的視覺暫留原理，一秒鐘播放三十張訊框(Frame)造成連續的動作。當每秒播放的訊框數少於三十，便會有停頓的感覺。當我們從要播放的三十張訊框中抽走一張時，抽走和前一張愈相似的訊框，對人類視覺所造成的影響就愈小。因此我們希望在丟棄訊框時，先丟棄和前一張訊框較相似的訊框。我們將這個以相似度為基礎的SFD演算法，稱為SimFD (Similarity based Frame Discard)。

三、相似度 (Similarity)

為了評量訊框間的相似度，我們考量了PSNR和Color Histogram兩種評量方式：

A、 PSNR (Peak Signal Noise Ratio)

PSNR原本被用來估計一張重建後的圖片(Reconstructed Picture)和原始圖片的差異。其基本的用意是用一個數字來反映重建後圖片的品質，數字愈高代表重建後圖片的品質愈好，也就是愈接近原始圖片。

假設我們有一張原始圖片，$f(i,j)$，包含了N * N 像素(pixel)，和一張重建後的圖片，$F(i,j)$，這張重建後的圖片可能是由原始圖片的壓縮版本解壓縮而

得到，或是經由某種可能有資料遺失的媒介而得到。為了評比這張重建圖片的品質，我們先計算平均平方誤差(Mean Squared Error，MSE) :

$$MSE = \frac{\sum [f(i,j) - F(i,j)]^2}{N^2}$$

然後再計算 PSNR

$$PSNR = 10Log_{10}\left(\frac{255^2}{\sqrt{MSE}}\right)$$

因為 PSNR 和相似度成正比關係，我們將兩張訊框，*F(i)*、*F(j)*的相似度定義為

$$Similarity(F(i), F(j) = PSNR(F(i), F(j))$$

PSNR 也可用來評量一段影片和其重建後本的差異。假設原始影片由 n 張訊框組成，f(i)，而重建後影片為 F (i)，則其 PSNR 為

$$PSNR_F = \sum_{i=1}^{n} PSNR(f(i), F(i))$$

其中 $PSNR(f(i), F(i))$ 為原始影片第 *i* 張訊框和重建後影片第 *i* 張訊框的 PSNR 值。

B、Color Histogram

在電腦圖學領域中，Color Histogram 是一張影像的另一種表示, 以每種顏色佔有幾個像素，繪出長條圖以表示這張影像的色彩分布。

Color Histogram 並不限於特定的 Color Space。在此以 RGB 圖片為例，每個像素由紅(R)、綠(G)、藍(B)三個顏色組成，各以一個 0~255 的數字表示該種顏色的深淺。我們準備 R[256]、G[256]和 B[256]三個陣列，然後檢查每個像素，數出每個顏色出現的次數，例如檢查到一像素的 RGB 為(23,155,255)時，就將 R[23]、G[155]和 B[255]的值加一。兩張圖片愈相似，它們的 Color Histogram 就愈接近。因此我們定義兩張圖片，X 和 Y，的相似度為

$$Similarity(X, Y) = \frac{C_{sim}}{\sum |h_x(i) - h_y(i)|}$$

其中 $h_x(i)$、$h_y(j)$ 為 X、Y 的 Color Histogram。

傳送到使用者端的影片可能因伺服器主動丟棄或在網路上遺失部分訊框，為了評量遺失部分訊框的影片和原始影片的相似度，我們採用影片相似度(Video Similarity)的概念。

我們以 *v* 表示原始影片，含有 *f* 張訊框，*v(i)*為原始影片中的第 *i* 張訊框，*V* 為使用者端接收到的的影片，*r(i)*表示是否有正確地收到第 i 張訊框 - v(i)，V(i) 為使用者所看到，當成是第 i 張訊框的訊框，即在第 i 張訊框應播放的時間，最接近的已播放訊框:

$$V(i) = v(j), j \le i \wedge r(j) = true$$

使用者端接收到的影片和原始影片的相似度定義為 ：

$$Similarity(v,V) = \sum Similarity(v(i),V(i)) \Big/ f$$

假設原始影片的訊框為<I0 , B1 , B2 , P3 , B4 , B5 , P6 , B7, B8>，但伺服器丟棄了全部的 B 訊框，使用者端只收到 I0、P3、P6 , 於是使用者將看到的訊框為<I0 , P3, P 6>，由於訊框播放時間的設計，每個訊框在畫面上的出現時間將持續到下個訊框出現，因此我們將出現的訊框視為<I0 , I0 , I0 , P3 , P3 , P3 ,P6 , P6 , P6 >，故影片相似度為

(Similarity(I0,I0) + Similarity(B1,I0) + Similarity(B2,I0) + Similarity(P3,P3) + Similarity(B4,P3) + Similarity(B5,P3) + Similarity(P6,P6) + Similarity(B7,P6) + Similarity(B8,P6)) / 9

四、SimFD (Similarity-based Frame Discard)

SimFD 可分成兩部分：

1. 預先處理 (Preprocessing)：一般儲存在伺服器的影片檔案，都是以壓縮過的格式儲存。為五比較訊框間的相似度，必須先將影片檔解壓縮，還原成原始資料(Raw Data)後才可進行評量。由於解碼(Decode)的技術進步，即時性地解碼、比較訊框相似度是可行的。但為了減輕伺服器的負荷，將運算資源用於提供更多更佳的串流服務，我們將相似度的比較向前移到串流服務開始之前，由伺服器利用閒置的運算資源或藉由其他主機的協助，預先處理影片檔案，並將訊框間相似度的比較結果記於訊框相似度報告(Frame Similarity Report，FSR)中，FSR 可單獨地以文字檔案存在，或附於影片壓縮格式的標頭中。之後當伺服器在選擇要丟棄的訊框時，便可直接讀取 FSR，取得各訊框的相似度資訊。如此一來，我們先將每張訊框和前張訊框

的相似度記錄於訊框相似度報告(FSR)中。在比較相似度之前，必須先將訊框解碼，得到原始資訊後，才能進行相似度比較。為了減輕串流伺服器的負荷，我們將此步驟向前移到串流服開始之前，例如伺服器新增影片之時或伺服器的負擔較輕時。

2． 線上策略 (Online Policy)：當串流開始，伺服器便讀取 FSR，並根據訊框類型及依賴關係，在網路頻寬的限制下，丟棄部分訊框。

以下我們詳細介紹這兩部分的運作：

A、 預先處理 (Preprocessing)

基本上，SimFD 將訊框相似度視為權重的一部分，預先處理的目的就是事先計算好訊框相似度，然而訊框相似度並不只是跟前一張訊框比較，還需要依訊框類型不同進行必要的改變。

MPEG 的三種訊框類型，I、P 和 B，重要性和檔案大小都是 I 最高，P 次之，B 最小。因為訊框間的依賴關係，如果在丟棄訊框時丟棄了一張 I，就相當於丟棄了整個 GOP，而丟棄了一張 P，就相當於丟棄了這張 P 之後，一直到下一張 I 之間的所有訊框。只有 B 訊框是不被其他訊框依賴的，丟棄 B 訊框並不會影響到之後的訊框，因此成為丟棄訊框時的首選。當所有的 B 訊框都已被丟棄，接下來的考慮為 P 訊框，最後才是 I 訊框。

假設當我們把 B 訊框都丟完了，頻寬仍然不夠傳輸剩下的 I、P 訊框，此時就必須在 P 訊框中挑選犧牲者。需要特別注意的是，當我們開始考慮 P 訊框時，勢必已經丟光所有的 B 訊框，因此如果在訊框相似度報告中紀錄 P 訊框和其前一訊框的相似度是沒有意義的，因為如果該訊框的前一訊框為 B 訊框，則該 B 訊框一定已被丟棄，如此一來便失去和前一訊框比較、計算相似度的意義。是故 P 訊框在訊框相似度報告中應紀錄其和前一 I 訊框或 P 訊框的相似度。因此在考量 P 訊框的相似度時，應考量其前一張 I 訊框或 P 訊框，而非播放順序上的前一張訊框。

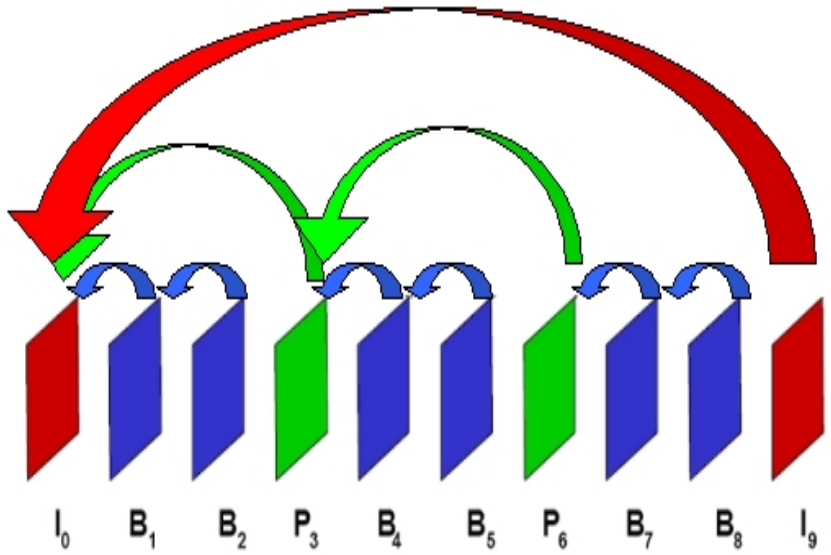
同理，I 訊框在訊框相似度報告中應紀錄其和其前一張 I 訊框的相似度。因當 SimFD 開始考慮要丟棄 I 訊框的時候，所以的 P、B 訊框都應該已被丟棄了。

由以上討論，訊框相似度報告中的內容應紀錄每張訊框和其前一張"有效"訊框的相似度。"有效"意指在SimFD執行時，實際上未被丟棄的訊框。每個訊框，$Frame_i$ ，在訊框相似度報告中擁有一對數字，< i , s >，i表示該訊框為第幾個訊框，s為該訊框和其前一有效訊框，Frame pre，的相似度。

1． $Frame_i$ 為I訊框時，$Frame_{pre}$為 $Frame_i$ 的前一個I訊框，

2． $Frame_i$ 為P訊框時，$Frame_{pre}$為 $Frame_i$ 的前一個I訊框，或P訊框

3． $Frame_i$ 為B訊框時，$Frame_{pre}$為 $Frame_i$ 的前一個I、P或B訊框



B、線上策略 (Online Policy)

SimFD 在串流服務開始之後，以 window 為單位，根據傳送速度控制模組(Rate Control Module)回報的適當傳送速度，估計目前 window 中的訊框是否可以順利地在播放時間限制下完成傳輸。一個 window 由一至數個 GOP 組成，window 的大小動態地依網路穩定性做調整。如果此 window 可順利地完成傳輸，SimFD 就可等待此 window 的傳輸完成後，檢查下一個 window。如果無法傳送整個 window，SimFD 先算出可傳輸的資料量當做目標大小，然後選擇性地丟棄部分訊框，直到 window 中的訊框總資料量符合目標大小，才傳送 window 中剩下的訊框。

- 丟棄 B 訊框

由於訊框間不同的重要性，SimFD 先從 B 訊框開始考慮。因為沒有任何訊框必須依賴於 B 訊框才能完成解碼，在 B 訊框間挑選犧牲者時只需考慮其和前一個訊框的相似度，故從相似度最高的 B 訊框開始丟棄，直到 window 的資料量小於目標大小。

- 丟棄 P 訊框

在 window 中所有的 B 訊框都被丟棄後，SimFD 開始考慮 P 訊框。然而 P 訊框彼此間也有依賴關係，每個 P 訊框必須在前一個 P 或 I 訊框的存在下才能正常地解碼，因此在同一個 GOP 中，無論相似度高低，都必須先丟棄最後的那張

P 訊框。如果 window 中只包含一個 GOP，那麼就從最後的 P 訊框開始丟棄，其次是倒數第二張，直到 GOP 中的 P 訊框全部被丟棄為止。window 中若是有兩個以上的 GOP，則可在各 GOP 中的最後一個 P 訊框間挑選和其前一有效訊框相似度最高者當犧牲者。

- 丟棄 I 訊框

在 P、B 訊框全部都被丟棄，卻仍然無法滿足目標大小的情況下，SimFD 將在剩下的 I 訊框間依其和前一個 I 訊框的相似度做出選擇，最壞的情況是網路頻低到整個 window 都被迫丟棄，如此 SimFD 便直接檢查下一個 window。

- PBP

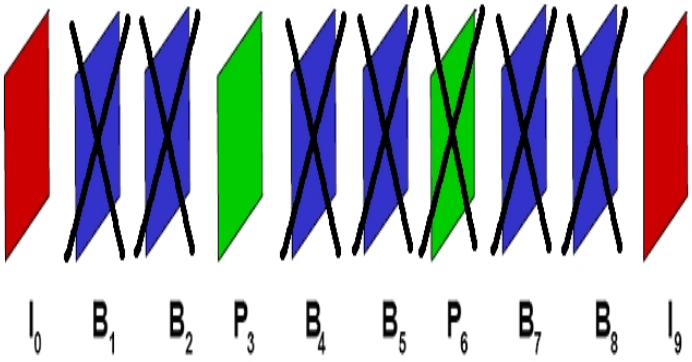
三種訊框類型中，I 訊框所帶資訊最多，檔案也最大。P 次之，檔案大小約為 I 訊框的一半，B 最小，約為 P 訊框的三分之一到四分之一。一般來說，除非有特別指定編碼器的選項(Encoding Option)，P 訊框的總數約為 B 訊框的一到兩倍。在一個包含三十張訊框的 GOP 中，除了第一張是 I 訊框外，剩下二十九張訊框中 P、B 約各佔一半。若我們將一 GOP 中的 B 訊框取走，則該 GOP 在播放時的訊框率(Frame Rate)只剩一半，但檔案大小卻只減少了約 20%。雖然表面上看起來傳送 B 訊框比傳送 I、P 訊框有效率，但那是因為 B 訊框必須仰賴 I、P 訊框的資訊才能正確地解碼，所以還是必須先傳送 I、P 訊框，不然 B 訊框也無法正確地解碼。

SimFD 的目的是在網路可傳輸量的限制下，儘可能傳送對服務品質幫助最大的訊框。依照上述丟棄訊框的法則，SimFD 會儘量先傳輸 I、P 訊框，就像是將 I、P 訊框塞入大小為網路可傳輸量的背包中，直到背包塞不下為止。然而我們發現，往往在背包，網路可傳輸量，已塞不下任何一個 I、P 訊框時，卻還有空間可以放得下數個 B 訊框 - 因為 B 訊框的檔案大小遠小於 I、P 訊框。

為了善加利用這些剩餘的空間，我們提出 PBP, 用 B 訊框填滿網路可傳輸量。然而盲目地將 B 訊框加回 window 中幫助並不大，因為 B 訊框必須在其依賴訊框存在時才能正確解碼。我們先挑出那些依賴訊框都將被傳送的 B 訊框，再依那些 B 訊框和前其一訊框的相似度，由低到高一一檢查，在網路可傳輸量的限制下加回 window 中。

例如在圖四中的GOP中，除了I_0和P_3外，其他的訊框都已經被丟棄。但因為B_1和B_2依賴的訊框，I_0和P_3，都將被傳送，所以如果B_1、B_2能傳送到使用者端，將可被正確解碼。故SimFD將試著在不超過網路可傳輸量的限制下，將B_1、B_2加回。

- 動態調整 window 大小



圖四 PBP 之一例

window 包含的 GOP 個數愈多，在選擇要丟棄的訊框時，可選擇到最不傷害服務品質的訊框，提高 SimFD 的效能。在理想的網路狀況下，極端的情況是一個 window 便包含整部影片全部的 GOP，在通盤的考量下，可做出最佳的決策。然而實際上網路狀態時常變動，在 window 開始時的估計往往和後續的頻寬變化有落差。例如當 window 開始時估計的頻寬可以傳輸約三分之二的訊框，但到了中途頻寬卻剩下一半，於是約有一半原本預估可順利到達的訊框被網路丟棄或延遲。

為了避免在網路不穩定時估計頻寬和實際狀況的落差太大，又可在網路情形良好時提高 SimFD 的效能，我們決定根據網路的穩定度調整 window 包含的 GOP 數，在網路穩定時增加 window 的大小，在網路頻寬變動較大時調降 window 的大小。

我們以 R_{pre}表示前次window開始時，傳送速率控制模組(Rate Control Module)所估計的適當傳送速率，R為此次的適當傳送速率。W為目前的window大小，W_{incre}為window每次增加的數值，W_{decre}為每次減少的數值，T_{stable}為用來判斷網路穩定度的臨界值。則：

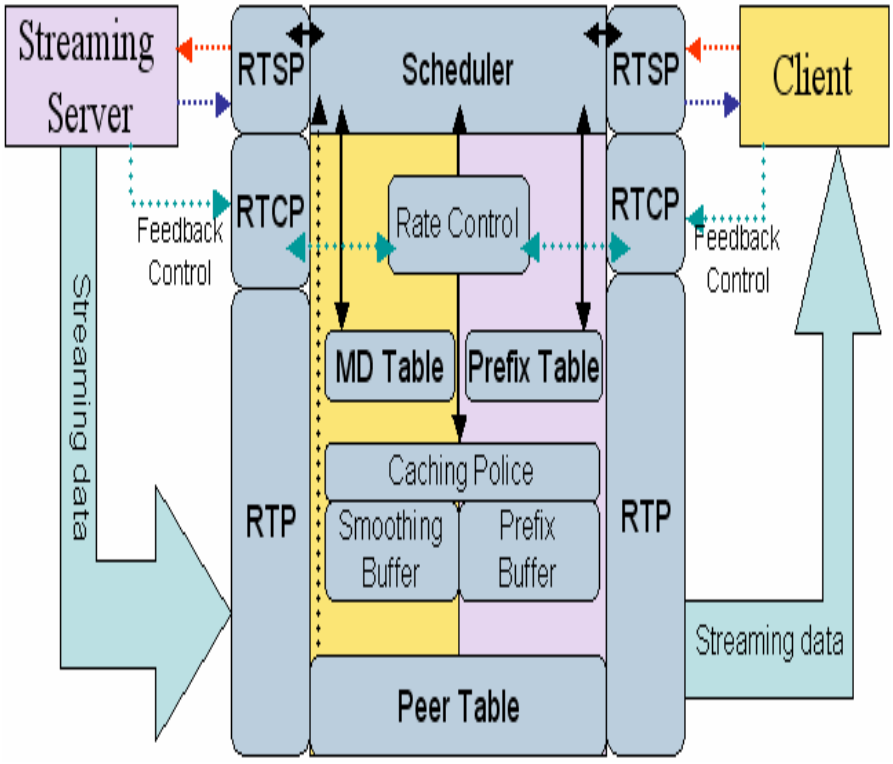
$$
\begin{aligned}
&if\ (|R - R_{pre}| < T_{stable}) \\
&\quad W = W + W_{incre} \\
&else \\
&\quad W = W - W_{decre}
\end{aligned}
$$

- Lopt (SimFD with Local Optimization)

在 SimFD 的線上策略(串流開始)中，在考慮訊框類型及訊框間依賴關係後，SimFD 先丟棄和前一訊框較相似的訊框。SimFD 所考慮的訊框相似度資訊是由

讀取訊框相似度報告(Frame Similarity Report)而來，訊框相似度報告中所記載的相似度是每個訊框和其前一有效訊框的相似度，有效訊框指的是前一個具有同等級或較高等級重要性的訊框。

【MPEG-4/7保證服務品質的代理伺服器】



圖五、代理伺服器的架構圖

圖五為新一代代理伺服器的架構圖。假設網路多媒體應用程式的客戶端有指定代理伺服器，則所有對原伺服器(Origin Server)所發出的要求，將會先傳送給代理伺服器(Proxy Server)，再查詢代理伺服器(Proxy Server)所暫存的相關資料，對原客戶端的要求做適當調正，然後才真正對原伺服器(Origin Server)發出要求，並回應客戶端。對於任何一個要求，代理程式(Proxy Server)會分析系統現況，如果可用網路頻寬、系統資源均滿足要求則允入，否則將拒絕要求。之後，代理伺服器(Proxy Server)將先到片頭暫存表(Prefix Table)檢查客戶端所要求的資料是否有暫存在片頭緩衝區(Prefix Buffer)內，若已存在，則到媒體描述表(Multimedia Description Table)內，取出此媒體內容(Content)的媒體資訊(Media information)及所要求的服務品質(Service Quality)，然後建立一條適當的連線，直接將片頭資料由代理伺服器傳向客戶端，並向原伺服器(Origin Server)要求後續的媒體資料，再將由原伺服器所接收到的資料儲存在平滑暫存區(Smoothing Buffer)以確保資料供給的穩定性。反之，若無法在片頭暫存表(Prefix Table)找到客戶端所要求的資料，則代理程式直接向原伺服器要求客戶端所需要的資料，並將片頭部分儲存在片頭緩衝區(Prefix Buffer)，更新片頭暫存表(Prefix Table)及媒體描述表(Multimedia Description Scheme Table)，且在接受資料的同時，將

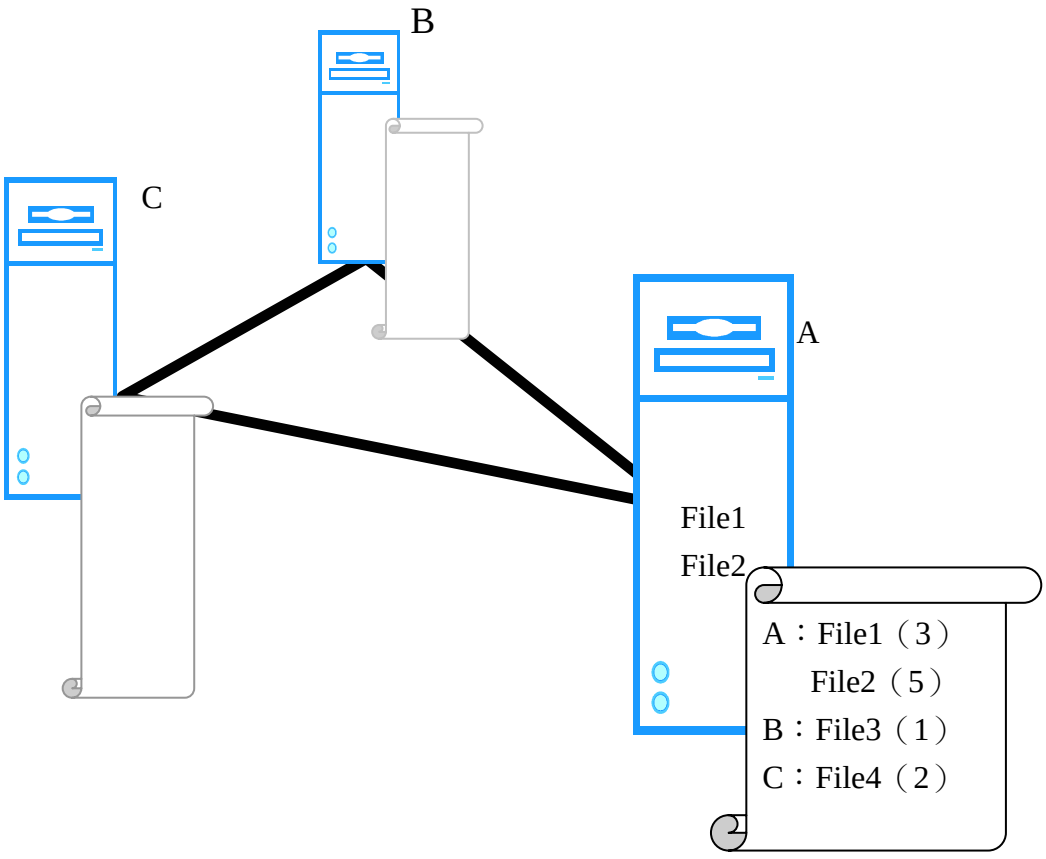
所接受到的資料傳送給客戶端。至於資料的暫存及取代則由暫存決策器(Caching Policyer)依特定的演算法來決定。

我們規劃將以兩年的時間完成【MPEG-4/7 保證服務品質的媒體內容代理伺服器】的設計。在本年度完成的項目有：

暫存決策器(Caching Policyer)

多媒體資料所需的檔案空間很大，而代理伺服器(Proxy Server)的硬碟空間有限，所以必須選擇性的暫存媒體資料。我們將面臨一個問題：到底要暫存哪些媒體呢? 我們將依據媒體內容的熱門程度、使用時間、所需資源及媒體內容間的相關性，來設計暫存決策演算法。演算法將參考媒體內容被點選的次數，擁有較大的點選次數的媒體資料將優先暫存：以此概念決定該儲存哪些資料。此功能模組並無確切特定的演算法可供參考,所以值得我們花多點心力去研究。

此外，因為代理伺服器被建構成對等式系統(Peer-to-Peer system)的架構，所以每一台代理伺服器需決定本身所要暫存的媒體，並與其他的代理伺服器相



圖六：代理伺服器之間的協同工作

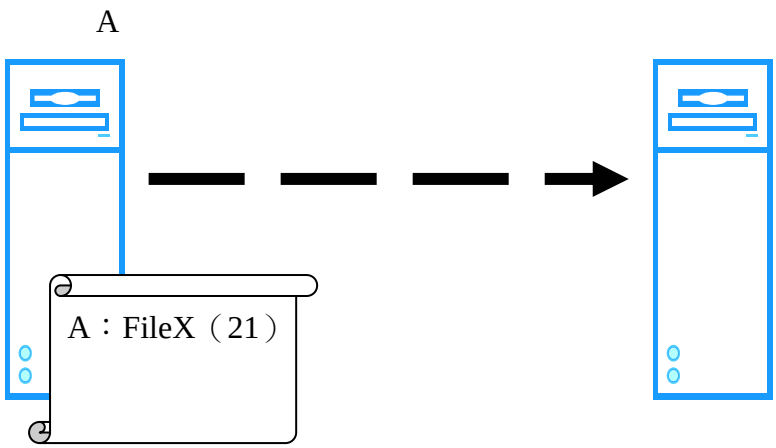
互合作，儲存不同的媒體，以有效利用儲存空間。暫存決策器便負責執行這部分的協調與計算。圖六展示了多台代理伺服器之間的協同工作情形。

正如圖六所展示的，我們讓一定數目的代理伺服器（*A*、*B*、*C*）組成一個互相工作的群組。為了讓伺服器之間能夠彼此協力工作，在群組當中每一台伺服器會定期交換本身所儲存的媒體資訊。該資訊除了包含媒體種類、名稱外，還包含每一個媒體被點選的次數，伺服器整體負載（load）等等。藉由交換這些資料，每一台伺服器上即能夠保持一份在該伺服器與其伙伴中，所儲存各樣媒體的列表。以圖六的伺服器*A*為例，伺服器*A*將會接收到來自伺服器*B*與伺服器*C*的媒體種類訊息；藉由此訊息，伺服器*A*將可以更新儲存在其中的伙伴資訊表，以保持最新的系統狀態供暫存決策器參考使用。圖六的系統狀態即為：檔案一（點選次數：三）、二（點選次數：5）存放在伺服器*A*中；檔案三（點選次數：一）存放在伺服器*B*中：檔案四（點選次數：二）存放伺服器*C*中

在每一台代理伺服器上，暫存決策器將根據群組中所有的伙伴資訊，以下列參數為基礎，決定代理該伺服器應該暫存的媒體種類與數目：

1. 媒體的歡迎度（popularity）：

   我們以媒體的被點選次數當作媒體的「歡迎度」。代理伺服器皆會監視儲存在其中的媒體，並記錄每一個媒體受到使用者要求的次數，此次數即所謂的「歡迎度」（popularity）。概念上，較受歡迎的媒體應該有較多的伺服器暫存，以減輕單一伺服器的負擔；有鑑於此，當媒體的歡迎度達到一定的門檻（popular threshold）時，暫存決策器將會發出要求，將該媒體複製至其他的代理伺服器上，使得更多的使用者能就近取得該媒體。



圖七：暫存決策器所觸發的檔案轉移，其中歡迎門檻為 20

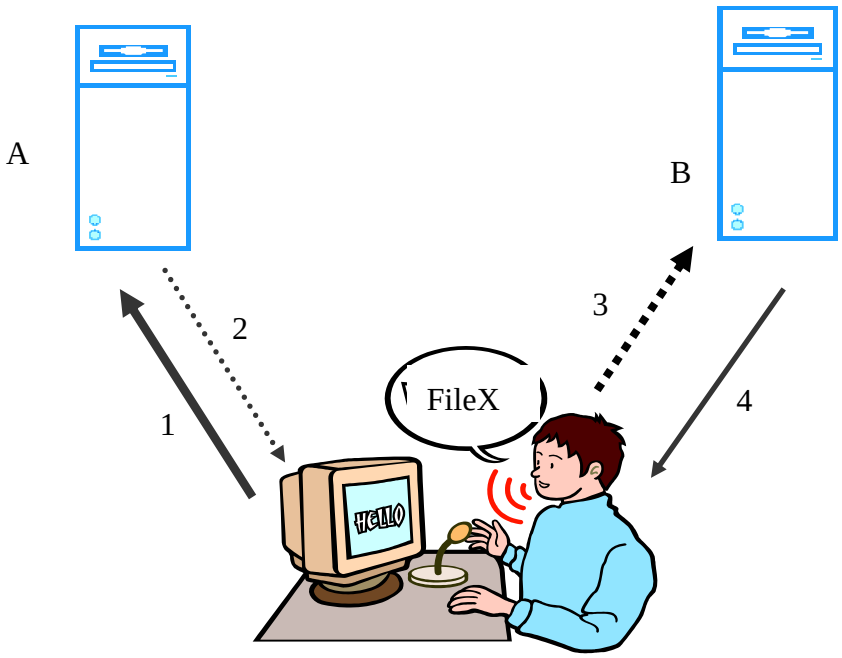
以圖七為例，當代理伺服器*A*的的媒體檔案*X*被點選的次數等於21時，(假設歡迎門檻為20)，暫存決策器便會觸發檔案轉移動作，將媒體*X*複製至其他的代理伺服器。

2. 伺服器的負載（load）：

當有一份媒體需要被代理伺服器儲存，暫存決策器將會根據群組中各個伺服器的負載狀況，選擇負擔最輕的機器儲存。如此一來，每一台代理伺服器所受到的服務要求數目將不會相差太多。這樣做的理由是為了達到負載平衡。此外，當代理伺服器中的儲存空間接近臨界值時，該伺服器將會根據各媒體的歡迎度，將不常被使用者要求的媒體從中刪除，以釋放儲存空間。

轉向器(Redirector)

網路上常常會發生同一份資料被複製多份，分別散置在不同的伺服器上。這將導致客戶端所選擇的伺服器，可能不是離他最近或傳輸品質最好的伺服器。本計畫所設計的代理伺服器(Proxy Server)將擔任起為客戶端選取最恰當伺服器的角色。因MPEG-7對於不同的媒體內容(Content)給定全世界唯一的代碼，代理伺服器(Proxy Server)可針對個別的媒體內容(Content)紀錄傳輸品質較好的數個伺服器，如此，代理伺服器(Proxy Server)將可很容易為客戶端找到最佳的伺服器。



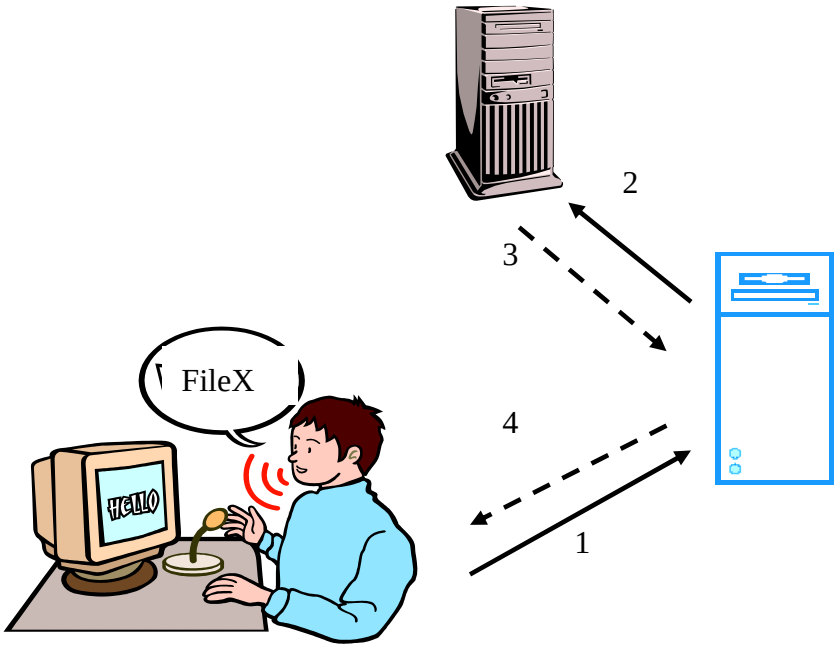
圖八：代理伺服器之間的轉向（redirection）

當代理伺服器收到使用者的要求時，將會檢查儲存在其中的伙伴列表，在當中選擇儲存有該媒體的伙伴（即代理伺服器）。若鄰近的代理伺服器可以提供服務，此結果將被告知使用者端，此即為「轉向」（redirection）。使用者藉此可向新的代理伺服器要求服務，讀取該媒體資料。

圖八為轉向功能的簡單示意圖。圖中使用者起初向最近的代理伺服器A發出媒體檔案*X*的要求（箭號1）。伺服器*A*在收到要求後，發現本身並未暫存檔案*X*，於是檢查伙伴列表，發現媒體檔案*X*暫存在代理伺服器B中。於是代理伺服器*A*向使用者發出轉向通知（箭號2），告知使用者應向代理伺服器*B*索取資料。使用者在接收到此訊息後，將自動的向代理伺服器B發出要求（箭號3），資料變由代理伺服器*B*傳至使用者（箭號4）。

另一方面，當使用者向代理伺服器要求某媒體檔案，若原代理伺服器及所有鄰近的代理伺服器都無法提供此服務時，則透過暫存決策器，決定負責暫存該媒體的代理伺服器。此代理伺服器運用即時轉傳功能，向原媒體來源伺服器（source）發出服務需求，代理伺服器從來源伺服器收到影片串流後，即時轉傳給使用者。此即為所謂「線上暫存」（on-line cache）或是「即時轉傳」：一方面向媒體來源（source）要求媒體資料並暫存，一方面將所暫存的資料以串流方式服務使用者。圖九即為代理伺服器中，即時轉傳功能的示意圖。

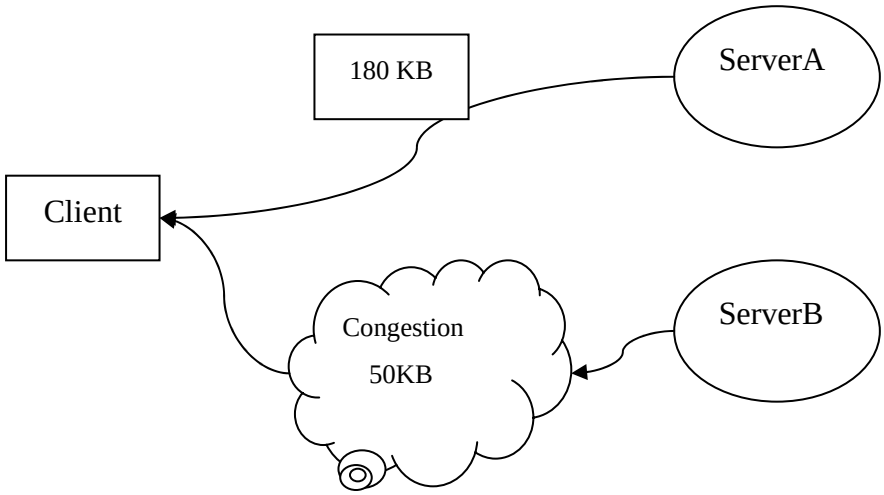
當使用者向代理伺服器發出媒體檔案*X*的要求（箭號1），而伺服器卻無法從伙伴列表中找出適當的伙伴。因此，為了提供使用者服務，代理伺服器將直接



圖九：代理伺服器的即時轉傳

向來源伺服器索取資料（箭號2），而來源伺服器將循一般服務流程，將資料送往代理伺服器（箭號3）。當代理伺服器收到媒體資料的片段，便回直接再轉傳至使用者端。（箭號4）

多源傳輸(Multi-Source transmission)

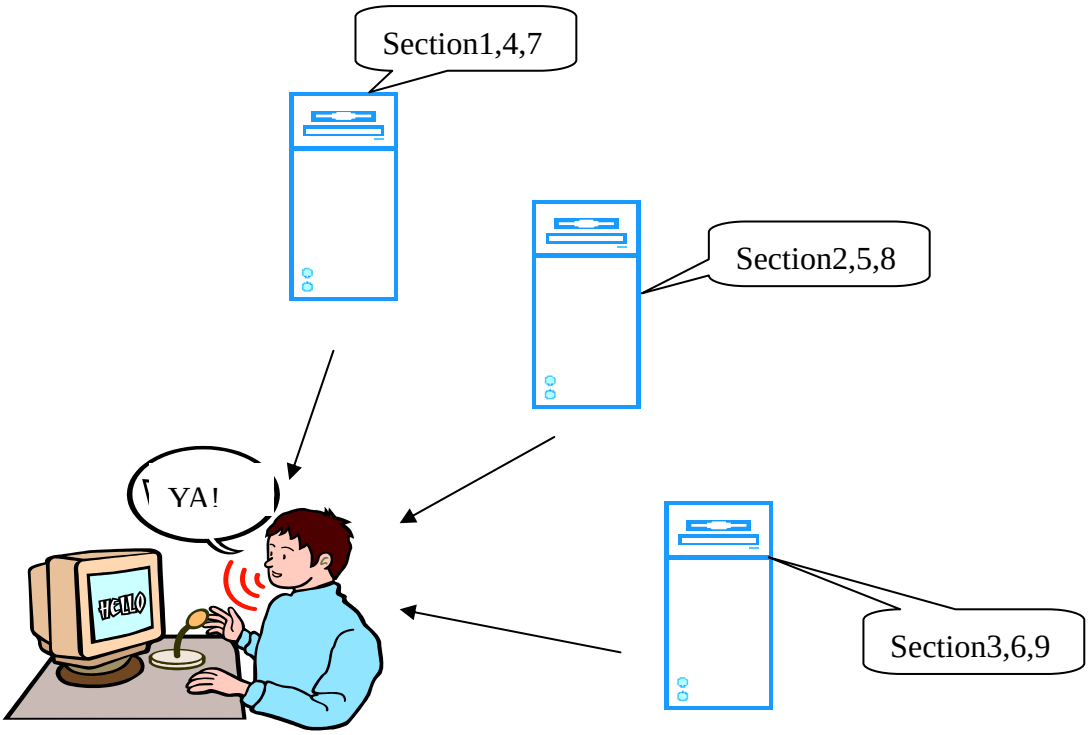


當客戶端發出要求時，可能會發現不只一部主機擁有該媒體（音樂、電影）。若客戶端頻寬夠大可容納多條串流，則透過排程器，使客戶端能同時得到多台代理伺服器的服務，享受較大頻寬的服務品質。此外，由於無線網路的連線品質會時時隨著使用者跟網路橋接器間的距離、阻隔物、周圍使用者的多寡而改變，無線網路的頻寬和品質常常無法負荷串流影片所需的要求；如果一位使用者可以同時連接到多台網路橋接器，讓串流影片所需要的頻寬平均分攤，使用者所收到的效果將會比只跟一台網路橋接器相連來的好。

假設我們把影片切成三部份，每一個部份分別從不同的網路橋接器連到相同或不同的串流伺服器上，分均在無線網路上的每一條串流所花的頻寬就只會是原本的三分之一；相反地，原本必須要在三秒內傳完三秒鐘的資料量，在每條串流上只需在三秒內傳完一秒鐘的資料量，使用者就可以順利播出所要觀看的串流影片了。

為了完成多源傳輸的目標，並且不額外增加代理伺服器與網路的負擔，各個伺服器的資料傳送必須是順序交錯。以圖十為例，使用者使用了三台代理伺服器進行多源傳輸。而此三台所傳送的資料區端分別錯開，而以交錯的順序傳送。如此一來，雖是多台伺服器同時服務一位使用者，但是伺服器系統與網路

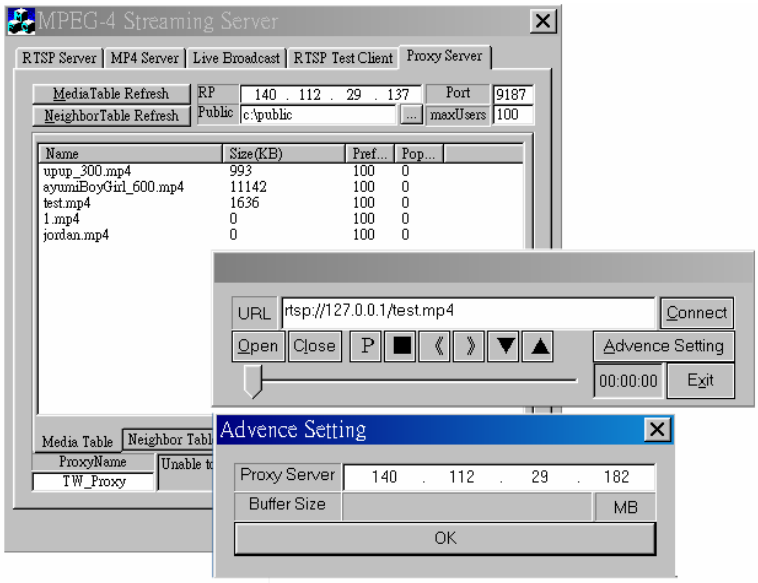
的整體負載卻未增加（因為並未傳送多餘重複的資料區段），另一方面，使用者



圖十：多源傳輸的交錯傳送

卻能得到較大的頻寬享受。

圖八為目前代理伺服器與客戶端之使用介面。



圖八、代理伺服器與客戶端之使用介面

四、 總結本計畫第三年的成果摘要

預期工作項目	實際工作成果	說明
完成頻寬調節傳輸模組中，執行層次刪除(Layer-dropping)與框架刪除(Frame-dropping)的相關元件。	完成速率調節（Rate shaping）模組中加強了SDF(Selective Frame Discard）功能	符合
將FEC code加入頻寬調節傳輸模組中，並增加交錯傳送的功能。	已完成FEC模組，交錯傳送功能含至多源傳輸模組中	符合
代理伺服器中，挑選最佳伺服器的轉向功能	已完成	符合
代理伺服器中，暫存及取代的決策機制	已完成	符合

五、 結論

本分項計畫『MPEG-4/7 保證服務品質之媒體內容傳輸技術』第三年計畫順利完成了預定目標，包含在【MPEG-4/7 保證服務品質的媒體內容代理伺服器】的技術中，完成暫存決策器、轉向功能與多源傳輸模組。在【RTP/RTCP 傳輸模組】方面，在 rate-shaping 技術上，加上了 SFD 模組，以提高使用者對網路的使用品質。此三年所發展各種技術，皆用以適應不同的網路環境，迄今可為圓滿達成。

六、 參考文獻

[1] Peer-to-peer multimedia streaming and caching service, Jeon, W.J.; Nahrstedt, K., Multimedia and Expo, 2002.Proceedings. 2002 IEEE International Conference on, Volume: 2, 26-29 Aug. 2002

[2] A framework approach to accommodation of multimedia communications for training systems, Shengru Tu, Liang Xu, and Ying Wu, Multimedia Software Engineering, 2002. Proceedings. Fourth International Symposium on, 11-13 Dec. 2002, Page(s): 232 -239

[3] TCP-like flow control algorithm for real-time applications, Seung-Gu Na, Jong-Suk Ahn, Networks, 2000. (ICON2000). Proceedings. IEEE International Conference on, 5-8 Sept. 2000, Page(s): 99 -104

[4] Scheduling real-time traffic in IP-based cellular networks, Lee, K.S. and Zarki, M.E., Personal, Indoor and Mobile Radio Communications, 2000. PIMRC 2000. The 11th IEEE International Symposium on, Volume: 2 , 18-21 Sept. 2000, Page(s): 1202 -1206 vol.2.

七、 相關論文成果

[1] Eric Wu, Hsu-Te Lai, Meng-Feng Tsai, Cheng-Fu Chou, "Low Latency and Efficient Packet Scheduling for Streaming Applications", In the Proceeding of IEEE International Communications Conference, Jun. 2004.

[2] Ching-Ju Lin, Cheng-Fu Chou, "Hybrid WLAN for Data Dissemination Applications", To appear In the Porceeding of IEEE/ACM International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems, Oct. 2004,

[3] L. Cheung, C.F. Chou, L. Golubchik, Y. Yang, "A Fault Tolerance Protocol For Uploads: Design and Evaluation", To appear in the Proceeding of International Symposium on Parallel and Distributed Processing and Applications, Dec. 2004.

Low Latency and Efficient Packet Scheduling for Streaming Applications

Eirc Hsiao-Kuang Wu*, Hsu-Te Lai*, Meng-feng Tsai *, Cheng-Fu Chou†

*Department of Computer Science and Information Engineering
National Central University, Chung-Li, Taiwan, R.O.C.
†Department of Computer Science and Information Engineering
National Taiwan University, Taipei, Taiwan, R.O.C.
hsiao@csie.ncu.edu.tw

Abstract— **Adequate bandwidth allocations and strict delay requirements are critical for real time applications. Packet scheduling algorithms like Class Based Queue (CBQ), Nested Deficit Round Robin (Nested-DRR) are designed to ensure the bandwidth reservation function. However, they might cause unsteady packet latencies and introduce extra application handling overhead, such as allocating a large buffer for playing the media stream. High and unstable latency of packets might jeopardize the corresponding Quality of Service since real-time applications prefer low playback latency. Existing scheduling algorithms which keep latency of packets stable require knowing the details of individual flows. GPS (General Processor Sharing)-like algorithms does not consider the real behavior of a stream. A real stream is not perfectly smooth after forwarded by routers. GPS-like algorithms will introduce extra delay on the stream which is not perfectly smooth. This thesis presents an algorithm which provides low latency and efficient packet scheduling service for streaming applications called LLEPS.**

I. Introduction

The number of new real-time applications is growing dramatically as more and more broadband services are available. The current best-effort service for Internet delivery of data cannot be sufficient for new multimedia applications. A variety of new multimedia applications such as voice over IP, video on demand and teleconferencing rely on network traffic scheduling algorithms in switches and routers to assure performance bounds to satisfy different QoS requirements. Typical VoIP (Voice over IP) applications require low latency and less bandwidth; VoD (Video on Demand) applications require a huge volume of bandwidth, but they can tolerate higher latency than VoIP applications. Unlike bursting data traffic applications, real-time applications usually do not try to consume all available bandwidth. In most cases, the required maximum bandwidth for a real-time application can be expected, such as a typical streaming application.

Since most of existing routers only provide best-effort services, the traffic of streaming applications will compete with other traffic of data services like FTP, HTTP and so on. The packets of streaming applications may experience higher latency or be dropped because of the competition among different kinds of traffic. Therefore, the quality of streaming application can not be guaranteed and new technologies are necessary to offer quality of service (QoS) for these real time applications.

A number of research solutions propose to enforce the fairness among Internet flows. The proposed fairness schemes makes that each flow consumes almost the similar bandwidth. They prevent the competition between flows and assure no starvation of any flow. The above researches assume that all flows have end to end congestion control mechanisms to adapt to the same compromised bandwidth assignments. The real-time application can only survive with enough bandwidth and different real-time applications may demand different bandwidths. Therefore, enforcing each flow getting the same bandwidth is not applicable.

Differentiated services (DiffServ) [1][2] and integrated services (IntServ) [3] are proposed to make the network to be QoS aware in IP networks in the mid 1990s. RSVP [4] is proposed for reservation for the IntServ approach and it defines how to allocate bandwidth hop by hop. Bandwidth allocation makes that each flow can get different bandwidth. Real-time applications can allocate bandwidth as they need, then they don't need to compete with other flows anymore.

Bandwidth reservation can ensure the delay of packets of real-time applications bounded into some value. Class Based Queue (CBQ) [5] has proposed a link sharing scheme which can be extended for bandwidth reservation implementation. However, CBQ does not consider the latency of packets and its service could be bursty for delay-bounded service constraints.

In addition to bandwidth guarantee, the latency of packets is required to be low and stable for some particular real-time applications. Bandwidth guarantee ensures no packets of streaming applications being dropped. Stable and low latency makes the playback latency of streaming-applications low. Therefore, *bandwidth guarantee* and *stable* and *low latency* are the two major requirements for stream applications. The bursting service providing by CBQ can not keep the latency of packets stable. Besides CBQ, other packet scheduling algorithms providing bandwidth reservation have been developed. They can be separated into three categories: GPS (Generalized Processor Sharing) like, frame based and regulative.

GPS is a concept of how multiple tasks (sessions) sharing a single processor in an OS. The sessions can get different

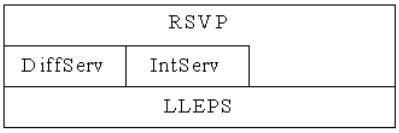


Fig. 1. DiffServ, IntServ, RSVP and LLEPS

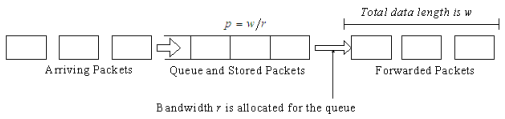


Fig. 2. Parameters of a queue

service rates at the same time, and they are served simultaneously. For applying GPS to the network technologies, it should be not noted that the service of GPS is bit by bit. However, the network service is packet by packet. Due to the different service units of GPS (bit-based) and Internet (packet-based), the algorithms WFQ [6] and WF^2Q [7] have been developed. Since they are all simulating GPS, they are classified as GPS like algorithms. The GPS like algorithms can extend the bandwidth sharing into bandwidth reservation and assure the packet behaviors smooth. In other words, GPS like packet scheduling algorithms mostly consider the latency of packets and bandwidth sharing ratio.

Frame based packet scheduling algorithms like DRR [8], NDRR [9] are different from GPS. They adopt different approaches to provide bandwidth sharing (reservation) services. They service all queues round by round and have no complex control mechanisms. The major advantage of frame based packet scheduling algorithms is efficiency. The disadvantage is that they only consider the queues which are always backlogged. That makes the latency of packets of the queues not always backlogged unstable.

The algorithms [10][11] are regulative packet scheduling algorithms. In regulative packet scheduling algorithms, each queue has a regulator which controls forwarding of packets. They can keep the latency of packets stable and the provide bandwidth reservation, but they are not scalable and hard to implement. Since the regulator demands to know the details of the flow, each regulator is only designed or configured for a particular flow. The regulative packet scheduling algorithm suffers the same problem of Integrated Service Framework [3].

II. Overview of LLEEPS

LLEPS is a packet scheduling algorithm. It can be the foundation of DiffServ, IntServ and RSVP. The relationship of them is represented in Figure 1. LLEPS can provide underlining packet scheduling service for them.

Whenever the bandwidth of bottleneck is not overwhelmed for on-going distinct multimedia applications, the consequent queuing delays of packets will be small and stable. Since the propagation time and transmission time of a packet on a link are constant values, the major cause of jitter is the queuing delay. In other words, as the queuing delay is small, the jitter will also be small. Streaming applications (ex: Mpeg Video Streaming) are not like bursty data oriented applications, such as TCP. They will not try to consume the bandwidth as much as possible. They usually consume the bandwidth as much as they need. LLEPS assumes streaming applications will not utilize more bandwidth than they need. The latency of packets can be bounded into a L/R value if the flow never uses more bandwidth than reserved bandwidth. L denotes the size of the packet and R denotes the reserved bandwidth.

LLEPS maintains multiple queues and each queue is used only for a session. Upon receiving a packet of a session, LLEPS puts it into a specific queue. The scheduler chooses the queue with the highest priority and forwards packets for the queue.

$[t_s,t_f]$ denotes a time interval from time t_s to t_f. The reserved bandwidth for the session *i* is r_i. The bandwidth of the link is B. The link can forward W bytes in the time interval $[t_s,t_f]$. The number of sessions in the system is n. The data length consumed by the session *i* from t_s to t is w_i^t. Under the condition that the inequality (3) holds, LLEPS can work. If the session i does not consume more bandwidth than r_i, the inequality (4) always holds.

$$W = B \times (t_f - t_s) \tag{1}$$

$$w_i = r_i \times (t_f - t_s) \tag{2}$$

$$\sum_{i=1}^{n} w_i \leq W \tag{3}$$

$$w_i^t \leq w_i \tag{4}$$

LLEPS gives the session *i* a higher priority when equation (4) is true. It denotes that LLEPS can guarantee bandwidth for each session even when other sessions consume bandwidth more than the bandwidth of the link.

$$w_i^t \leq r_i \times (t - t_s) \tag{5}$$

$$p_i = w_i^t / r_i \tag{6}$$

If the session *i* temporarily consumes more bandwidth than the reservation, the inequality (5) will not be true. LLEPS determines the priority of each session from this concept. Let p_i be the priority of the session *i*. The equation (6) shows the way the value of p_i is determined. The lower p_i denotes the higher priority in LLEPS. Figure 2 shows the relationship of *p, w, r* and the queue.

The value of p_i can be an indicator to reflect the greedy status for a particular session. At any moment, p_i of each queue should be similar. As a session is more aggressive than others, its p_i value will be larger than others'. According to the property of p_i, LLEPS only needs to forward packets of

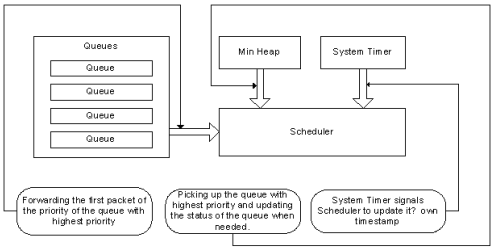


Fig. 3. Components of LLEPS

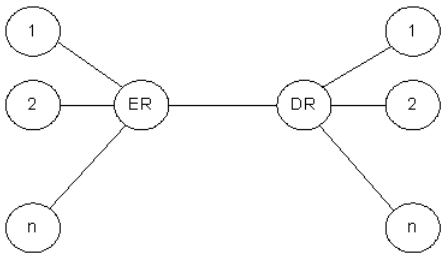


Fig. 4. Topology of Simulation

the queue with a lower p_i value than others. The behavior of LLEPS is able to keep the p_i value of each queue similar.

LLEPS provides the packet forwarding service based on the concept of time slice. LLEPS defines a time interval [t_s,t_f] and tries to reserve bandwidth for each session. It can guarantee the bandwidth in the current time interval, the next time interval and so on. Let *T* be the length of [t_s,t_f] and LLEPS must update the usage status of each queue every *T* seconds. It ensures that each session can get the bandwidth reserved for it in each time slice.

A. Components of LLEEPS

LLEPS contains these components: a system timer, a scheduler, queues and a min heap. Figure 3 introduces the basic architecture of LLEPS and the relationships between the three components.

1) Queues:

There are numbers of queues in the system. The relationship between queues and sessions is one to one. A session may contain multiple flows. When a packet arrives, LLEPS classifies it by the IP address or the port number of UDP or TCP and puts it into the corresponding queue. The way to classify packets could be flexible for distinct requirements.

There are several properties for a queue: Q, w and ts. Q value is the data length LLEPS will reserve for the session (queue) in a time slice. The value of w is the number of bytes that the queue has been served in the time slice. The value of ts is the time stamp of a queue and it denotes the time that usage status of the queue has been updated. These properties can be utilized to reflect the usage status of the queue.

LLEPS uses the value of Q and the value of w to determine the priority of the queue and uses ts to determine that the usage status of the queue should be updated or not.

2) Min Heap:

Min Heap is used to store the indexes of queues. It makes the scheduler able to pick the queue with the highest priority quickly.

3) System Timer:

System Timer is used to update the time stamp of the scheduler. After the scheduler picks a queue, it learns that the usage status of the queue should be updated accordingly as the time stamp of the queue is older than the time stamp of the scheduler.

System Timer updates the time stamp of the scheduler periodically. The period is defined as tick.

4) Scheduler:

The scheduler picks the queue with the highest priority to serve. After forwarding a packet from the queue, the scheduler will update the usage status w of the queue and put it into the Min Heap. The scheduler will update the ts value of the queue if the time stamp of the queue is smaller than the time stamp of the scheduler before putting the queue into the Min Heap.

There is a trick between ts and the Min Heap. While comparing the priority values of two queues, the queue with a smaller ts value will get a higher priority. This ensures that the scheduler can pick the queue which needs to be updated the w and ts values without any packet forwarding.

III. Simulation Experiments

A. Simulation Environment

The simulations are implemented using Network Simulator 2. The topology used in the simulation is presented by Figure 23. The circle with a number is a node and it denotes a sender or a receiver. The number *n* on it denotes that it's *n*th pair of a sender and a receiver. The circle with *ER* or *DR* is a router. The bottleneck of the simulations is the link between the two routers in Figure 4. The bandwidth of the links between nodes and routers is 100Mbps and the propagation delay is 1 millisecond. The link between two routers is 1Mbps and 15 milliseconds. LLEPS and Nested-Deficit Round Robin are applied on the simplex link from *ER* to *DR*.

B. Bandwidth Sharing

The bandwidth sharing experiment is to show the ability of LLEPS to provide bandwidth sharing. In the bandwidth sharing experiment, the same scenario is applied for LLEPS and NDRR. Three UDP flows are created: 500Kbps for UDP 1, 300Kbps for UDP 2 and 200Kbps for UDP 3. Each UDP flow runs with 10Mbps. UDP 1 starts at 0th second, stops at 15th second and starts at 25th second again. UDP 2 starts at 5th second. UDP 3 starts at 10th second. In NDRR, Q_min is set to 1500 bytes. The packet size of the test is set to 500 bytes.

The receiving rate of the three UDP flows should be 5:3:2 at any time in the idea case. Figure 5 shows the result of LLEPS and Figure 6 shows the result of NDRR. The ratio of

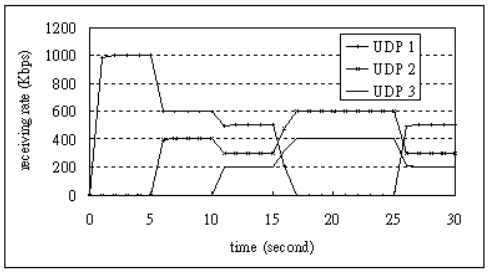


Fig. 5. Bandwidth Sharing Test/LLEPS

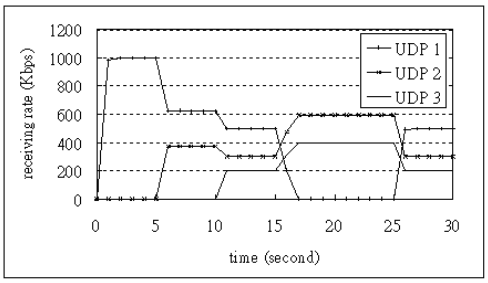


Fig. 6. Bandwidth Sharing Test/NDRR

reserved bandwidth for the three UDP flows is 5:3:2. Figure 6 demonstrates that NDRR always keeps the sharing ratio of all flows the same as the ratio of reserved bandwidth. However, LLEPS does not always keep the sharing ratio but it ensures that all flows can get reserved bandwidth.

In Figure 5 from 5th second to 10th second, only UDP flow 1 and UDP flow 2 are running. The sharing ratio between the two UDP flows should be 5:3. But the real sharing ratio in LLEPS is 3:2. In the experiment, after LLEPS updating the queue for UDP flow 1 and UDP flow 2, the value w is always equal to the value Q because the two UDP flows using more bandwidth than the reserved bandwidth all the time. The updating of all queues happens at the same time and UDP flow 2 may sent one or two more packets before the updating. The unfairness is never memorized by LLEPS. That's the reason why the sharing ratio between the two flows is not 5:3.

C. Bandwidth Reservation

The goal of the experiment is to test the ability of LLEPS to provide bandwidth guarantee. LLEPS allocates 200Kbps for UDP flow 1 and 300Kbps for UDP flow 2. The bit rate of UDP flow 1 and UDP flow 2 are 200Kbps and 300Kbps. UDP flow 1 starts at 5th second and UDP flow 2 starts at 10th second. There are four background TCP flows competing with the two UDP flows.

Figure 7 presents the result of the experiment. Before the starting of the two UDP flows, four TCPs take all the bandwidth of the bottleneck. After the first UDP flow starts, it gets the bandwidth reserved to it. The second UDP flow also gets the reserved bandwidth. Since LLEPS allocates bandwidth

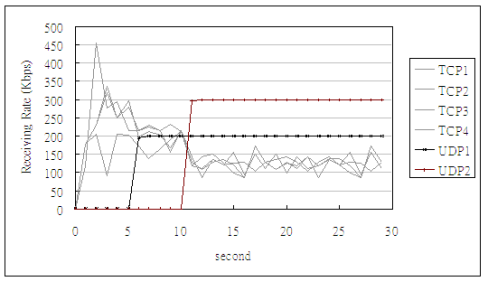


Fig. 7. Bandwidth Reservation Test of LLEPS

for the two UDP flows, the bandwidth for the two flows will not be used by the four TCPs. The four TCPs share the left bandwidth 500Kbps. The result shows that LLEPS is able to provide bandwidth guarantee.

D. Buffer Under Run Problem

As described, NDRR and other frame-based packet scheduling algorithms do not perform well when some queues are not always active. The delay of packets is not stable.

In the simulation, ten flows are running and both LLEPS and NDRR allocate 100Kbps for each of the ten flows. The bit rate of the first nine UDP flows is 1Mbps and the bit rate of the tenth UDP flow is only 80Kbps. The queue belonging to the tenth UDP flow will be empty at most of time and other queues will be always active. The first 9 UDP flows are background traffic and keep the bottleneck link busy.

The 10th UDP flow is used to simulate a streaming application. Most streaming applications like VoD and VoIP do not have special control mechanism to control the behavior (bit rate or something else) of the stream. A CBR (Constant Bit Rate) MP3 (MPEG Audio Layer 3) broadcasting is just like a CBR UDP flow. This is why an UDP flow is able to simulate a streaming application here.

This paper uses *Delay Jitter* to represent the stability of latency of packets. The definition of *Delay Jitter* is the same as RTP. R_i is the receiving time of packet *i*. S_i is the sending time of packet i. D_{i-1}, i denotes the delay variation between two consecutive packets. Ji is the smoothed function of D_{i-1}, i.

$$D_{i-1,i} = (R_i - R_{i-1}) - (S_i - S_{i-1}) \tag{7}$$

$$J_i = 15/16 \times J_{i-1} + 1/16 \times |D_{i-1,i}| \tag{8}$$

In the experiment, only the J_i of the 10th UDP flow is plotted. The packet size is set to 100 bytes, 500 bytes and 1400 bytes. Qmin in NDRR is set to 1500 bytes. In LLEPS, *tick* is set to 250 milliseconds.

Figure 8, Figure 9 and Figure 10 are the results of the simulations. The solid line is the result of LLEPS and the dashed line is the result of NDRR. In these simulations, no packet of the 10th UDP flow is dropped. In Figure 8, the transmission time of a packet on the bottleneck link is 0.8 milliseconds. In Figure 9, it's 4 milliseconds. In Figure 10,

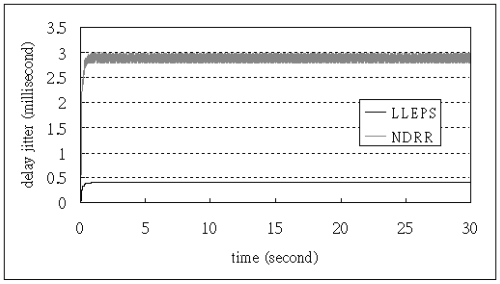


Fig. 8. Delay Jitter with 100 bytes packet size

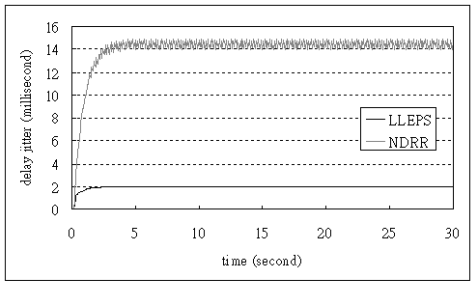


Fig. 9. Delay Jitter with 500 bytes packet size

it's 11.2 milliseconds. If the transmission time is longer, the delay jitter will also become longer. In LLEPS, the delay jitter is always approximate to 50 percent of the transmission time of a packet. The experiment shows that LLEPS can handle the problem much better than NDRR.

IV. Conclusion And Future Work

This paper introduces a new scheduling protocol (LLEPS) to provide bandwidth guarantee service efficiently. In addition to ensuring bandwidth guarantee services, LLEPS does not introduce much delay jitter effects for packets. NDRR does not address *"Buffer Under Run Problem"* but LLEPS does. Since LLEPS always serves the queue with the highest priority, it provides preemption mechanism for packets. However, in NDRR, the service sequence of each queue is constant. The continuous work of LLEPS is to design a better way to determine the priority of each session. LLEPS determines the priority of each session based on the corresponding transmission rate. LLEPS estimates the transmission rate of a session in a time interval. When the time interval is too small, the estimation will become very unstable and no meaning for LLEPS. The continuous efforts will address the issue.

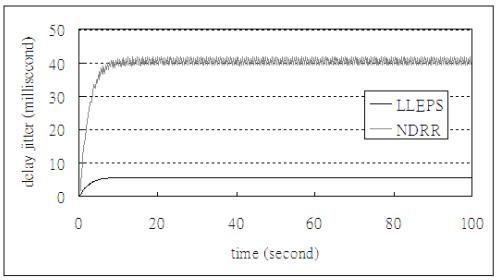


Fig. 10. Delay Jitter with 1400 bytes packet size

References

[1] S. Blake, D. Black, M.Carlson, E. Davies, Z. Wang and W. Weiss, "An architecture for differentiated services", RFC 2475, December 1998.
[2] K. Nichols, V. Jacobson and L. Zhang, "Two-bit differentiated services architecture for the Internet", IETF RFC 2638, July 1999.
[3] R. Braden and D. Clark, "Integrated Services in the Internet Architecture: An Overview", RFC 1633, July 1994.
[4] R. Braden, L. Zhang, S. Berson, S. Herzog, and S. Jamin, "Resource ReSerVation protocol (RSVP) - Version 1 Functional Specification", RFC 2205, September 1997.
[5] S. Floyd and V. Jacobson, "Link-Sharing and resource management models for packet networks", IEEE/ACM Transactions on Networking, August 1995.
[6] A. Demers, S.Keshav and S.Shenker, "Design and Analysis of a fair queueing algorithm", Proceeding of ACM SIGCOMM, September 1989. algorithms", Proc. SIGCOMM'96, August 1996.
[7] Jon C.R. Bennett and Hui Zhang, "WF2Q: Worst-case Fair Weighted Fair Queueing", INFOCOM '96, Proceedings IEEE, Mar 1996
[8] M. Shreedhar and George Varghese, "Efficient fair queueing Using deficit round-robin", IEEE Transactions on Networking, Jane 1996.
[9] Salil S. Kanhere and Harish Sethu, "Fair, Efficient and Low-Latency Packet Scheduling using Nested Deficit Round Robin", Proceedings of the IEEE Workshop on High-Performance Switching and Routing (HSPR), May 2001
[10] H. Zhang and D. Ferrari, "Rate-Controlled static-priority queueing", Proc. IEEE INFOCOM '93, September 1993
[11] S. Iatrou and I. Starvrakakis, "A Dynamic Regulation and Scheduling Scheme for Real-Time Traffic management", IEEE/ACM Transactions on Networking, February 2000.

Hybrid WLAN for Data Dissemination Applications*

Ching-Ju Lin and Cheng-Fu Chou
Department of Computer Science and Information Engineering,
National Taiwan University, Taipei, Taiwan, ROC
{shauk, ccf}@cmlab.csie.ntu.edu.tw

Abstract

In traditional infrastructure-based IEEE 802.11 wireless LAN network, all the mobile nodes associated themselves with an access point (AP) share the same wireless channel with other mobile devices. In such wireless network environment, there are two major inefficiencies: (a) load imbalance among different APs: when the distribution of the location of mobile nodes is non-uniform, this might result in a situation where a lot of mobile nodes access a single AP and the rest of APs are idle, and (b) poor resource utilization: most of the mobile nodes contend for the wireless channels of APs while there are some unused and available wireless channels, which can be used to set up ad hoc groups. In this paper, we propose the hybrid WLAN (H-WLAN) framework to address these two problems. First, our H-WLAN framework provides the adaptive service selection module for the mobile node to figure out a better wireless channel and the server to get its service. Second, the popularity-aware channel management module is able to efficiently integrate the ad hoc mode with the infrastructure mode by allowing mobile nodes to set up a new ad hoc group and then share their resources with other nodes within that group. The simulation results show that our H-WLAN framework can substantially (a) achieve high utilization and load balance among wireless channels with lower control overhead, and (b) provide better quality of service for the mobile nodes.

1. Introduction

Wireless local area networks have come into great use in recent years. Although a variety of wireless network technologies have been in the market, wireless LANs based on the IEEE 802.11 standard are the most popular. There are two operation modes, i.e., the infrastructure mode and the ad hoc mode, specified in IEEE 802.11 standard. In the ad hoc mode, all mobile nodes are free to directly communicate with each other in a peer-to-peer manner. Another option of the operation mode in WLAN is the infrastructure mode, where each mobile node associates itself with an access point (AP) within its direct transmission range. The AP is connected to a wired network and provides a conduit to the external network for the mobile nodes. Communications can only take place between the wireless node and the AP but not directly between the wireless nodes.

In this paper, we consider the problem of distributing popular data at the server to users in a wireless LAN environment. The data dissemination problems correspond to a set of important applications. These applications include distribution of course slides in a class, dissemination of useful information in the info-station, online documents distribution in a conference, and many more. In a traditional infrastructure-based WLAN, all communications will go through the APs, but an AP can only provide a certain amount of bandwidth (e.g., 802.11b is 11 Mbps) for all users within its transmission coverage. In such wireless network environment, there are two major inefficiencies: (a) load imbalance problem among the APs: when the distribution of the location of the mobile nodes is non-uniform, this might lead to a situation where a lot of nodes access a single AP but other APs are idle, and (b) poor resource utilization: most of the nodes contend for the wireless channel of the AP while there are still some unused but available wireless channels, which can be used to set up ad hoc groups. When the number of users increases within an AP's cell, the contention for the wireless channel will lead to more collisions, degradation in system throughput, and longer data delivery time for the mobile nodes. To increase the system throughput and reduce the user's response time, a simple solution is to allocate more bandwidth for users by installing more APs, which use different wireless channels, but this approach is expensive and not flexible to deal with the "temporary" fluctuation of the network load.

In a wireless network environment, an important observation for the data dissemination applications is that most local users are interested in nearly the same set of objects.

* This work was partially supported by the National Science Council and the Ministry of Education of ROC under the contract No. NSC92-2622-E-002-002 and 89E-FA06-2-4-8.

In other words, since the requested objects of local users are similar, the mobile nodes can form a local ad hoc group to communicate with each other for sharing those objects without going through the APs until they have to switch back to the infrastructure mode. Furthermore, setting up the local ad hoc groups, which use different channels from the APs, can mitigate the contention and collisions in the infrastructure-based APs channels. Therefore, if we can efficiently integrate the ad hoc mode and the infrastructure mode, such hybrid WLAN is able to increase both the system and user's throughput, and provides a better and consistent service.

However, there are several important issues that need to be addressed in order to to design a high-performance hybrid WLAN framework. The first one is the service selection problem between different ad hoc groups and infrastructure-based networks. The second one is channel allocation problem including when to set up a new ad hoc group and which channel should be assigned to that group. An inappropriate service selection or channel allocation mechanism may lead to congestion in some channels or servers while others might be idle. In our H-WLAN framework, we use the *adaptive service selection module* for the mobile node to figure out a better channel-server selection to get its service, while achieving higher throughput for the mobile node and the load balance between the wireless channels. The *popularity-aware channel management module* is used to determine the proper wireless channel allocation to fast and correctly react the actual network condition. The simulation results show that our H-WLAN can result in significant performance improvements. That is, the H-WLAN can efficiently integrate the ad hoc and infrastructure modes and achieve load balance among wireless channels, better channel utilization, and higher throughput for both the system and the mobile nodes.

The remainder of the paper is organized as follows. In section 2, we review the related works on integrating the ad hoc wireless network with infrastructure-based WLAN. The H-WLAN framework is described in Section 3. Section 4 presents a performance study of our proposed framework. At last, the conclusion is given in Section 5.

2. Related Work

In general, most of works on ad hoc networks focus on providing routing technologies to improve the efficiency of content distribution applications. In [9] the authors propose two schemes, i.e., CacheData and CachePath, that the mobile nodes cache the popular data or data path for others in the same ad hoc network whereas our H-WLAN system aims to improve data access in the infrastructure-based WLAN environment with the help of multiple ad hoc groups.

On the other hand, some works based on infrastructure mode network aim to improve the performance of content dissemination and provide the QoS. In WICAT system [7], each mobile node has a list of interested objects, and these objects may be located at different infostations. Only when a mobile node moves close to an infostation, it is able to download those objects that are available at that infostation and match user's preference. In [8], their framework supports a map-on-the-go application, i.e., users move between infostations, and each infostation provides relevant maps to these users. These two systems focus on designing an efficient infrastructure-based wireless system to distribute useful objects and information for users; however, our framework is a hybrid wireless system integrating the ad hoc mode with infrastructure modes.

Other frameworks or systems, include [6], [1], [10], NUMI [5], and M^2-WLAN [2], use peer-to-peer communication to assist the infrastructure-based system for data exchange or distribution. [10] did an initial study on using non-cooperative mechanism, e.g., peer-to-peer communication, to improve the efficiency of content distribution for the infostation system. Two encountered mobile nodes check the file content lists of each other, identify some files they need, and exchange those files. In NUMI [5], the APs collaborate among themselves and can predict the mobile node's location in the near future. Using this information, the AP is able to select another mobile node, which is likely to move around the vicinity of the destination mobile node, to piggyback the requested data for the destination mobile node. However, how to predict the moving behavior of mobile nodes has the great effect on the performance of the NUMI system. In the M^2-WLAN framework [2], besides communicating with the APs, the mobile node can also switch to ad hoc mode to connect with another node under the administration of the AP, the central manager that assigns channels and keeps the information for all mobile nodes. Both NUMI and M^2-WLAN frameworks are for host-centric applications while our H-WLAN framework is for data dissemination applications. In addition, the peer-to-peer communication is a single-hop connection in NUMI and M^2-WLAN systems but can be a multi-hop connection in our H-WLAN framework, i.e., the transmission coverage of our framework can be extended by relaying data through multiple nodes.

Lastly, much work (as described in [3]) has been done in addressing server selection problems for services replicated across wide-area networks. We note that our work differs from them in (i) our adaptive service selection module can maintain a better tradeoff between the load balancing nodes across the service and locating a "closer" service to a node, and (ii) how to manage channel allocation in the H-WLAN framework has a great impact on the performance of the service selection scheme.

3. Hybrid WLAN Framework

In this section, we first give an introduction about how a mobile node can switch its connection between the infrastructure mode and ad hoc mode in the IEEE 802.11 standard and then present our H-WLAN framework.

In a MAC frame header, ToDS and FromDS bits represent the frame to be transmitted into a Distribution System (DS) or from a DS and indicate the meaning of the four address fields in the MAC frame. If both ToDS and FromDS bits are zero, this packet is exchanged between two wireless nodes without an AP involved. If a mobile node runs in the infrastructure-mode, the ToDS bit is set to one and FronDS is set to zero. This means that the mobile node sends a packet to another node via the AP. In this work we assume that a mobile node can change some parameters setting to decide which mode (or wireless channel) it would like to use. That is, a mobile node can switch between the infrastructure mode and ad hoc mode dynamically. Usually there are multiple wireless channels available in a wireless network. For instance, there are 14 channels defined in the DSSS PHY frequency channel plan in the 802.11 standard. Our goal is to make use of the available but unused wireless channels to form some ad hoc groups for assisting the data distribution to relieve the traffic load at the APs and to improve the performance for both the system and the users.

The main challenge of H-WLAN framework is how to efficiently integrate ad hoc mode into the traditional infrastructure-based WLAN environment. To address this challenge in the H-WLAN framework, we propose two major components (a) the adaptive service selection module, and (b) popularity-aware channel management module. After a mobile node joins the wireless network and wants to get the objects, it first uses the service selection module to decide it can get the requested objects from which server through which channel. In popularity-aware channel management module, the H-WLAN system is able to decide when and how to set up a new ad hoc group and to allocate which channel to this group. If a mobile node would like to share its objects with other nodes, the H-WLAN system determines if it is necessary to set up a new ad hoc group or just notifies this node to join the existing ad hoc group to share its resource using the same wireless channel. Next, we discuss each of them in detail accordingly.

3.1. Adaptive Service Selection Module

To choose a proper server, the service selection module consists of two phases: one is the resource discovery phase and the other is the server-channel selection phase. The resource discovery phase is that the system first provides the potential server list, i.e., a list of all accessible servers which can provide the requested object, to the request mobile node. After the mobile node acquires the potential server list, it can use that information and choose a proper server. For instance, when the mobile node in the infrastructure mode wants to get an object via the wireless network, it might contact with one of the APs within its direct transmission range and then obtain the potential server list from that AP. In order to provide such information, we use the APs as an information-managing center, and each AP maintains a group-information table, which contains three fields: the object id, the wireless channel and the available servers. Each tuple in the server list of object i, as shown in Table 1, contains two fields: the first field is the wireless channel used by these corresponding servers in the available server field, and the second field is the available servers, i.e., the nodes hold object i at that moment and in this ad hoc group. The discussions of the maintenance issues, such as how to set up, delete and update the group-information table, are presented in the following subsections.

object id	channel id	available server id
3	1	9, 8
	2	7
7	1	9

Table 1. Group Information Table.

channel id	group leader	status
1	9	on
2	7	on
3	null	off

Table 2. Channel Status Table.

3.1.1. Resource Discovery Phase: Next, we describe how a mobile node, according to its location, obtains the potential server list below:

1) Within the AP coverage: when the mobile node is within some APs' transmission range, there are two possible cases as follows. If this mobile node is in some ad hoc group, it first needs to switch back to the infrastructure mode without affecting ongoing communication. That is, the mobile node examines if it is involved in the process of any ongoing communications. If so, the mobile node has to wait for the end of those communications and then switch back to infrastructure mode. Otherwise, the mobile node is able to switch back to the infrastructure mode immediately. After the mobile node switches back to the infrastructure mode, the following procedures are the same as those in the infrastructure mode case. If this mobile node is in the infrastructure mode, it sends out the request message to one of those APs and asks for the potential server list.

2) Beyond the AP coverage: when the mobile node is beyond the transmission range of any AP, it randomly joins an ad hoc group within its direct transmission range and then sends out the request message to the leader of that ad hoc group. After receiving the request message, the group leader processes that message, generates the corresponding request message and switches to the infrastructure mode to send that request message to the AP for that request mobile node. After getting the source list, the group leader forwards the resource information back to the mobile node. Since the group leader is the relay node for those mobile nodes which cannot access any AP, one sufficient condition for the group leader is that it should locate within transmission range of some APs.

As the AP receives the request, it cannot tell if this request is from the intermediate node such as the group leader or not. The AP retrieves the potential server list from the group information table and returns that information to the client. The potential server list records all members that hold the requested objects in each different ad hoc group. In the traditional infrastructure-based 802.11 WLAN, if a mobile node is out of any APs' transmission range, it is unable to set up any wireless connection even if there are some mobile nodes within its transmission range. The intuition behind our work is the mobile node still has the chance to get its requested objects via its neighbor nodes in an ad hoc group even though it cannot connect to any AP directly. Therefore, the data delivery range in H-WLAN is extended via combining the ad hoc mode, and it also lets more users be able to get their requested objects.

3.1.2. Server-Channel Selection Phase: After a mobile node gets the potential server list from the AP, it is likely that there is more than one server in the list. In this subsection, we discuss how the mobile node to choose the suitable server to get its requested objects. Our goal is to maintain the tradeoff between the load balance and the proximity among all available wireless channels so that the system utilization and users' QoS can be improved. Five various methods are presented to select a proper channel-server pair, and the motivations and characteristics of each method as follows:

1) Signal-Strength Method (wlan): In this method, the mobile node connects to the AP with the best signal strength. This method is simple and widely used in current infrastructure-based WLAN. However, as the number of mobile nodes increases and the distribution of the mobile node's location is non-uniform, some APs might have heavy traffic load but the others might be still idle, i.e., it leads to the load imbalance problem. We include this signal-strength method and use it as a baseline approach to compare with.

2) Log-Based Method (wlan-nc): The log-based method is that each AP keeps track of the total number of current flows and includes this information in its broadcasting beacon message. A mobile node might receive several beacon messages from different APs, and can connect to the light-loaded server in order to avoid the congestion. The motivation for this method is to explore the benefit of such simple bookkeeping method.

3) Random-based Method (random): When there is more than one server in the potential server list, the mobile node randomly chooses a server from it and then checks if this server is reachable. If not, the mobile node continues to select another one until it can establish a connection with the chose server. The random-based method is straightforward and able to achieve load balance among the servers. However, the selected server may be far from the request mobile node, or the channel that they use may happen to be congested. We illustrate this effect in the performance evaluation section.

4) Ping-based Methods (ping, pr2, and r2p): To investigate more dynamic and adaptive source selection schemes, we consider the ping-based methods. The ping-based method is that the mobile node pings every source candidates in the potential list and then selects one server after receiving their response messages. The idea of such scheme is to use the response time of the small probing message to examine the condition of the transmission path between the mobile node and the server. In other words, if the mobile node receives the reply from the source quickly, it indicate that the quality of connection between the mobile node and the server is good enough. We also consider other two variations of ping-based methods — the ping-random2 (pr2) method and the random2-ping (r2p) method. The ping-random2 (pr2) method is similar to the ping-based method except that the mobile node has to choose one randomly from two servers that their ping response messages are most quickly returned. The random2-ping (r2p) method is that the mobile node randomly selects two servers from the potential server list and then only sends out the ping messages to these two servers rather than all servers in the list. In pr2 and r2p methods, the mobile node can not only select the proper server dynamically but free from connecting to the same server with other competing mobile nodes. Obviously, there is a number of other selection schemes that could be built or just reconstructed from the above ones. However, this set is reasonably representative for us to make comparisons.

3.2. Popularity-Aware Channel Management Module

In this section, we discuss some important issues in the channel management module including (i) when and how to create an ad hoc group, (ii) when and how to join an ex-

M : Total number of wireless channels;
O : Total number of objects;
p_i : The popularity of object i; $i = 1, 2, \ldots, O$
B_i : Available bandwidth in channel i; $i = 1, 2, \ldots, M$
$B_i = 1/M$;

if $(p_i >= B_j)$ **then**
    Assign channel j for object i;
    $p_i = p_i - B_j$;
    $B_j = 0$;
else if $(B_j > p_i > 0)$ **then**
    Assign channel j for object i;
    $B_j = B_j - p_i$;
    $p_i = 0$;
else
    reject;
end if

Figure 1. Channel management algorithm.

isting ad hoc group, and (iii) what should be done before a node leaves the ad hoc group.

3.2.1. Create an Ad Hoc Group: After the mobile node has got its requested objects and is willing to share those objects with other users, it can send the control message to the AP. After receiving this message, the AP determines if it is necessary to construct a new ad hoc group for those objects according to the channel management protocol. Since the popularity of accessed objects may be skew, an inappropriate channel assignment in the system may lead to poor resource utilization and system performance. We propose a popularity-aware channel management module and give its details as follows.

The main idea of the popularity-aware scheme is to allocate all available channels to each object in proportion to its popularity. A popular object should take advantage of more channels than a non-popular one to meet more access requests from mobile nodes. The detail about how to allocate channels is described in Figure 1. The bandwidth of all wireless channels is normalized to 1, and will be fully utilized in our popularity-aware management protocol. Moreover, each object will get the suitable amount of bandwidth based on its popularity. When an AP receives the request from certain node, which wants to share object i with the popularity p_i, it will check if p_i is greater than 0 first. If so, the AP randomly chooses a channel, called channel j, which there is still free bandwidth in and allocate it to object i. The amount of bandwidth that could be allocated to this object is max(p_i, B_j) (the max(a,b) function returns the larger one from two input values), and p_i and B_j will both be reduced by max(p_i,B_j) after getting available bandwidth. After updating the value of p_i and B_j, the object i could not get any more resource if its popularity p_i has been reduced to 0, and the channel j cannot be allocated to other objects if B_j has been reduced to 0.

3.2.2. Join an Ad Hoc Group: In H-WLAN framework, there are several conditions for a mobile node to join an existing ad hoc group. We describe them as follows:

1) Retrieving an object: The mobile node decides to get objects from another mobile node in some existing ad hoc group.

2) Sharing a wireless channel: The mobile node would like to share its objects but there are no more available wireless channels to build a new ad hoc group.

3) Out of range of APs: The mobile node cannot connect to any AP directly, i.e., there is no AP within its direct transmission range.

After deciding to join which ad hoc group, the mobile node has to send new state message including the group identification and the object list to the AP. The APs update their group information table and channel status table respectively after receiving this message.

3.2.3. Leave an Ad Hoc Group: After joining an ad hoc group, the mobile node might want to break the connection or get some objects that cannot be found in the current ad hoc group. Thus, the mobile node has to leave its ad hoc group to get the requested objects from other ad hoc groups or APs. In order to let the hybrid system work correctly, the mobile node should notify the AP to update all relevant information such as group information table and channel status table and take back the wireless channel before leaving. For instance, if some objects are only held on this leaving mobile node, the server has to delete this object field. If the group leader would like to leave, the APs have to appoint another group member as the new group leader before allowing it leaving or else this ad hoc group would be forced to be closed.

4. Performance Evaluation

We evaluate the performance of our H-WLAN framework by simulation, implemented in NS2 [4] with CMU wireless extension. In the simulation topology, there are 3 APs uniformly distributed in a $450m \times 400m$ area. The data rate of a wireless channel is $11Mbps$ and the transmission range of all mobile nodes, including the APs, is $250m$. The total number of objects in the simulation is 10. The object size is uniformly distributed between $100KB$ and $1MB$. Another important observation is that users usually request the objects at some specific periods, such as the period before a meeting or a class, and hence the inter-arrival time of request is getting smaller while the scheduled time of a presentation (or a class) approaches. To consider the above observation, we use a 2-phase arrival process with 2 different arrival rates in the simulations. For this 2-phase arrival process, it has a lower mean arrival rate λ_l for the first k arrivals and a higher mean arrival rate λ_h for the remaining arrivals. Moreover, let the random variables of x_l and

x_h represent the interarrival times in the first phase and second phase and they are exponential distributions with means $\lambda_l = 1$ and $\lambda_h = 5$ $(/sec)$ accordingly.

The performance metrics used in the remainder of this section are: (a) the mean response time to get a requested object, (b) the variance of mean response time, (c) the ratio of control overhead, which is exchanged for keeping some information such as potential server list, to total bandwidth consumption, and (d) the makespan, the time needed to complete the transfer of all requested objects. These metrics will reflect the quality-of-service characteristics that would be of general interest to the wireless LAN system and users.

4.1. Service Selection Study

First, we investigate the performance of different wireless network systems under Poisson arrival case and then Bulk arrival case. Each AP uses a different wireless channel and in this simulations there are 2 additional wireless channels, which are available for setting up 2 new ad hoc groups to share the objects locally. We use the pure infrastructure-based WLAN as the baseline to be compared with the performance of H-WLAN.

4.1.1. Poisson arrival case: The request arrival pattern is the Poisson arrival. Each user needs 5 objects sequentially and the order of requested objects is randomly selected from all possible 10 objects. There are total 30 users and each of them expects to get 5 different objects. We evaluate the performance of different selection schemes respectively as follows:

1) Hybrid mod vs. Infrastructure mode: From Figure 2 and 3, the simulation results show that the hybrid system, i.e., random, ping, pr2, or r2p, outperforms the traditional infrastructure-based system, i.e., wlan, in terms of mean response time and makespan time. Such performance improvement for users is due to the fact that the hybrid system is able to make use of available wireless channels to increase both the system's capacity and the user's throughput. As the makespan time is concerned, the r2p method also has smaller mankspan than other methods; namely, using the ping-based methods, the mobile node can finish its schedule earlier. In addition to the wlan method, we consider an enhanced method for improving the performance of pure infrastructure mode, i.e., wlan-nc, which each AP keeps track of the number of connections and the AP with lowest load is assigned to next user's request. The comparison, illustrated in Figure 2 and 3, shows that our hybrid system, such as ping, pr2, and r2p, can still perform better than wlan-nc because of taking advantage of additional available wireless channels.

2) Ping-based vs. Random: Another important and interesting question arises which selection scheme is better and

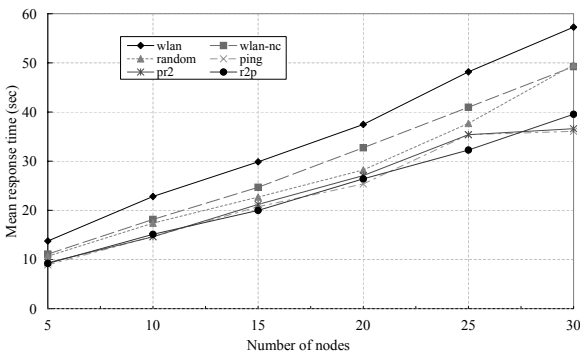


Figure 2. Poisson arrival: Mean response time.

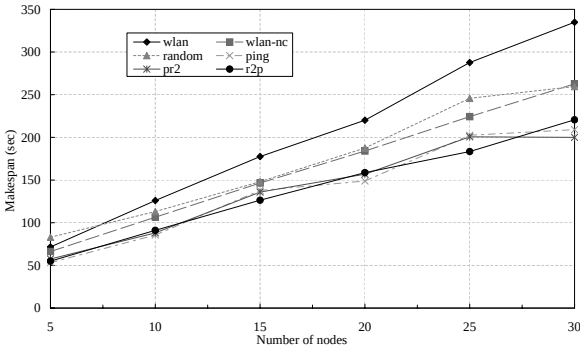


Figure 3. Poisson arrival: Makespan.

well-performing for the hybrid system. In Figure 2 and 3, we can see that the adaptive selection methods such as ping, pr2 and r2p have smaller mean response time or makespan time than the random method. This can be explained as follows. In all adaptive selection methods, the mobile node can choose a better server and channel to use according to the network condition at that moment; nevertheless, in the random method, it just blindly chooses one of feasible servers to connect. In other words, in the random method, it is more likely to choose the worst (highest-loaded) or farthest server to connect. In our simulations, we do not consider the scenario with other background traffic, but, however, in a more practical network environment, it is not easy for the random method to achieve the load balance if there exist different amount of background traffic running on the servers. On the contrary, the mobile nodes, in the ping-based methods, can choose the suitable server dynamically via the information from ping response messages. Therefore, we will focus our discussion on ping-based methods, i.e., ping, pr2, and r2p, and tell the differences among them in the following sections.

3) Control Overhead Comparison: In this experiment, we assume that the infrastructure-based WLAN, i.e., the wlan method, does not require additional wireless band-

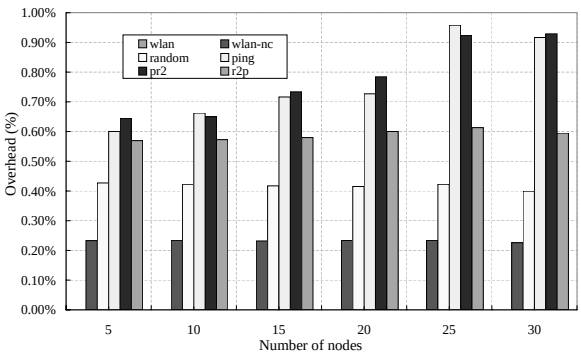


Figure 4. Poisson arrival: Overhead.

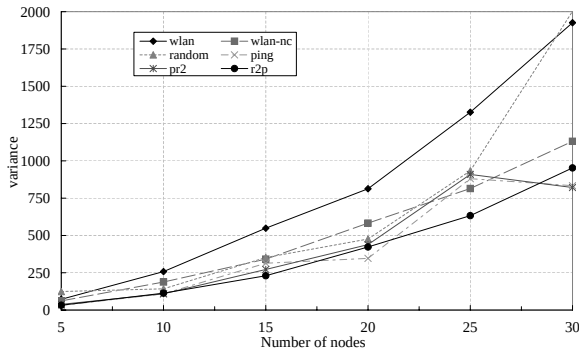


Figure 5. Poisson arrival: Variance.

width for exchanging control messages. In contrast, both the wlan-nc method and the hybrid methods do require additional wireless bandwidth to exchange information like the server list. The bandwidth requirements are normalized by all transferred data packets including the requested objects and the control messages. As the results shown in Figure 4, the bandwidth overhead of r2p is no more than 0.6% while the ping and pr2 methods result in the bandwidth overheads that don't exceed 1%. It shows that our H-WLAN can have a substantial performance improvement but the additional bandwidth requirement is small relatively.

4) Variance Comparison: As illustrated in Figure 5, the random method causes the higher variance compared with the rest of selection methods. This is not surprising since the random method is to select a server randomly from the server list without any further information; namely, the probability is equal for a mobile node to choose the nearest server or the farthest one. The rest of adaptive hybrid methods such as ping, pr2, and r2p can provide fair and consistent service (mean response time) compared with the random method under this simulation setup.

4.1.2. Bulk arrival case: After the above discussions, an important and interesting question arises: which service selection method is most suitable for the hybrid system. Thus,

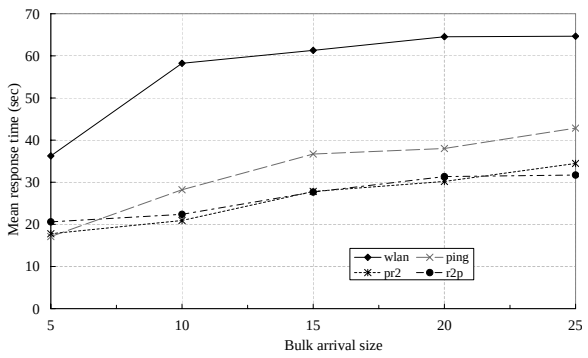


Figure 6. Bulk arrival: Mean response time.

we consider a more bursty arrival pattern and observe the performances under the bursty network condition. In this simulation, all mobile nodes want to obtain the same single object but the number of each arriving nodes may be greater than 1. That is, the bulk arrival size of first 5 arrivals is 1 and the bulk arrival size of the remaining arrivals is sb, where sb is from 5 to 25. There are total 30 mobile users who are uniformly distributed within the simulation area.

As shown in Figure 6, both pr2 and r2p methods can maintain lower mean response time compared with the ping method, where we can make the following observations. As the arrival is not bursty, i.e., the bulk arrival size is small, the ping method can make a good decision to choose a suitable server. However, when the bulk arrival size increases, it is not a good idea for all the users to choose its best server since they might choose the same one and swamp it. Consequently, the mobile node can avoid selecting the worst server and at the same time keep a load balance among all available servers. This is why the pr2 and r2p perform better than the ping method does.

4.2. Channel Management Study

Since r2p method performs better than or as well as other server-channel selection schemes and only needs small control overhead, we use it to evaluate the performance of popularity-aware channel management module. In this experiment, the popularity of objects is the Zipf distribution, given in Eq. (1), and we set $\theta = 0$ and $K = 10$.

$$\begin{aligned} &\text{Prob[request of object } i] = \tfrac{c}{i^{(1-\theta)}} \\ &\forall i = 1, 2, \ldots, K \text{ and } 0 \le \theta \le 1 \\ &\text{Where } c = \tfrac{1}{H_K^{(1-\theta)}} \text{ and } H_K^{(1-\theta)} = \textstyle\sum_{j=1}^{K} \tfrac{1}{j^{(1-\theta)}} \end{aligned} \quad (1)$$

Each mobile node desires to get a single object, which is randomly selected from 10 possible objects. There are 5 additional wireless channels could be allocated for all objects to construct the ad hoc group. We use the first-k and round-

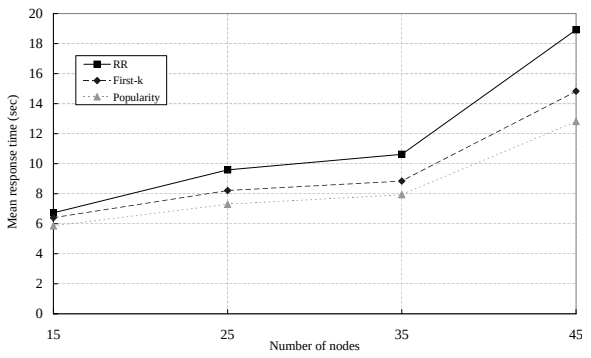


Figure 7. Channel management.

robin (RR) method as the base case for comparing with our popularity-aware mechanism.

In first-k method, all available channels can only be allocated to first k request objects, and each object may get multiple channels if it requests more than once and the request is still first-k ones. In this simulation setup, if one channel can be only allocated to single object, 5 wireless channels can only respond to first 5 requests. When the popularity of object is uniform distribution, first-k method might perform worse owing to the low utilization of bandwidth, i.e., the channels are not fully utilized but also not able to be allocated to other objects. On the other hand, it gets the acceptable performance, as shown in Figure 7, when the popularity of object is non-uniform, since the first 5 requests, which get the wireless channels to set up new ad hoc groups, may happen to share the more popular object and use bandwidth adequately.

In the RR method, the APs allocate the channel by round-robin whether all bandwidth in this channel is fully used or not, and authorize objects to get the resource by round-robin as well, i.e., if an object was once authorized to get the channel, it is able to get additional channel unless all other objects have been allowed to get the channel once. It is fair for every objects but not a good scheme for this case that the popularity of objects is the Zipf distribution. As illustrated in Figure 7, the round-robin method has the worst performance, compared to others, due to improper management. The more popular objects can not get sufficient bandwidth to use, but at the same time the bandwidth that is allocated to non-popular objects is still idle. Moreover, the results show that our popularity-aware mechanism has the best performance under the skew popularity. The lower mean response time also illustrates that the users can get a better QoS and the total amount of bandwidth can be utilized more effectively.

5. Conclusions

In this paper, we discussed the issues about the integration of ad hoc mode wireless network with infrastructure-based WLAN for the data dissemination applications. The main contributions of H-WLAN are as follows. First, we have shown that it is promising to use extra available channels to set up the ad hoc groups for aiding the traditional infrastructure-based WLAN, and the simulation results also illustrate that the QoS of users and the throughput and utilization of system are both substantially enhanced in H-WLAN. Second, the adaptive service selection module and the popularity-aware channel management module are proposed in this work. Both modules can provide the efficient resource management and consistent service. The mobility issue of nodes is beyond the focus of this work since the mobility may lead to separation in an ad hoc group and termination in current connections. Our future work plans to take the mobility issue into consideration.

References

[1] Y. Bejerano. Effi cient integration of multi-hop wireless and wired networks with qos constraints. In *Proceedings of ACM Mobicom*, Sep. 2002.

[2] J. Chen, S. G. Chan, J. He, and S. Liew. Mixed-mode wlan: The integration of ad hoc mode with wireless lan infrastructure. In *Proceedings of IEEE Globecom*, Dec. 2003.

[3] W. C. Cheng, C.-F. Chou, L. Golubchik, and S. Khuller. A performance study of bistro, a scalable upload architecture. In *In ACM SIGMETRICS Performance Evaluation Review*, pages 31–39, March 2002.

[4] http://www.isi.edu/nsnam/ns/. *The Network Simulator - ns-2*.

[5] S. B. Kodeswaran, O. Ratsimor, A. Joshi, T. Finin, and Y. Yesha. Using peer-to-peer data routing for infrastructure-based wireless networks. In *PerCom*, page 109120, Mar. 2003.

[6] Y. Sun and E. M. Belding-Royer. Application-oriented routing in hybrid wireless networks. In *Proceedings of ACM Mobicom*, May 2003.

[7] P. U. WICAT. *Information Project*. http://wicat.poly.edu/infostation.htm.

[8] T. Ye, H.-A. Jacobsen, and R. H. Katz. Mobile awareness in a wide area wireless network of info-stations. In *Mobile Computing and Networking*, pages 109–120, 1998.

[9] L. Yin and G. Cao. Supporting cooperative caching in ad hoc networks. In *Proceedings of IEEE INFOCOM*, Mar. 2004.

[10] W. H. Yuen, R. D. Yates, and S.-C. Mau. Noncooperative content distribution in mobile infostation networks. In *IEEE WCNC*, Mar. 2003.