行政院國家科學委員會專題研究計畫 成果報告

調變頻譜分析於音樂曲風及樂器音色之自動分類辨識之研究
研究成果報告(精簡版)

計 畫 類 別 ：個別型
計 畫 編 號 ：NSC 99-2221-E-216-048-
執 行 期 間 ：99 年 08 月 01 日至 100 年 10 月 31 日
執 行 單 位 ：中華大學資訊工程學系

計 畫 主 持 人：李建興
共 同 主 持 人：連振昌、陳建宏
計畫參與人員：碩士班研究生-兼任助理人員：廖伯鈞
碩士班研究生-兼任助理人員：方仁政
碩士班研究生-兼任助理人員：鄭永坤
博士班研究生-兼任助理人員：林懷三
其他-兼任助理人員：莊清乾

報 告 附 件 ：出席國際會議研究心得報告及發表論文

公 開 資 訊 ：本計畫可公開查詢

中 華 民 國 101 年 01 月 10 日

中 文 摘 要 ： 本計畫我們提出新的調變頻譜特徵並將其應用於音樂曲風之自動分類，我們分別對八度音程頻譜對比值(OSC)聲譜圖、MPEG-7 之正規化聲音頻譜封包(NASE)聲譜圖及梅爾倒頻譜(MFCC)聲譜圖做調變頻譜分析，分別產生 OSC 調變聲譜圖(modulated OSC spectrogram)、NASE 調變聲譜圖(modulated NASE spectrogram)及 MFCC 調變聲譜圖(modulated MFCC cepstrogram)，然後自每一個調變頻譜圖中擷取新的調變頻譜特徵值，包含調變頻譜能量值、調變頻譜平滑度、調變頻譜質心、調變頻譜波谷值及調變頻譜對比值，接著以主軸分析(principal component analysis, PCA)演算法來選取適當之調變頻譜特徵值並降低特徵向量維度，再以多重特徵向量來表示同一種音樂曲風，最後再以線性區別分析(linear discriminant analysis, LDA)演算法或非參數區別分析(nonparametric discriminant analysis, NDA)演算法來提升辨識率。實驗結果比較我們所提出之方法及 2004 年音樂曲風分類競賽之前五名參賽者，還有其他具備相同實驗設定之論文，我們所提出之方法得到最佳之辨識率(89.44%)，比 2004 年音樂曲風分類競賽之優勝者(84.07%)還高 5.37%。

中文關鍵詞： 音樂曲風分類、調變頻譜、主軸分析演算法、線性區別分析演算法、非參數區別分析演算法

英 文 摘 要 ： In this project, a new set of modulation spectral features are proposed for automatic music genre classification based on long-term modulation spectral analysis on the spectral (OSC and MPEG-7 NASE) as well as cepstral (MFCC) features. First, modulation spectral analysis is performed respectively on the OSC spectrogram, NASE spectrogram, and MFCC spectrogram to obtain the corresponding modulation spectrograms. Second, each modulation spectrogram is decomposed into several logarithmically-spaced modulation subbands. From each modulation subband, the modulation subband energy, modulation spectral centroid, modulation spectral flatness, modulation spectral valley, and modulation spectral contrast are computed as the modulation features of each subband. Principal component analysis (PCA) and linear discriminant analysis (LDA) or nonparametric discriminant analysis (NDA) are then employed to reduce the feature dimension and improve the

classification accuracy. Experiments conducted on the music database employed in the ISMIR2004 Audio Description Contest (ISMIR2004) have shown that the proposed approach can achieve a classification accuracy of 89.44%, which is better than the winner of the contest by 5.37%.

英文關鍵詞： music genre classification, modulation spectrum, principal component analysis (PCA), linear discriminant analysis (LDA), nonparametric discriminant analysis (NDA)

行政院國家科學委員會補助專題研究計畫▓ 成 果 報 告
□期中進度報告

調變頻譜分析於音樂曲風及樂器音色之自動分類辨識之研究

計畫類別：☑ 個別型計畫　　☐ 整合型計畫
計畫編號：NSC 99-2221-E-216-048-
執行期間：2010 年 08 月 01 日 至 2011 年 10 月 31 日

計畫主持人：李建興
共同主持人：連振昌、陳建宏
計畫參與人員：林懷三、方仁政、廖伯鈞、莊清乾

成果報告類型(依經費核定清單規定繳交)：☑精簡報告　☐完整報告

本成果報告包括以下應繳交之附件：
☐赴國外出差或研習心得報告一份
☐赴大陸地區出差或研習心得報告一份
☑出席國際學術會議心得報告及發表之論文各一份
☐國際合作研究計畫國外研究報告書一份

處理方式：除產學合作研究計畫、提升產業技術及人才培育研究計畫、列管計畫及下列情形者外，得立即公開查詢
☐涉及專利或其他智慧財產權，☐一年☐二年後可公開查詢

執行單位：中華大學資訊工程學系
中　華　民　國　101 年　01 月　10 日

摘要

本計畫我們提出新的調變頻譜特徵並將其應用於音樂曲風之自動分類，我們分別對八度音程頻譜對比值(OSC)聲譜圖、MPEG-7 之正規化聲音頻譜封包(NASE)聲譜圖及梅爾倒頻譜(MFCC)聲譜圖做調變頻譜分析，分別產生 OSC 調變聲譜圖(modulated OSC spectrogram)、NASE 調變聲譜圖(modulated NASE spectrogram)及 MFCC 調變聲譜圖(modulated MFCC cepstrogram)，然後自每一個調變頻譜圖中擷取新的調變頻譜特徵值，包含調變頻譜能量值、調變頻譜平滑度、調變頻譜質心、調變頻譜波谷值及調變頻譜對比值，接著以主軸分析(principal component analysis, PCA)演算法來選取適當之調變頻譜特徵值並降低特徵向量維度，再以多重特徵向量來表示同一種音樂曲風，最後再以線性區別分析(linear discriminant analysis, LDA)演算法或非參數區別分析(nonparametric discriminant analysis, NDA)演算法來提升辨識率。實驗結果比較我們所提出之方法及 2004 年音樂曲風分類競賽之前五名參賽者，還有其他具備相同實驗設定之論文，我們所提出之方法得到最佳之辨識率 (89.44%)，比 2004 年音樂曲風分類競賽之優勝者(84.07%)還高 5.37%。

一. 報告內容

1. 前言

對於音樂曲風之分類，其擷取之特徵向量可分為短時距特徵(short-term features)和長時距特徵(long-term feature)兩類。短時距特徵是從一段較短時間(通常是一個音框)之音樂訊號中所擷取之特徵向量，一般而言是屬於較低階之音樂特徵。最常用來做為音樂曲風分類之音樂特徵可分為三類：音色(timbre)、節奏(rhythm)及音高(pitch)。

音色特徵通常呈現演奏之樂器或聲音來源之特性，譬如音樂、語音、及環境聲音等。通常較常使用之音色特徵有以下幾種：

(1) 低能量特徵 (low-energy feature, LEF)

此特徵之定義是將連續數個音框看成一個紋理視窗(texture window)，首先計算每個音框之能量值：

$$E_{RMS}(n) = \left(\frac{1}{M}\sum_{m=0}^{M-1}(x[n\times M+m])^2\right)^{1/2}$$

其中，M 表示每個音框的樣本數目，然後計算紋理視窗中所有音框能量的平均值：

$$\overline{E}_{RMS} = \frac{1}{N}\sum_{n=0}^{N-1}E_{RMS}(n)$$

其中，N 表示每個紋理視窗中的音框數目，此低能量特徵是計算在此紋理視窗中有

多少百分比之音框其能量值低於所有音框之能量平均值：

$$LEF = \frac{1}{N}\sum_{n=1}^{N} LEI(n)$$

其中

$$LEI(n) = \begin{cases} 1, & E_{\mathrm{RMS}}(n) \geq \overline{E}_{RMS} \\ 0, & otherwise \end{cases}$$

(2) 越零率 (zero-crossing rate)

計算經過振幅為零的水平線的次數，可以用來表示時間域上的頻率程度，其特徵擷取公式如下：

$$Z_t = \frac{1}{2}\sum_{n=1}^{N} | sign(x[n]) - sign(x[n-1]) |$$

其中 t 表示音框索引值，N 為音框大小，且

$$sign(x[n]) = \begin{cases} 1, & x[n] \geq 0 \\ 0, & x[n] < 0 \end{cases}$$

(3) 頻譜質心 (spectral centroid)

主要用來表示聲音訊號組成頻率的平均值，可用來表示聲音的明亮(brightness)程度。其計算公式如下：

$$C_n = \frac{\sum_{m=1}^{N/2} M_n[m] \times m}{\sum_{m=1}^{N/2} M_n[m]}$$

其中 C_n 為頻譜質心，N 為音框大小，n 為音框索引值(frame index)，$M_n[m]$ 為第 m 個頻率係數之強度值。

(4) 頻譜頻寬 (spectral bandwidth)

用來描述頻譜圖的形狀，其值越小表示頻率分布越集中在頻譜質心，可以用其來描述頻率分布是集中在頻譜質心亦或是分散在各個頻率。

$$D_n = \sqrt{\frac{\sum_{m=1}^{N} (m - C_n)^2 \times M_n[m]}{\sum_{m=1}^{N} M_n[m]}}$$

(5) 頻譜滑動率 (spectral rolloff)

也是用來描述頻譜圖形狀的另一種用法。

$$\sum_{m=1}^{R_n} M_n[m] = TH \times \sum_{m=1}^{N} M_n[m]$$

其中 R_n 為頻譜滑動率，$TH = 0.85$ 為此公式常用的數值

(6) 頻譜變遷度 (spectral flux)

用來描述前後連續兩個音框間的頻譜差異性，可用來分割聲音片段。

$$F_n = \sum_{m=1}^{N} (N_n[m] - N_{n-1}[m])^2$$

其中 N_n[m]為將 M_n[m]正規化的值，其計算公式如下：

$$N_n[m] = \frac{M_n[m]}{\sum_{m=1}^{N} M_n[m]}$$

(7) 梅爾倒頻譜係數 (Mel-frequency cepstral coefficients, MFCC)

梅爾倒頻譜係數之概念是用一組模擬人類聽覺系統之梅爾濾波器來過濾每一臨界頻帶之聲音訊號，並對每一頻帶之對數能量頻譜值(logrithmic spectra)做離散餘弦轉換(discrete cosine transform, DCT)，即可求得每一音框之梅爾倒頻譜係數。

(8) 八度音程頻譜對比值 (octave-based spectral contrast, OSC)

八度音程頻譜對比值[1]是用來描述每一個八度音程的子頻帶中，頻譜波峰和波谷的強度值之差異值，如此可以大略的反映聲音訊號之泛音(harmonic)和非泛音(non-harminic)的分佈狀況。

節奏特徵主要是描述一首音樂之節奏特性，通常是由一段音樂中的節拍統計圖(beat histogram)中擷取其節奏特徵，包括所有節拍的強度、主節拍的速度及強度、主節拍和次節拍之速度間距，以及主節拍和次節拍的相對強度值。預估主節拍速度和其對應強度的方法可參考[2, 3]。Tzanetakis 提出從一首音樂之音高統計圖(pitch histogram)中擷取音高特徵的方法[4]，其特徵包括頻率、音高強度值和音高間距。此一音高統計圖可以使用各種音高偵測演算法來統計得到[5, 6]，而旋律與泛音也廣泛地由音樂家用來研究音樂的結構，因此，Scaringella 等人提出藉由描述每一小段音樂片段之音高分佈來擷取旋律與泛音的方法[7]，此一方法類似旋律或泛音分析器，但不用事先決定較高階之音樂特性，如基頻、和弦或音樂調性。

通常要表示一整首音樂之特性時，必需將短時距的特徵向量整合起來構成長時距之特徵向量，整合的方式包括計算所有短時距特徵向量之平均值及標準差，或者以自我回

歸模型[8]或調變頻譜[9-11]來分析。

(1) 平均值和標準差

最常被使用來整合短時距特徵向量的方法是計算所有特徵向量之平均值和標準差，然而以此統計資料來描述一整首音樂並無法顯示音樂訊號隨時間變化之特性。

(2) 自我回歸模型 (autoregressive (AR) model)：

Meng 等人以 AR 模型來分析音樂訊號隨著時間變化的特性[8]，他們提出以對角自我回歸模型(diagonal autoregressive model, DAR)與多變量自我回歸模型(multivariate autoregressive model, MAR)分析來整合短時距特徵向量。在 DAR 模型裡，將每一個短時距特徵值視為一個獨立的 AR 模型，並計算所有短時距特徵值的平均值、標準差和每一個 AR 模型的回歸係數作為長時距特徵向量。在 MAR 模型中，將短時距特徵向量以一個多變量自我回歸模型來表示。MAR 模型和 AR 模型最大的不同在於 MAR 模型考慮了特徵值間之關聯性，因此，在 MAR 模型下所擷取的長時距特徵向量則包含所有短時距特徵向量的平均值、共變異數矩陣和 MAR 模型的回歸係數。

(3) 調變頻譜分析 (modulation spectrum analysis)

調變頻譜分析是要觀察沿著時間軸上頻率的變化情形，此方法最早是由 Kingsbury 提出用來做語音的辨識[9]，其研究顯示人類聽覺最敏感的調變頻率大約在 4Hz 左右。Sukittanon 也使用調變頻譜分析來辨識分析音樂之內容[10]，其實驗顯示對每一子頻帶正規化後之調變頻率特徵受到旋積雜訊干擾之影響較小。Shi 同樣使用調變頻譜分析來描述音樂訊號之長時距特性[11]，用以擷取音樂之節拍速度以對不同情感之音樂類型做分類。

本計畫中，我們將提出新的調變頻譜特徵並將其應用於音樂曲風之自動分類，這些調變頻譜特徵值包含調變頻譜能量值、調變頻譜平滑度、調變頻譜質心、調變頻譜波谷值及調變頻譜對比值，接著以主軸分析(principal component analysis, PCA)演算法來選取適當之調變頻譜特徵值並降低特徵向量維度，再以多重特徵向量來表示同一種音樂曲風，最後再以線性區別分析(linear discriminant analysis, LDA)演算法或非參數區別分析(nonparametric discriminant analysis, NDA)演算法來提升辨識率。

2. 研究目的與研究方法

本計畫應用調變聲譜圖於音樂曲風之自動辨識分類，調變聲譜圖主要是對八度音程頻譜對比值(octave spectral contrast, OSC)、MPEG-7 之正規化聲音頻譜封包(NASE)及梅爾倒頻譜(MFCC)聲譜圖做調變頻譜分析，分別產生 OSC 調變聲譜圖、NASE 調變聲譜圖及 MFCC 調變聲譜圖，然後我們將每一個調變頻譜分解成對數間距之調變頻帶，接著自每一調變頻帶中擷取五種調變特徵值：調變頻譜能量值、調變頻譜平滑度、調變頻譜質心、調變頻譜波谷值及調變頻譜對比值。然後以主軸向量分析演算法來選取適當之調變頻譜特徵值並降低特徵向量維度，再以多重特徵向量來表示同一種音樂曲風，最後再以線性區別分析演算法或非參數區別分析演算法來辨識輸入之音樂檔案是屬於何種類別之音樂曲風。

本計劃之音樂曲風自動分類辨識系統包含訓練階段和辨識階段兩部分，訓練階段是由四個主要模組所組成: 調變頻譜特徵擷取、主軸向量分析演算法、多重特徵向量分群演算法、及線性區別分析演算法或非參數區別分析演算法。辨識階段是由四個主要模組所組成: 調變頻譜特徵擷取、主軸向量分析轉換、線性區別分析轉換或非參數區別分析轉換、和分類。

2.1 調變頻譜特徵擷取

首先，我們先建構聲音訊號之聲譜圖，包括 OSC 聲譜圖、NASE 聲譜圖及 MFCC 聲譜圖，然後對各式各樣聲譜圖以調變頻譜分析來描述其隨時間之變化趨勢以擷取辨識特徵。

2.1.1 OSC 聲譜圖

OSC 是用來描述一音樂訊號之頻譜特性[1]，首先將音樂訊號依據八度音程之觀念將其分解為 B 個(本計劃中 $B = 9$)子頻帶，每一子頻帶之頻率範圍請參考表一，然後分別計算每一子頻帶之頻譜波峰和波谷的強度值，一般而言，頻譜波峰主要反映聲音訊號之泛音(harmonic)成份，而波谷相當於非泛音(non-harminic)或雜訊成份，因此頻譜波峰值和波谷值的之差異值可以大略的反映聲音頻譜的對比分佈狀況。對於一音樂訊號，我們先將其切割成一個個音框，然後以傅立葉轉換得到每一音框之聲音頻譜，接下來以八度音程之帶通濾波器(octave scale band-pass filter) 將一音框之聲音頻譜分解為 B 個子頻帶，然後再對每一子頻帶計算其頻譜對比特徵。假設$(x_{b,1}, x_{b,2}, \ldots, x_{b,N_b})$代表第 b 個子頻帶之強度頻譜，N_b 代表所有位於第 b 個子頻帶中之傅立葉轉換係數之數目，假設此一子頻帶之強度頻譜已經依據其強度值由大至小排序過，也就是說 $x_{b,1} \geq x_{b,2} \geq \ldots \geq x_{b,N_b}$，第

b 個子頻帶之頻譜波峰及波谷的強度值就可以下列公式來預估：

$$Peak_b = \log(\frac{1}{\alpha N_b}\sum_{i=1}^{\alpha N_b} x_{b,i})$$

$$Valley_b = \log(\frac{1}{\alpha N_b}\sum_{i=1}^{\alpha N_b} x_{b,N_b-i+1})$$

其中α為鄰近區之參考因子(本計劃中設$\alpha = 0.2$)，第 b 個子頻帶之頻譜對比值可定義為

$$SC_b = Peak_b - Valley_b.$$

對每一音框，我們取所有子頻帶之頻譜波谷值($Valley_b$, $1 \le b \le B$)及頻譜對比值(SC_b, $1 \le b \le B$)為此一音框之 OSC 特徵向量，然後我們將所有音框之 OSC 係數(包含所有子頻帶之頻譜波谷值及頻譜對比值)沿著時間軸串接起來構成二維之影像圖，稱為 OSC 聲譜圖。

表一. 八度音程之每一子頻帶之頻率範圍 (Sampling rate = 44.1 kHz)

OSC Subband	Low Frequency (Hz)	High Frequency (Hz)
1	0	100
2	100	200
3	200	400
4	400	800
5	800	1600
6	1600	3200
7	3200	6400
8	6400	12800
9	12800	22050

2.1.2 NASE 聲譜圖

在MPEG-7標準中，是以對數之頻率間格來描述音訊訊號的頻譜圖。由於人類的聽覺系統對於頻率的敏感度是呈現對數之對應關係，所以我們使用對數頻率來取頻率間距，如此可以兼顧描述性與簡潔性。聲音頻譜封包(audio spectrum envelope, ASE)在MPEG-7標準裡普遍用於表示原始聲音訊號中每一子頻帶之功率頻譜[12-14]，主要是描述介於*loEdge* (預設62.5Hz)與*hiEdge* (預設為16000Hz)間的頻譜資訊，將介於*loEdge*與*hiEdge*間的頻率再分解為B個子頻帶，而每一子頻帶的頻寬解析度是以八度音(octave)解析度為基準，以1000Hz為中心上下區分，總計介於[*loEdge*, *hiEdge*]間之子頻帶數目為 $B = 8/r$，其中r是八度音的解析度，其範圍是介於1/16倍八度音至8倍八度音之間：

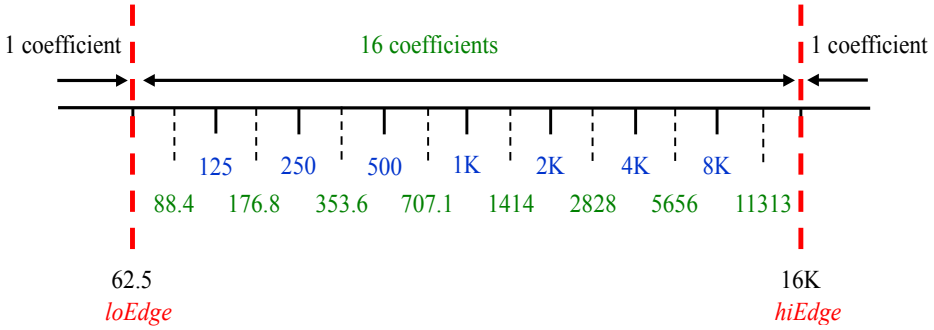
$$r = 2^j \text{ octaves}, \ -4 \le j \le 3$$

在本計劃中，我們採用之r值為1/2，因此$B = 16$，而每一子頻帶之邊界頻率(f_{edge})的公式如下：

$$f_{edge} = 2^{rm} \times 1000$$

其中m是整數。此外又加上兩個額外的子頻帶，一個為0Hz到*loEdge*的頻帶能量總合，一

個為*hiEdge*到取樣頻率一半的頻帶能量總合，因此整個頻譜範圍可分解為 $(B+2)$ 個子頻帶，圖一為一個八度音解析度之邊界頻率f_{edge}分隔圖。



圖一. 八度音頻帶濾波器(頻譜解析度 $r = 1/2$)

NASE在MPEG-7標準中是針對每一個音框之ASE係數轉換至分貝之刻度單位後做正規化之動作，然而對一段音樂訊號而言，可能包含了許多音框，所以我們將所有音框之NASE係數沿著時間軸串接起來構成二維之影像圖，稱為NASE聲譜圖。我們取所有音框之NASE係數值($NASE(b)$, $0 \le b \le B+1$)及RMS值($R(b)$)沿著時間軸串接起來構成二維之影像圖，稱為NASE聲譜圖。

2.1.3 MFCC 聲譜圖

梅爾倒頻譜係數已經廣泛應用於語音辨識上[15-17]，事實上，梅爾(mel)是用以表示人類聽覺系統對一個音調(tone)感覺上的音高或頻率的計算單位，在人類聽覺系統中，對於一個音調的實際頻率(physical frequency)之反應並不是完全呈線性變化，而實際頻率(f)和梅爾頻率(mel)之間的對應關係在頻率低於 1 KHz 時是呈線性變化，但在高頻的部份則是呈現對數變化，兩者間之對應關係式如下：

$$mel = 2595 \log_{10}(1 + \frac{f}{700})$$

$$f = 700(10^{\frac{mel}{2595}} - 1)$$

人類之聽覺系統可將聲音之頻率分為一個個臨界頻帶(critical band)，位於同一臨界頻帶內之頻率聲音對人耳聽起來是相似的，因此我們可以用一組梅爾濾波器來過濾每一臨界頻帶之音樂訊號，並對每一頻帶之對數能量頻譜值(logrithmic spectra)做離散餘弦轉換(discrete cosine transform, DCT)，即可求得每一音框之 MFCC 係數，然後將所有音框之 MFCC 係數沿著時間軸串接起來構成二維之影像圖，稱為 MFCC 聲譜圖。

對於每一種聲譜圖(OSC 聲譜圖、NASE 聲譜圖及 MFCC 聲譜圖)，我們將分別對其做調變頻譜分析來描述各聲譜圖中之頻譜(或倒頻譜)係數隨時間變化之辨識特徵。

2.1.4 調變頻譜分析(modulation spectral analysis)

調變頻譜特徵主要用來描述一聲音訊號中某一特定頻率(或頻帶)隨時間變化之過程。首先，對時間域之聲音訊號做傅立葉轉換後可以得到在頻率域中之短時間聲音的頻譜係數。若對一特定頻譜係數追蹤其隨時間變化之軌跡，稱之為流動頻譜(running spectrum)，而對流動頻譜進行頻率分析之結果就稱為調變頻譜。

傳統的方法是在流動頻譜上使用帶通濾波器來擷取此一頻帶，但這方法不能正確地獲得調變頻譜，因此，Wada 等人提出調變頻譜控制(modulation spectrum control)方法[18]，對每一個頻率中的調變頻譜乘上一個權重值，以消除不需要的調變頻譜成份。此一權重值法可以避免帶通濾波器之一些限制，如濾波器係數的數目、延遲的時間、穩定度、和相位的誤差等，因此更適合用在聲音辨識系統。Kanedera 等人[19, 20]將調變頻譜中 0 至 40 Hz 之頻率範圍分成若干子頻帶，實驗在不同調變子頻帶之係數對於聲音辨識上之重要性，此一結果顯示在調變頻率 1 至 16 Hz 之範圍內所擷取之特徵有較好之辨識效果。Vuuren 與 Hermansky[21]同樣也將調變頻率分成若干子頻帶，且對調變頻率較低頻介於 0 至 1 Hz 部份更細分成三個子頻帶，實驗結果顯示在低頻 0.5 至 1 Hz 之子頻帶範圍內其重要性亦不差，其實驗結果顯示取調變頻譜之頻率介於 0.5 和 16 Hz 範圍內對語音辨識之重要性較大。

在本計劃中，我們擬將調變頻譜之觀念應用於不同之特徵值，包括 OSC 係數、NASE 係數及 MFCC 係數等，計算各種特徵值之調變頻譜，然後將調變頻譜切割成數個調變子頻帶(每個調變子頻帶之調變頻率分佈範圍可參考表二)，再從每一個調變子頻帶內擷取調變特徵值，最後將所有係數之調變特徵值串接起來構成二維之影像圖，稱為調變聲譜圖，包含 OSC 調變聲譜圖、NASE 調變聲譜圖及 MFCC 調變聲譜圖。

表二. 調變子頻帶之頻率範圍

Modulation Frequency Band	Modulation frequency interval (Hz)
0	[0, 0.5)
1	[0.5, 1)
2	[1, 2)
3	[2, 4)
4	[4, 8)
5	[8, 16)
6	[16, 32)
7	[32, 64)

首先，假設 $\mathbf{x}_n = [x_n(1), x_n(2), \ldots, x_n(D)]^{\mathrm{T}}$ 表示擷取自第 n 個音框之特徵向量，此特徵向量可以是第 n 個音框之 OSC 特徵向量、NASE 特徵向量或 MFCC 特徵向量，沿著時間軸對相同特徵值之連續 W 個音框(稱為分析視窗)做 FFT 轉換，即可得到其調變頻譜係數：

$$M_t(m,d) = \sum_{n=0}^{W-1} x_{(t\times W/2)+n}(d) e^{-j2\pi\frac{m}{W}n}, \quad 0 \le m < W, \quad 0 \le d < D,$$

其中 $M_t(m, d)$表示第 t 個分析視窗之調變頻譜，m 代表調變頻率索引值。接著我們將調變頻譜分為 J 個對數間距之調變子頻帶(modulation subband)，每個調變子頻帶之調變頻率分佈範圍可參考表二($J = 8$)，然後我們自每一個調變子頻帶中擷取五種調變頻譜特徵值：調變頻譜能量值(modulation subband energy, MSE)、調變頻譜平滑度(modulation spectral flatness, MSF)、調變頻譜質心(modulation spectral centroid, MSCEN)、調變頻譜波谷值(modulation spectral valley, MSV)及調變頻譜對比值(modulation spectral contrast, MSC)。

(1) 調變頻譜能量值之定義如下：

$$MSE(j,d) = \sum_{m=\Phi_{j,l}}^{\Phi_{j,h}} |M_t(m,d)|$$

其中 $\Phi_{j,l}$ 及 $\Phi_{j,h}$ 分別表示第 j 個調變子頻帶之下界頻率索引值和上界頻率索引值。

(2) 調變頻譜平滑度用來表示每一調變子頻帶中各個調變頻率之能量分布是否很平均，其定義如下：

$$MSF(j,d) = \frac{\sqrt[\Phi_{j,h}-\Phi_{j,l}+1]{\prod_{m=\Phi_{j,l}}^{\Phi_{j,h}} |M(m,d)|}}{\frac{1}{\Phi_{j,h}-\Phi_{j,l}+1}\sum_{m=\Phi_{j,l}}^{\Phi_{j,h}} |M(m,d)|}$$

(3) 調變頻譜質心表示每一調變子頻帶之能量質心，其定義如下：

$$MSE(j, d) = 10\log_{10}(1 + \sum_{m=\Phi_{j,l}}^{\Phi_{j,h}} (\overline{M}(m, d))^2)$$

(4) 調變頻譜波谷值(modulation spectral valley, MSV)及調變頻譜波峰值(modulation spectral peak, MSP)之定義如下：

$$MSP(j,d) = \max_{\Phi_{j,l} \le m \le \Phi_{j,h}} |M(m,d)|$$

$$MSV(j,d) = \min_{\Phi_{j,l} \le m \le \Phi_{j,h}} |M(m,d)|$$

(5) 調變頻譜對比值則定義為調變頻譜波峰值與調變頻譜波谷值之差異值：

$$MSC(j,d) = MSP(j,d) - MSV(j,d)$$

因此每一特徵向量之所有調變頻譜對比值、調變頻譜波谷值、調變頻譜平滑度、調變頻譜質心及調變頻譜能量值可構成五個 $D \times J$ 之矩陣，每一矩陣可視為二維之影像圖，統稱為調變聲譜圖。然後以主軸向量分析演算法來選取適當之調變頻譜特徵值並降低特徵向量維度，再以多重特徵向量來表示同一種音樂曲風，最後再以線性區別分析演算法或非參數區別分析演算法來辨識輸入之音樂檔案是屬於何種類別之音樂曲風。

2.2 主軸向量分析演算法(principal component analysis, PCA)

PCA 是先計算所有訓練資料之特徵向量的平均變異數矩陣 $E[\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}]$之 eigenvalue 及 eigenvector [22]，並以 eigenvector 當作基底來做線性轉換，而 eigenvalue 的大小可以決定其對應之 eigenvector 轉換後之特徵所保留之資訊量大小，eigenvalue 越大表示資料作線性轉換後，特徵的變異數值會越大，而變異數的大小又表示了分佈的寬廣，資料分佈越廣表示所保留之資訊量越大，也就是說，以 eigenvalue 值較大之 eigenvector 做為線性轉換之基底，轉換後的特徵分佈範圍會比以 eigenvector 較小的 eigenvector 轉換後的分佈範圍來得大。PCA 之進行步驟如下：

步驟 1：計算平均向量

$$\mathbf{m} = E[X]$$

其中 X是所有訓練資料之集合，$X = \{\mathbf{x}_i \mid i = 1 \ldots N\}$，m 是所有訓練資料的平均向量，$N$是訓練資料的數量。

步驟 2：令平均向量為 0

$$\mathbf{x}_i^{'} = \mathbf{x}_i - \mathbf{m}$$

步驟 3：求取平均變異數矩陣，C

$$\mathbf{C} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i^{'}(\mathbf{x}_i^{'})^{\mathrm{T}}$$

步驟 4：求取變異數矩陣 C 的 eigenvalue 及 eigenvector 並將其依 eigenvalue 值由大至小重新排序

步驟 5：設定臨界值α_{PCA}(表示所要保留的資訊量程度)，以計算轉換後維度 d

$$\sum_{i=1}^{d}\lambda_i \geq \alpha_{PCA} \times \sum_{i=1}^{D}\lambda_i$$

其中λ_i表示第 i 大之 eigenvalue，D 為轉換前之維度

步驟 6：以所保留之 d 個 eigenvector 對所有資料作線性轉換

$$\mathbf{x}_{\mathrm{PCA}} = \mathbf{A}_{\mathrm{PCA}}^{T}\mathbf{x}_i^{'}$$

其中 $\mathbf{A}_{\mathrm{PCA}}$ 為此 d 個較大 eigenvector 構成之 PCA 轉換矩陣。

2.3 多重特徵向量分群演算法

一般而言，自同一類別之音樂中所擷取之特徵向量呈現出極多樣之變化，因此我們必需以多重特徵向量才足以表示同一類別之音樂。本計畫中我們將以 c-means 分群演算法將屬於同一類別音樂之所有訓練特徵向量分成 K 小群，所有小群之群中心即構成某一類別音樂之所有代表特徵向量，因此全部有 $K \times C$ 個代表特徵向量，C 為音樂類別之數目。

2.4 線性區別分析演算法(linear discriminant analysis, LDA)

LDA 演算法之目的是將一個高維度的特徵向量轉換成一個低維度的向量，並且增加辨識的準確率[22]。LDA 演算法的主要精神是要把相同類別之間的距離最小化，並且把不同類別之間的距離給最大化，所以，必需決定一個轉換矩陣來將高維度的特徵向量

轉換成低維度向量，透過這樣的轉換我們能夠增強不同類別之間的差異性。最常使用的轉換矩陣主要依據 Fisher criterion J_F 來求得：

$$J_F(A) = tr((A^{\mathrm{T}} S_W A)^{-1}(A^{\mathrm{T}} S_B A))$$

其中，S_W 和 S_B 分別代表的是相同類別之散佈矩陣(within-class scatter matrix)和不同類別之散佈矩陣(between-class scatter matrix)，而相同類別之散佈矩陣的定義如下：

$$S_W = \sum_{j=1}^{C} \sum_{i=1}^{N_i} (\mathbf{x}_i^j - \boldsymbol{\mu}_j)(\mathbf{x}_i^j - \boldsymbol{\mu}_j)^{\mathrm{T}}$$

而 $\mathbf{x}_i^j$ 代表在類別 j 中的第 i 個特徵向量，$\boldsymbol{\mu}_j$ 為第 j 類的平均向量(mean vector)，C 為類別的數目，N_j 為類別 j 裡的特徵向量個數。而不同類別之散佈矩陣之定義如下：

$$S_B = \sum_{j=1}^{C} p_j (\boldsymbol{\mu}_j - \boldsymbol{\mu})(\boldsymbol{\mu}_j - \boldsymbol{\mu})^{\mathrm{T}},$$

$\boldsymbol{\mu}$ 為所有類別的平均向量，p_j 為第 j 類的事先機率。LDA 演算法的目的是要去求出能夠使不同類別之散佈矩陣和相同類別之散佈矩陣的比值為最大值之轉換矩陣 $\mathbf{A}_{\mathrm{LDA}}$，而其維度大小為 $n \times d$：

$$\mathbf{A}_{\mathrm{LDA}} = \arg\max_{A} \frac{tr(A^{\mathrm{T}} S_B A)}{tr(A^{\mathrm{T}} S_W A)}.$$

此一轉換矩陣，可經由求出 $S_W^{-1} S_B$ 的 eigenvectors 來得到，而 $\mathbf{A}_{\mathrm{LDA}}$ 之(C-1)個行向量為前(C-1)個最大 eigenvalue 值所對應之 eigenvector。

在我們決定出最佳的轉換矩陣 $\mathbf{A}_{\mathrm{LDA}}$ 後，我們以 $\mathbf{A}_{\mathrm{LDA}}$ 將每一 n 維的特徵向量轉換為(C-1)維之向量。令 $\mathbf{x}_{\mathrm{PCA}}$ 為 PCA 轉換後之特徵向量，則 LDA 轉換後之特徵向量如下:

$$\mathbf{x}_{\mathrm{LDA}} = \mathbf{A}_{\mathrm{LDA}}^{\mathrm{T}} \mathbf{x}_{\mathrm{PCA}}$$

2.5 非參數區別分析演算法(nonparametric discriminant analysis, NDA)

LDA 演算法假設每一類別之特徵向量是呈現高斯分佈，若是資料之分佈非高斯分佈，其辨識率會隨之下降。因此 Fukunaga 提出 NDA 演算法來解決此一問題[23]，其原始方法僅適用於兩種類別之分類上。在 NDA 中，相同類別之散佈矩陣的定義同 LDA，而不同類別之散佈矩陣之定義如下：

$$S_b^{NDA} = \sum_{l=1}^{N_1} w(1,l)\left(\mathbf{x}_l^1 - m_2(\mathbf{x}_l^1)\right)\left(\mathbf{x}_l^1 - m_2(\mathbf{x}_l^1)\right)^T + \sum_{l=1}^{N_2} w(2,l)\left(\mathbf{x}_l^2 - m_1(\mathbf{x}_l^2)\right)\left(\mathbf{x}_l^2 - m_1(\mathbf{x}_l^2)\right)^T$$

其中 $\mathbf{x}_l^i$ 代表第 i 類別之第 l 個特徵向量，$w(i,l)$為其權重函數值，$m_j(\mathbf{x}_l^i)$ 代表與 $\mathbf{x}_l^i$ 相近之 k 個特徵向量之平均值：

$$m_j(\mathbf{x}_l^i) = \frac{1}{k}\sum_{p=1}^{k} NN_p(\mathbf{x}_l^i, j)$$

其中 $NN_p(\mathbf{x}_l^i, j)$ 代表第 j 類別中與 $\mathbf{x}_l^i$ 第 p 個接近之特徵向量。

Li 等人將兩種類別之 NDA 演算法加以演繹成多種類別之 NDA 演算法(multiclass nonparametric discriminant analysis, MNDA) [24]，其定義之不同類別之散佈矩陣如下：

$$\mathbf{S}_b^{MNDA} = \sum_{i=1}^{C}\sum_{\substack{j=1\\ j\neq i}}^{C}\sum_{l=1}^{N_i} w(i,j,l)(\mathbf{x}_l^i - m_j(\mathbf{x}_l^i))(\mathbf{x}_l^i - m_j(\mathbf{x}_l^i))^T$$

其中 $w(i,j,l)$ 權重函數值之定義如下：

$$w(i,j,l) = \frac{\min\{d^\alpha(\mathbf{x}_l^i, NN_k(\mathbf{x}_l^i, i)), d^\alpha(\mathbf{x}_l^i, NN_k(\mathbf{x}_l^i, j))\}}{d^\alpha(\mathbf{x}_l^i, NN_k(\mathbf{x}_l^i, i)) + d^\alpha(\mathbf{x}_l^i, NN_k(\mathbf{x}_l^i, j))}$$

其中 $d(\mathbf{v}_1, \mathbf{v}_2)$代表特徵向量 $\mathbf{v}_1$ 及 $\mathbf{v}_2$ 之歐幾里德(Euclidean distance)距離，參數 α 控制權重函數值隨歐幾里德距離變化之幅度。依據 MNDA 所定義之權重函數值，當某一特徵向量接近其分類邊界時其權重函數值接近於 0.5，如果此特徵向量遠離分類邊界，則其權重函數值會接近於 0，也就是說 MNDA 會加強靠近分類邊界之特徵向量之權重函數值。令 $\mathbf{A}_{\text{NDA}}$ 為 NDA 之轉換矩陣，$\mathbf{x}_{\text{PCA}}$ 為 PCA 轉換後之特徵向量，則 NDA 轉換後之特徵向量如下:

$$\mathbf{x}_{\text{NDA}} = \mathbf{A}_{\text{NDA}}^{\text{T}}\mathbf{x}_{\text{PCA}}$$

2.6 分類辨識

在辨識的部份中，我們先將輸入之音樂訊號切割為一個個固定長度(1 或 2 秒鐘)之分析視窗，相臨之分析視窗間重疊約 0.5 秒，而且以一個分析視窗為基本辨識單位，自每一分析視窗中擷取特徵向量以分辨其屬於每一種類音樂之相似度，最後再將所有分析視窗之相似度整合計算以辨識此一輸入之音樂訊號是屬於何種音樂類別。假設第 t 個分析視窗之特徵向量為 $\mathbf{x}_t$，計算此一向量和每一辨識音樂類別之所有代表特徵向量($\mathbf{x}_{k,d}$, 1

$\le k \le C, 1 \le d \le N_k,$ 其中 C 為資料庫中音樂之種類數目，N_k 為第 k 種音樂類別之代表特徵向量數目)之間的距離，在這裡的距離公式是歐幾里德距離，假設輸入之音樂訊號總計可切割為 T 個分析視窗，令 $d_{t,k}$ 表示 $\mathbf{x}_t$ 和第 k 種音樂類別之所有 N_k 個代表特徵向量之間的最小距離：

$$d_{t,k} = \min_{1 \le d \le N_k} d(\mathbf{x}_t, \mathbf{x}_{k,d}),\ 1 \le t \le T,\ 1 \le k \le C$$

最終辨認之音樂種類代表編碼 s 可由下列公式來決定：

$$s = \arg\min_{1 \le k \le N} \prod_{t=1}^{T} d_{t,k}$$

3. 實驗結果與討論

在實驗中所使用之音樂資料庫為 2004 年音樂曲風分類競賽(*ISMIR2004 Music Genre Classification Contest*)所使用之音樂資料庫[25]，此資料庫中有 1458 首音樂檔案，其中有一半 729 首音樂檔案用於訓練，另外一半 729 首音樂檔案用於辨識，這些音樂檔案之取樣頻率為 44100 Hz，壓縮之位元率為 128 kbps，音訊範圍大小為 16 bits 且為立體聲之 MP3 檔案，在本實驗中，我們先將每一壓縮檔案轉換為 44100 Hz、16 bits 之單聲道音樂檔案。這些音樂檔案總共分為六種類別：古典音樂(*Classical*)、電子音樂(*Electronic*) 、爵士/藍調音樂(*Jazz/Blue*)、重金屬/龐克音樂(*Metal/Punk*)、搖滾/流行音樂(*Rock/Pop*)、及世界音樂(*World*)，總計用於訓練及辨識之古典音樂檔案分別有 320/320 首，電子音樂檔案分別有 115/114 首，爵士/藍調音樂檔案分別有 26/26 首，重金屬 /龐克音樂檔案分別有 45/45 首，搖滾/流行音樂檔案分別有 101/102 首，世界音樂檔案分別有 122/122 首。

為了與 2004 年音樂曲風分類競賽之參賽者之實驗結果比較，我們實驗中也是採用相同 50:50 之訓練檔案及辨識檔案比例，但是因為每一音樂類別之檔案數目不盡相同，因此其整體之辨識率定義如下：

$$CA = \sum_{1 \le c \le C} P_c \times CA_c,$$

其中 P_c 為第 c 種音樂類別之出現機率，CA_c 為第 c 種音樂類別之辨識率。

表格三比較各種調變特徵向量之辨識率，由此表格可以看出當 $\alpha_{PCA} = 0.99$ 時將所有特徵向量整合來計算距離時以 LDA/NDA 辨識可以得到最佳之辨識率(88.07%)。

表格四與表格五比較以多重特徵向量和 LDA 或 NDA 來辨識各種調變特徵之辨識率之辨識率，由此表格可以看出當以 LDA 來辨識可以得到之最佳辨識率為 89.30%，以 NDA 來辨識可以得到之最佳辨識率為 89.44%。

表三.以 LDA/NDA 辨識各種調變特徵之辨識率(%)

α_{PCA}	Feature Set	LDA	NDA
0.98	MMFCC	82.30	81.62
	MOSC	80.52	80.38
	MNASE	80.38	80.52
	MMFCC+MOSC+MNASE	86.83	**87.65**
0.99	MMFCC	81.48	83.26
	MOSC	82.17	81.62
	MNASE	80.38	81.34
	MMFCC+MOSC+MNASE	**88.07**	**88.07**

表四.以多重特徵向量和 LDA 辨識各種調變特徵之辨識率(%)

α_{PCA}	Feature Set	*K*=2	*K*=3	*K*=4	*K*=5	*K*=6	*K*=7	*K*=8	*K*=9
0.98	MMFCC	82.44	83.95	82.58	84.77	83.68	84.36	84.50	84.09
	MOSC	80.93	82.99	81.07	82.99	82.99	82.58	83.13	83.13
	MNASE	81.34	82.30	82.72	83.40	82.85	83.68	83.54	82.72
	MMFCC+MOSC+MNASE	89.03	**89.30**	88.48	87.65	87.65	86.97	87.65	87.79
0.99	MMFCC	83.68	83.54	83.68	85.19	84.22	85.60	85.46	84.09
	MOSC	83.26	83.68	83.54	82.72	83.13	84.36	84.50	85.46
	MNASE	81.89	82.72	83.40	84.22	84.09	83.95	84.50	83.81
	MMFCC+MOSC+MNASE	88.07	88.07	**88.89**	88.48	88.61	88.34	87.79	88.20

表五.以多重特徵向量和 NDA 辨識各種調變特徵之辨識率(%)

α_{PCA}	Feature Set	*K*=2	*K*=3	*K*=4	*K*=5	*K*=6	*K*=7	*K*=8	*K*=9
0.98	MMFCC	82.17	82.58	82.17	84.09	84.09	84.64	84.77	83.95
	MOSC	82.03	83.40	82.30	82.44	82.44	82.03	83.68	83.26
	MNASE	80.80	82.17	82.44	82.99	82.03	83.54	83.68	82.85
	MMFCC+MOSC+MNASE	89.30	88.20	**89.44**	87.93	87.79	86.42	88.20	87.65
0.99	MMFCC	84.09	84.64	84.77	83.95	83.68	85.05	84.50	84.77
	MOSC	83.68	84.09	82.44	82.85	83.40	83.68	84.22	85.60
	MNASE	82.03	84.09	83.26	84.77	83.95	83.81	83.81	83.13
	MMFCC+MOSC+MNASE	88.34	88.34	**89.30**	88.07	87.52	86.69	88.07	88.61

表格六比較我們所提出之方法及 2004 年音樂曲風分類競賽之前五名參賽者，還有

其他具備相同實驗設定之論文，由此表格中我們可以發現我們所提出之方法得到最佳之辨識率(89.44%)，比 2004 年音樂曲風分類競賽之優勝者(84.07%)還高 5.37%。

表六、對於 2004 年音樂曲風分類競賽之音樂資料庫之辨識率比較

References	*CA*
Our proposed approach	89.44%
Our previous approach [26]	86.83%
Y. Song *et al.* [27]	84.77%
T. Lidy & A. Rauber [28]	79.70%
E. Pampalk (winner)	84.07%
K. West (2nd rank)	78.33%
G. Tzanetakis (3rd rank)	71.33%
T. Lidy & A. Rauber (4th rank)	70.37%
D. Ellis & B. Whitman (5th rank)	64.00%

二. 參考文獻

[1] D. N. Jiang, L. Lu, H. J. Zhang, J. H. Tao, and L. H. Cai, “Music type classification by spectral contrast feature”, in *Proc. of IEEE Int. Conf. on Multimedia & Expo*, Vol. 1, pp. 113-116, 2002.

[2] M. E. P. Davies and M. D. Plumbley, “Beat tracking with a two state model”, in *Proc. ICASSP*, Vol. 3, pp. 241-244, 2005.

[3] W. A. Sethares, R. D. Robin, and J. C. Sethares, “Beat tracking of musical performance using low-level audio feature”, *IEEE Trans. Speech and Audio Processing,* Vol. 13, No. 12, Mar. 2005, pp. 275-285.

[4] G. Tzanetakis, A. Ermolinskyi, and P. Cook, “Pitch Histogram in Audio and Symbolic Music Information Retrieval”, in *Proc. IRCAM*, 2002.

[5] T. Tolonen and M. Karjalainen, “A computationally efficient multipitch analysis model”, *IEEE Trans. Speech and Audio Processing*, Vol. 8, No. 6, pp. 708-716, Nov. 2000.

[6] R. Meddis and L. O’Mard, “A unitary model of pitch perception”, *Journal of the Acoustical Society of America*, Vol. 102, No. 3, pp. 1811-1820, Sep. 1997.

[7] N. Scaringella, G. Zoia and D. Mlynek, "Automatic genre classification of music content: a survey", *IEEE Signal Processing Magazine*, Vol. 23, Issue 2, pp.133 - 141, Mar 2006.

[8] A. Meng, P. Ahrendt, J. Larsen, and L. K. Hansen, ”Temporal feature integration for music genre classification”, *IEEE Trans. Audio, Speech and Language Processing*, Vol. 15, No. 5, pp.1654-1664, July 2007.

[9] B. Kingsbury, N. Morgan, and S. Greenberg, “Robust speech recognition using the modulation spectrogram“, *Speech Communication*, Vol. 25, No. 1, pp.117-132, 1998.

[10] S. Sukittanon, L. E. Atlas, and J. W. Pitton, “Modulation-scale analysis for content identification”, *IEEE Trans. Signal Processing,* Vol. 52, No. 10, pp. 3023-3035, Oct. 2004.

[11] Y. Y. Shi, X. Zhu, H. G. Kim and K. W. Eom, "A tempo feature via modulation spectrum analysis and its application to music emotion classification", in *Proc. of 2006 IEEE Int. Conf. Multimedia and Expo (ICME)*, pp.1085-1088, July 2006.

[12] H. G. Kim, N. Moreau, and T. Sikora, “Audio classification based on MPEG-7 spectral basis representation”, *IEEE Trans. Circuits and Systems for Video Technology*, Vol. 14, No. 5, pp. 716-725, May 2004.

[13] H. G. Kim, N. Moreau, and T. Sikora, *MPEG-7 Audio and Beyond: audio content indexing and retrieval*, Wiley, 2005.

[14] B. S. Manjunath, P. Salembier, and T. Sikora, *Introduction to MPEG-7: Multimedia Content Description Interface*, Wiley, 2002.

[15] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[16] R. Vergin, D. O'Shaughnessy, and A. Farhat, “Generalized Mel frequency cepstral coefficients for large-vocabulary speaker-independent continuous-speech recognition”, *IEEE Trans. Speech and Audio Processing*, Vol. 7, No. 5, pp. 525-532, Sep. 1999.

[17] J. W. Picone, “Signal modeling techniques in speech recognition”, *Proceedings of the IEEE*, Vol. 81, pp. 1215–1247, 1993.

[18] N. Wada, N. Hayasaka, S. Yoshizawa, and Y. Miyanaga, **“**Direct control on modulation spectrum for noise-robust speech recognition and spectral subtraction**”,** in *Proc. IEEE Int. Symp. on Circuits and Systems*, 21-24 May, 2006, pp. 2533-2536.

[19] N. kanedera, T. Arai, H. Hermansky, and M. Pavel, “On the relative importance of various components of the modulation spectrum for automatic speech recognition”, *Speech Communication*, Vol. 28, Issue 1, May 1999, pp.43-55.

[20] N. kanedera, T. Arai, H. Hermansky, and M. Pavel, “On the importance of various modulation frequencies for speech recognition”, in *Proc. of ESCA*, 1997, pp. 1079-1082.

[21] S. V. Vuuren and H. Hermansky, “On the importance of components of the modulation spectrum for speaker verification”, in *Proc. of ICSLP*, Nov 1998.

[22] R. Duda, P. Hart, and D. Stork, *Pattern Classification*. New York:Wiley, 2000.

[23] K. Fukunaga, *Statistical Pattern Recognition*, Academic Press, 1990.

[24] Z. Li, D. Lin, and X. Tang, “Nonparametric Discriminant Analysis for Face Recognition”, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 31, no. 4,

pp. 755-761, Apr. 2009.

[25] http://ismir2004.ismir.net/ISMIR_Contest.html.

[26] C. H. Lee, J. L. Shih, K. M. Yu, and H. S. Lin, "Automatic Music Genre Classification Based on Modulation Spectral Analysis of Spectral and Cepstral Features", *IEEE Trans. on Multimedia*, vol. 11, no. 4, pp. 670-682, June 2009.

[27] Y. Song and C. Zhang, "Content-based information fusion for semi-supervised music genre classification", *IEEE Trans. on Multimedia*, vol. 10, no. 1, pp. 145-152, Jan. 2008.

[28] T. Lidy and A. Rauber, "Evaluation of feature extractors and psycho-acoustic transformations for music genre classification", in *Proc. 6th Int. Conf. on Music Information Retrieval*, 2005, pp. 34-41.

三. 計畫成果自評

本計劃完成音樂曲風之自動分類系統，能夠根據音樂的性質事先將音樂曲目分類為不同的曲風類型，有效率的管理龐大的音樂資料庫，此外也可做為音樂推薦系統使用，當使用者在選取一首喜愛的音樂時，可以將曲風相似之音樂曲目推薦給使用者，減少使用者搜尋性質相似之音樂所花的時間。當初提計劃書時預計以二年期之計劃來完成音樂曲風及樂器音色之自動分類辨識之系統，但是計劃只通過一年期，因此我們先完成音樂曲風自動分類系統，有關樂器音色之自動分類辨識系統則預計在後續之計劃中執行完成。目前我們已發表之相關論文如下：

期刊論文 (Journal Papers)：

[1] **C. H. Lee**, C. H. Chou, and J. C. Fang, "Automatic Music Genre Classification Using Modulation Spectral Features and Nonparametric Discriminant Analysis", *Journal of Information Technology and Applications*, Vol. 5, No. 2, June 2011, pp. 75-82.

[2] **C. H. Lee**, J. L. Shih, K. M. Yu, and H. S. Lin, "Automatic Music Genre Classification Based on Modulation Spectral Analysis of Spectral and Cepstral Features", *IEEE Trans. on Multimedia*, Vol. 11, No. 4, June 2009, pp. 670-682. (SCI, EI)

[3] **C. H. Lee**, C. C. Han, and C. C. Chuang, "Automatic Classification of Bird Species by Their Sounds Using Two Dimensional Cepstral Coefficients", *IEEE Trans. on Audio, Speech, and Language Processing*, Vol. 16, No. 8, Nov. 2008, pp. 1541-1550. (SCI, EI)

[4] **C. H. Lee**, C. H. Chou, C. H. Han, and R. Z. Huang, "Automatic Recognition of Animal Vocalizations Using Averaged MFCC and Linear Discriminant Analysis", *Pattern Recognition Letters*, Vol. 27, Issue 2, Jan. 2006, pp. 93-101. (SCI, EI)

[5] **C. H. Lee**, Y. K. Lee and R. Z. Huang, "Automatic recognition of bird songs using cepstral coefficients", *Journal of Information Technology and Applications*, Vol. 1, No. 1,

May 2006, pp. 17-23.

[6] J. L. Shih, **C. H. Lee**, and S. W. Lin, "Automatic classification of musical audio signals", *Journal of Information Technology and Applications*, Vol. 1, No. 2, Sep. 2006, pp. 95-105.

研討會論文 (Conference Papers)：

[1] **C. H. Lee**, C. H. Chou, C. C. Lien, and J. C. Fang, "Music Genre Classification Using Modulation Spectral Features and Multiple Prototype Vectors Representation", in *Proc. of the 4th International Congress on Image and Signal Processing (CISP'11)*, Oct. 15-17, 2011, Shanghai, China, pp. 2762-2766. (NSC-99-2221-E-216-048, EI)

[2] **C. H. Lee**, H. S. Lin, C. H. Chou, and J. L. Shih, "Modulation Spectral Analysis of Static and Transitional Information of Cepstral and Spectral Features for Music Genre Classification", in *Proc. of the 5th International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP'2009)*, Sep. 12-14, 2009, Kyoto, Japan, pp. 1030-1033. (EI)

[3] **C. H. Lee**, J. L. Shih, K. M. Yu, H. S. Lin, and M. H. Wei, "Fusion of Static and Transitional Information of Cepstral and Spectral Features for Music Genre Classification", in *Proc. of the IEEE Asia-Pacific Services Computing Conference*, Dec. 9-12, 2008, Yilan, Taiwan. (EI)

[4] **C. H. Lee**, J. L. Shih, K. M. Yu and H. S. Lin, "Modulation Spectral Analysis of Audio Features for Music Genre Classification", in *Proc. of the 21th IPPR Conference on Computer Vision, Graphics, and Image Processing*, Yilan, Aug. 24-26, 2008.

[5] C. H. Chou, **C. H. Lee** and H. W. Ni, "Bird Species Recognition by Comparing the HMMs of the Syllables", in *Proceedings of Second International Conference on Innovative Computing, Information and Control*, Kumamoto, Japan, Sep. 5-7, 2007. (EI)

[6] **C. H. Lee**, J. L. Shih, K. M. Yu and J. M. Su, "Automatic Music Genre Classification Using Modulation Spectral Contrast Feature", in *Proceedings of IEEE International Conference on Multimedia and Expo*, Beijing China, July 2007, pp. 204-207. (EI)

[7] **C. H. Lee**, C. C. Lien and R. Z. Huang, "Automatic Recognition of Birdsongs Using Mel-frequency Cepstral Coefficients and Vector Quantization", in *Proceedings of International MultiConference of Engineering and Computer Scientists*, Hong Kong, 2006, pp. 331-335.

[8] **C. H. Lee**, J. L. Shih, and S. W. Lin, "A novel approach to music genre classification", in *Proceedings of the 18th IPPR Conference on Computer Vision, Graphics, and Image Processing*, Taipei, Aug. 20-22, 2005.

行政院國家科學委員會補助國內專家學者出席國際學術會議報告

101 年 1 月 10 日

附件三

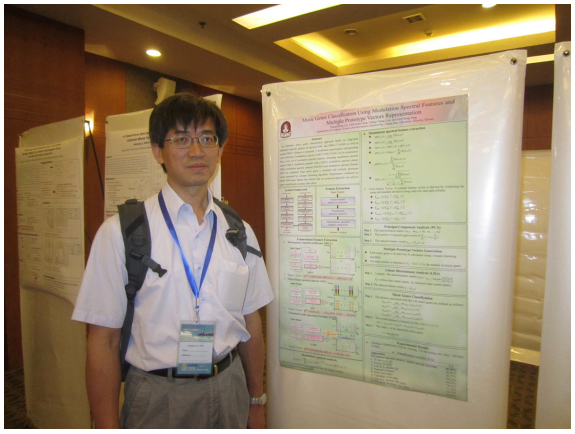
報告人姓名	李建興	服務機構及職稱	中華大學資訊工程系
會議時間地點	15-17 October, 2011 Shanghai, China	本會核定補助文號	計劃編號： NSC 99-2221-E-216-048-
會議名稱	(中文) (英文) The 4th International Congress on Image and Signal Processing (CISP 2011)		
發表論文題目	(中文) (英文) Music Genre Classification Using Modulation Spectral Features and Multiple Prototype Vectors Representation		

一、參加會議經過

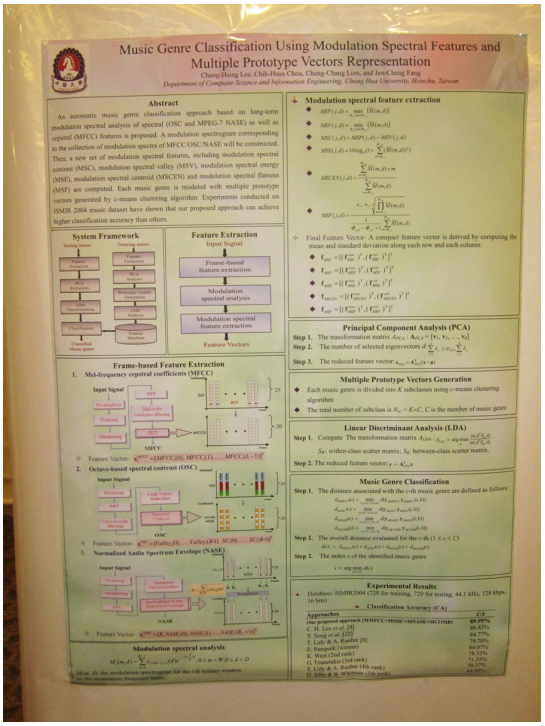
10/14 搭乘 16:30 長榮航空班機前往中國上海，下榻於 Radisson 上海新世界酒店。10/15 早上搭乘地鐵前往會議地點—光大會展中心國際大酒店，先註冊報到，中午用餐後，參加下午的 Session 1B(Image Processing Applications) 及 Session 2B(Feature extraction and machine vision in Images) 的演講。10/16 早上聆聽了 Session 3B (Speech and Language Processing) 及 Session 4B (Image enhancement and noise filtering)的演講。我們的論文則是安排於下午之 poster session，由於此一研討會是與 The 4th International Conference on BioMedical Engineering and Informatics 合辦，因此會場中除了有影像及訊號處理相關論文發表，還有許多和生物醫學有關之論文發表，相當熱絡。10/17 早上參加 Session 7B(Pattern Recognition in images) 及 Session 8B(signal processing applications) 的演講後，即返回飯店整理行李，準備搭機返回台灣。

會場-上海光大會展中心

會場報到處



poster session



發表之論文海報

二、與會心得

1. 本次會議是由中國東華大學主辦，會場附近無任何指示標誌，對於上海不熟悉者要花一些時間詢問路人方能抵達。
2. 主辦地點光大會展中心國際大酒店，有十幾個會議廳及宴會廳，因此開會者及參加宴會者容易混雜一起，弄錯會議室。
3. 大會午晚餐供應相當完善，讓與會者相當方便。
4. 此次會議參加的學者以亞洲籍(特別是中國籍)居多，較為美中不足。

三、考察參觀活動(無是項活動者省略)

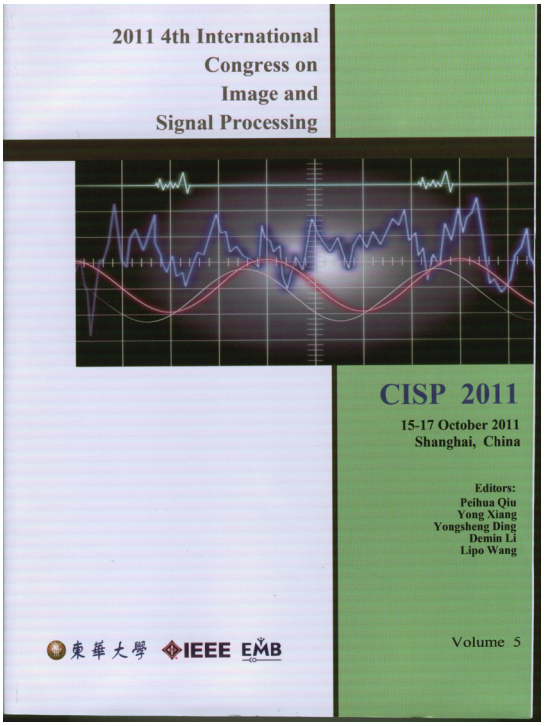
四、建議

1. 國內也常舉辦國際性研討會，可以於議程中安排半日或一日遊，讓外國學者增加認識台灣的機會，或於晚宴時安排有台灣特色的表演，對推展觀光也許有一些幫助。

五、攜回資料名稱及內容

論文光碟及論文輯

CISP 論文光碟



CISP 論文輯

六、其他

非常感謝國科會之補助得以參加該研討會。

Music Genre Classification Using Modulation Spectral Features and Multiple Prototype Vectors Representation

Chang-Hsing Lee, Chih-Hsun Chou, Cheng-Chang Lien, and Jen-Cheng Fang
Department of Computer Science and Information Engineering
Chung Hua University, Hsinchu, Taiwan

***Abstract*—In this paper, we will propose an automatic music genre classification approach based on long-term modulation spectral analysis of spectral (OSC and MPEG-7 NASE) as well as cepstral (MFCC) features. A modulation spectrogram corresponding to the collection of modulation spectra of MFCC/OSC/NASE will be constructed. The modulation spectrum is then decomposed into several logarithmically spaced modulation subbands. For each modulation subband, a new set of modulation spectral features, including modulation spectral contrast (MSC), modulation spectral valley (MSV), modulation spectral energy (MSE), modulation spectral centroid (MSCEN) and modulation spectral flatness (MSF) are then computed from each modulation subband. To cope with the problem that the feature vectors extracted from the music tracks of identical music genre might differ significantly, each music genre is modeled with a number of representative prototype vectors generated by c-means clustering algorithm. An information fusion approach which integrates both feature level fusion method and decision level combination method is then employed to improve the classification accuracy. Experiments conducted on ISMIR 2004 music dataset have shown that our proposed approach can achieve higher classification accuracy than other approaches with the same experimental setup.**

Keywords- Mel-frequency cepstral coefficients, modulation spectral analysis, music genre classification, normalized audio spectrum envelope, octave-based spectral contrast.

I. INTRODUCTION

The music genre classification problem is defined as genre labeling of music tracks. In general, automatic music genre classification plays an important and preliminary role in a music organization or music retrieval system. A new album or music track can be assigned to a proper genre in order to place it in the appropriate section of an online music store or music database. Thus, a number of supervised classification techniques have been developed for automatic classification of unlabeled music tracks [1-9].

To determine the music genre of a music track, some discriminating audio features have to be extracted through content-based analysis of the music signal. In general, short-term feature is first computed for each short-time frame. Then, the short-term features extracted from several consecutive frames are aggregated to from long-term features. Short-term features, typically describing the timbral characteristics of audio signals, are usually extracted from every short time window (or frame) during which the audio signal is assumed to be stationary. The timbral characteristics generally exhibit the properties related to instrumentations or sound sources such as music, speech, or environment sounds. The most widely used timbral features include zero crossing rate (ZCR), spectral centroid, spectral bandwidth, spectral flux, spectral rolloff, Mel-frequency cepstral coefficients (MFCC), discrete wavelet transform coefficients [2, 10], octave-based spectral contrast (OSC) [3, 4], MPEG-7 normalized audio spectrum envelope (NASE) [11], etc.

Generally, music genres not only correspond to the timbre of the music but also to the temporal structure of the music. That is, the time evolution of music signals will provide some useful information for music genre discrimination. To characterize the temporal evolution of a music track, long-term features can be generated by aggregating the short-term features extracted from several consecutive frames within a time window. The methods developed for aggregating temporal features include statistical moments [1, 5, 8, 12], entropy or correlation [12, 13], nonlinear time series analysis [12], autoregressive (AR) models or multivariate autoregressive (MAR) models [7], modulation spectral analysis [5, 8, 9, 12], etc

Once the features are extracted from a music track, a classifier will be employed to determine the music genre of the given music track. Several learning techniques, such as K-nearest neighbor (KNN) [1, 2], linear discriminant analysis (LDA) [2], Gaussian mixture models (GMM) [1, 2, 4], hidden Markov models (HMM) [10], Adaboost [14], and support vector machines (SVM) [2, 15], have been employed for audio classification.

In this paper, modulation spectral analysis [16] of MFCC [17], OSC [3, 4] and MPEG-7 NASE [11] will be employed to characterize the time-varying behavior of music signals. A modulation spectrogram corresponding to the collection of modulation spectra of MFCC/OSC/NASE will be constructed. The modulation spectrum is then decomposed into several logarithmically spaced modulation subbands. For each modulation subband, a new set of modulation spectral features will be computed for music genre classification: modulation spectral peak, modulation spectral valley, modulation spectral energy, modulation spectral centroid, and modulation spectral flatness.

II. PROPOSED MUSIC GENRE CLASSIFICATION SYSTEM

The proposed music genre classification system consists of two phases: the training phase and the classification phase.

This research was supported in part by the National Science Council of R.O.C. under contract NSC-99-2221-E-216-048.

978-1-4244-9306-7/11/$26.00 ©2011 IEEE

The training phase is composed of four main modules: modulation spectral feature extraction, principal component analysis (PCA) [18, 19], multiple prototype vectors generation, and linear discriminant analysis (LDA) [18, 19]. The classification phase consists of four modules: modulation sepctral feature extraction, PCA transformation, LDA transformation, and classification. A detailed description of each module will be described below.

A. Modulation Spectral Feature Extraction

1) Frame-based feature extraction: In this paper, the frame based feature vectors used to describe an audio frame include MFCC, OSC, and NASE. The feature vectors used to represent the *t*-th audio frame can be summarized as follows:

$$\mathbf{x}_t^{MFCC} = [MFCC_t(0), MFCC_t(1), \cdots, MFCC_t(L\text{-}1)]^{\mathrm{T}} \quad (1)$$

$$\mathbf{x}_t^{OSC} = [OSC_t(0), OSC_t(1), \cdots, OSC_t(2B_O\text{-}1)]^{\mathrm{T}} \quad (2)$$

$$\mathbf{x}_t^{NASE} = [R, NASE_t(0), NASE_t(1), \cdots, NASE_t(B_N+1)]^{\mathrm{T}} \quad (3)$$

where L is the length of MFCC feature vector, B_O is the number of octave scale filters, B_N is the number of logarithmic subbands and R is the RMS-norm gain value computed from the audio spectral envelope $\mathrm{ASE}_{db}(b)$ of all subbands:

$$R = \sqrt{\sum_{b=0}^{B_N+1} (ASE_{dB}(b))^2} \quad (4)$$

2) Modulation spectral analysis: The frame-based features can't characterize the variations of a sound within a long-time analysis window. In this study, we will apply long-term modulation spectral analysis to MFCC, OSC, and NASE to capture the time-varying behavior of the music signals.

Without loss of generality, let $\mathbf{x}_n = [x_n(1), x_n(2), \ldots, x_n(D)]^{\mathrm{T}}$ denote the feature vector extracted from the *n*-th audio frame of a music signal, where D is the length of the feature vector. The feature vector $\mathbf{x}_n$ can be the frame based MFCC/OSC/NASE feature vector, or a combination of these feature vectors by concatenating them together. By applying FFT on each feature value along the time trajectory within a texture window of length W, we can get the modulation spectrogram:

$$M_t(m,d) = \sum_{n=0}^{W-1} x_{(t\times W/2)+n}(d) e^{-j2\pi\frac{n}{W}m}, 0 \le m < W, 0 \le d < D \quad (5)$$

where $M_t(m, d)$ is the modulation spectrogram for the *t*-th texture window, m is the modulation frequency index. In this study, the window length W is 512 with 50% overlapping between two neighboring texture windows. By time averaging the magnitude modulation spectrograms of all texture windows, the representative modulation spectrogram of a music track can be derived as follows:

$$\overline{M}(m,d) = \frac{1}{T}\sum_{t=1}^{T} |M_t(m,d)|, \quad 0 \le m < W, 0 \le d < D \quad (6)$$

where T is the total number of texture windows in the music track.

3) Modulation spectral feature extraction: The averaged modulation spectrum of each feature value will be decomposed into J logarithmically spaced modulation subbands (in this paper, J = 8), Table I shows the frequency interval of each modulation subband. For each feature value, modulation spectral contrast (MSC) [9], modulation spectral valley (MSV) [9], as well as modulation spectral energy (MSE), modulation spectral centroid (MSCEN), and modulation spectral flatness (MSF) within each modulation subband are then evaluated:

$$MSP(j,d) = \max_{\Phi_{j,l} \le m < \Phi_{j,h}} \left(\overline{M}(m,d)\right) \quad (7)$$

$$MSV(j,d) = \min_{\Phi_{j,l} \le m < \Phi_{j,h}} \left(\overline{M}(m,d)\right) \quad (8)$$

$$MSE(j,d) = 10\log_{10}\left(1 + \sum_{m=\Phi_{j,l}}^{\Phi_{j,h}} (\overline{M}(m,d))^2\right) \quad (9)$$

$$MSCEN(j,d) = \frac{\sum_{m=\Phi_{j,l}}^{\Phi_{j,h}} \overline{M}(m,d) \times m}{\sum_{m=\Phi_{j,l}}^{\Phi_{j,h}} \overline{M}(m,d)} \quad (10)$$

$$MSF(j,d) = \frac{\sqrt[\Phi_{j,h}-\Phi_{j,l}+1]{\prod_{m=\Phi_{j,l}}^{\Phi_{j,h}} \overline{M}(m,d)}}{\frac{1}{\Phi_{j,h}-\Phi_{j,l}+1}\sum_{m=\Phi_{j,l}}^{\Phi_{j,h}} \overline{M}(m,d)} \quad (11)$$

where $\Phi_{j,l}$ and $\Phi_{j,h}$ are respectively the low modulation frequency index and high modulation frequency index of the *j*-th modulation subband, $0 \le j < J$. The MSPs correspond to the dominant rhythmic components, MSVs the non-rhythmic components, MSEs express the power of each modulation subband, MSCENs indicate the mass center of each modulation subband, and MSFs represent the modulation frequency distribution within a modulation subband. Further, the difference between MSP and MSV will reflect the modulation spectral contrast distribution:

$$MSC(j,d) = MSP(j,d) - MSV(j,d) \quad (12)$$

TABLE I. FREQUENCY RANGE OF EACH MODULATION SUBBAND.

Modulation subband index	Modulation frequency index range	Modulation frequency range (Hz)
0	[0, 3)	[0, 0.5)
1	[3, 6)	[0.5, 0.1)
2	[6, 12)	[1, 2)
3	[12, 24)	[2, 4)
4	[24 48)	[4, 8)
5	[48, 96)	[8, 16)
6	[96 192)	[16, 32)
7	[192, 256)	[32, 42.24)

As a result, all MSCs (MSVs, MSEs, MSCENs or MSFs) will form a $D \times J$ matrix. To derive a compact feature vector, the mean and standard deviation along each row (and each column) of the MSC, MSV, MSE, MSCEN, and MSF matrices will be computed. Let the modulation spectral feature values derived from the *d*-th ($0 \le d < D$) row of the MSC matrix be notated $u_{MSC}^{row}(d)$ and $\sigma_{MSC}^{row}(d)$. Thus, for a music track the modulation spectral feature vector derived from the D rows of

the_MSC matrix is of size 2D and can be represented as follows:

$$\mathbf{f}_{MSC}^{row} = [u_{MSC}^{row}(0), \sigma_{MSC}^{row}(0), \cdots, u_{MSC}^{row}(D-1), \sigma_{MSC}^{row}(D-1)]^{\mathrm{T}} \quad (13)$$

Similarly, the modulation spectral feature values can also be derived from each column of the MSC modulation feature matrix. Thus, the modulation spectral feature vector derived from the J columns of the MSC matrix can be represented as follows:

$$\mathbf{f}_{MSC}^{col} = [u_{MSC}^{col}(0), \sigma_{MSC}^{col}(0), \cdots, u_{MSC}^{col}(J-1), \sigma_{MSC}^{col}(J-1)]^{\mathrm{T}} \quad (14)$$

In this paper, these two modulation spectral feature vectors, $\mathbf{f}_{MSC}^{row}$ and $\mathbf{f}_{MSC}^{col}$, are concatenated together to yield the MSC modulation spectral feature vector of a music track, which is of size (2D+2J):

$$\mathbf{f}_{MSC} = [(\mathbf{f}_{MSC}^{row})^{\mathrm{T}}, (\mathbf{f}_{MSC}^{col})^{\mathrm{T}}]^{\mathrm{T}} \quad (15)$$

Similarly, the modulation spectral feature vectors derived from the MSV, MSE, MSCEN, and MSF matrices can be represented as follows:

$$\mathbf{f}_{MSV} = [(\mathbf{f}_{MSV}^{row})^{\mathrm{T}}, (\mathbf{f}_{MSV}^{col})^{\mathrm{T}}]^{\mathrm{T}} \quad (16)$$

$$\mathbf{f}_{MSE} = [(\mathbf{f}_{MSE}^{row})^{\mathrm{T}}, (\mathbf{f}_{MSE}^{col})^{\mathrm{T}}]^{\mathrm{T}} \quad (17)$$

$$\mathbf{f}_{MSCEN} = [(\mathbf{f}_{MSCEN}^{row})^{\mathrm{T}}, (\mathbf{f}_{MSCEN}^{col})^{\mathrm{T}}]^{\mathrm{T}} \quad (18)$$

$$\mathbf{f}_{MSF} = [(\mathbf{f}_{MSF}^{row})^{\mathrm{T}}, (\mathbf{f}_{MSF}^{col})^{\mathrm{T}}]^{\mathrm{T}} \quad (19)$$

4) Feature vector normalization: Since the dispersion is not identical for each feature value, a linear normalization will be employed to make the range of each feature value between 0 and 1:

$$F(m) = \frac{f(m) - f_{\min}(m)}{f_{\max}(m) - f_{\min}(m)} \quad (20)$$

where $F(m)$ denotes the normalized m-th feature value, $f_{\max}(m)$ and $f_{\min}(m)$ denote respectively the maximum and minimum of the m-th feature values of all training music tracks.

B. Principal Component Analysis (PCA)

PCA has been a widely used technique for dimensionality reduction [18, 19]. PCA is defined as the orthogonal projection of the data onto a lower dimensional vector space such that the variance of the projected data is maximized. First, the K-dimensional mean vector and $K \times K$ covariance matrix are computed for the set of K-dimensional training vectors $\mathbf{X} = \{\mathbf{x}_j, j = 1, \ldots, N\}$:

$$\boldsymbol{\mu} = \frac{1}{N}\sum_{j=1}^{N}\mathbf{x}_j \quad (21)$$

$$\sum = \frac{1}{N}\sum_{j=1}^{N}(\mathbf{x}_j - \boldsymbol{\mu})(\mathbf{x}_j - \boldsymbol{\mu})^{\mathrm{T}} \quad (22)$$

Second, the eigenvalues and corresponding eigenvectors of the covariance matrix are computed and sorted in a decreasing order of the eigenvalues. Let the eigenvector $\mathbf{v}_i$ be associated with eigenvalue λ_i, $1 \le i \le D$. The first d eigenvectors having the largest eigenvalues will form the columns of the $K \times d$ transformation matrix $\mathbf{A}_{PCA}$:

$$\mathbf{A}_{PCA} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_d] \quad (23)$$

The number of selected eigenvectors d can be determined by finding the minimum integer that satisfies the following criterion:

$$\sum_{j=1}^{d}\lambda_j \ge \alpha_{PCA}\sum_{j=1}^{D}\lambda_j, \quad (24)$$

where α_{PCA} determines how many percentage of information need to be preserved. The projected vector can be computed according to the transformation matrix $\mathbf{A}_{PCA}$:

$$\mathbf{x}_{PCA} = \mathbf{A}_{PCA}^{\mathbf{T}}(\mathbf{x} - \boldsymbol{\mu}) \quad (25)$$

C. Multiple Prototype Vectors Generation

In general, the characteristics of the music tracks of the same genre might differ significantly. That is, the feature vectors extracted from the music tracks of identical music genre will reveal many isolated manifolds in the feature space. As a result, modeling each music genre with a single feature vector is bound to fail. A better approach that takes into account such a situation is to model each music genre with a number of representative prototype vectors. These prototype vectors can be obtained by classifying all music tracks derived from identical music genre into some subclasses such that music tracks with similar feature vectors are clustered together. In this paper, the c-means clustering algorithm [18, 19] will be used to automatically divide all training feature vectors belonging to identical music genre into some subclasses. Each subclass consists of several similar feature vectors and its prototype vector is defined as their mean vectors.

Let C denote the total number of music genre in the database. In this paper, the c-means clustering algorithm will classify each music genre into K subclasses. Thus, the total number of subclass is $N_{sc} = K \times C$.

D. Linear Discriminant Analysis (LDA)

Linear discriminant analysis (LDA) [18, 19] aims at improving the classification accuracy at a lower dimensional feature vector space. LDA deals with the discrimination between various classes rather than the representation of all classes. The objective of LDA is to minimize the within-class distance while maximize the between-class distance. In LDA, an optimal transformation matrix that maps an H-dimensional feature space to an h-dimensional space ($h \le H$) has to be found in order to provide higher discriminability among various music classes.

Let $\mathbf{S_W}$ and $\mathbf{S_B}$ denote the within-class scatter matrix and between-class scatter matrix, respectively. The within-class scatter matrix is defined as:

$$\mathbf{S_W} = \sum_{c=1}^{N_{sc}}\sum_{n=1}^{N_c}(\mathbf{x}_{c,n} - \bar{\mathbf{x}}_c)(\mathbf{x}_{c,n} - \bar{\mathbf{x}}_c)^{\mathrm{T}}, \quad (26)$$

where $\mathbf{x}_{c,n}$ is the n-th feature vector labeled as class c, $\bar{\mathbf{x}}_c$ is the mean vector of class c, N_{sc} is the total number of music

subclasses, and N_c is the number of training vectors labeled as subclasses c. The between-class scatter matrix is given by:

$$\mathbf{S_B} = \sum_{c=1}^{N_{sc}} N_c(\overline{\mathbf{x}}_c - \overline{\mathbf{x}})(\overline{\mathbf{x}}_c - \overline{\mathbf{x}})^{\mathrm{T}}, \quad (27)$$

where $\overline{\mathbf{x}}$ is the mean vector of all training vectors. The most widely used transformation matrix is a linear mapping that maximizes the so-called Fisher criterion J_F defined as the ratio of between-class scatter to within-class scatter:

$$J_F(\mathbf{A}) = tr((\mathbf{A}^{\mathrm{T}}\mathbf{S_W}\mathbf{A})^{-1}(\mathbf{A}^{\mathrm{T}}\mathbf{S_B}\mathbf{A})). \quad (28)$$

From the above equation, we can see that LDA tries to find a transformation matrix that maximizes the ratio of between-class scatter to within-class scatter in a lower-dimensional space. In this study, a whitening procedure is integrated with LDA transformation such that the multivariate normal distribution of the set of training vectors becomes a spherical one [19]. First, the eigenvalues and corresponding eigenvectors of $\mathbf{S_W}$ are calculated. Let $\mathbf{\Phi}$ denote the matrix whose columns are the orthonormal eigenvectors of $\mathbf{S_W}$, and $\mathbf{\Lambda}$ the diagonal matrix formed by the corresponding eigenvalues. Thus, $\mathbf{S_W\Phi} = \mathbf{\Phi\Lambda}$. Each training vector $\mathbf{x}$ is then whitening transformed by $\mathbf{\Phi\Lambda}^{-1/2}$:

$$\mathbf{x}^w = (\mathbf{\Phi\Lambda}^{-1/2})^{\mathrm{T}}\mathbf{x} \quad (29)$$

It can be shown that the whitened within-class scatter matrix $\mathbf{S}_{\mathbf{W}}^{w} = (\mathbf{\Phi\Lambda}^{-1/2})^{\mathrm{T}}\mathbf{S_W}(\mathbf{\Phi\Lambda}^{-1/2})$ derived from all the whitened training vectors will become an identity matrix $\mathbf{I}$. Thus, the whitened between-class scatter matrix $\mathbf{S}_{\mathbf{B}}^{w} = (\mathbf{\Phi\Lambda}^{-1/2})^{\mathrm{T}}\mathbf{S_B}(\mathbf{\Phi\Lambda}^{-1/2})$ contains all the discriminative information. A transformation matrix $\mathbf{\Psi}$ can be determined by finding the eigenvectors of $\mathbf{S}_{\mathbf{B}}^{w}$. Assuming that the eigenvalues are sorted in a decreasing order, the eigenvectors corresponding to the (N_{sc}–1) largest eigenvalues will form the column vectors of the transformation matrix $\mathbf{\Psi}$. Finally, the optimal whitened LDA transformation matrix $\mathbf{A}_{WLDA}$ is defined as:

$$\mathbf{A}_{WLDA} = \mathbf{\Phi\Lambda}^{-1/2}\mathbf{\Psi} \quad (30)$$

$\mathbf{A}_{WLDA}$ will be employed to transform each H-dimensional feature vector to be a lower h-dimensional vector. Let $\mathbf{x}$ denote the H-dimensional feature vector, the reduced h-dimensional feature vector can be computed by:

$$\mathbf{y} = \mathbf{A}_{WLDA}^{\mathrm{T}}\mathbf{x}_{PCA} \quad (31)$$

E. Music Genre Classification

In the classification phase, let $\mathbf{y}_{\mathrm{MMFCC}}$, $\mathbf{y}_{\mathrm{MOSC}}$, $\mathbf{y}_{\mathrm{MNASE}}$ respectively denote the modulation spectral feature vectors extracted from MFCC, OSC, and NASE modulation spectrograms. At the stage of feature level fusion, a new combined feature vector $\mathbf{y}_{\mathrm{MCOMB}}$ is obtained by concatenating $\mathbf{y}_{\mathrm{MMFCC}}$, $\mathbf{y}_{\mathrm{MOSC}}$, and $\mathbf{y}_{\mathrm{MNASE}}$ together:

$$\mathbf{y}_{\mathrm{MCOMB}} = [\mathbf{y}_{\mathrm{MMFCC}}^{\mathrm{T}}, \mathbf{y}_{\mathrm{MOSC}}^{\mathrm{T}}, \mathbf{y}_{\mathrm{MNASE}}^{\mathrm{T}}]^{\mathrm{T}} \quad (32)$$

The same linear normalization using (20) is applied to each feature value. Each type of normalized feature vector is then transformed to be a lower-dimensional feature vector by using PCA transformation matrix $\mathbf{A}_{PCA}$, and LDA transformation matrix $\mathbf{A}_{WLDA}$. The classifier is then employed to compute the distances between the transformed feature vector and the representative feature vectors of all music classes. The distance between the input music track and the c-th music genre in terms of modulation MFCC feature is defined as follows:

$$d_{MMFCC}(c) = \min_{1\le i\le C, 1\le k\le K} d(\mathbf{y}_{MMFCC}, \mathbf{y}_{MMFCC}(i,k)) \quad (33)$$

where $\mathbf{y}_{MMFCC}$ and $\mathbf{y}_{MMFCC}(i,k)$ are the modulation MFCC feature vectors of the input music track and the k-th prototype vector of the i-th music genre, respectively. The distance between the input music track and every music genre in terms of modulation OSC, NASE, and combined feature (denoted by $d_{\mathrm{MOSC}}(c)$, $d_{\mathrm{MNASE}}(c)$, and $d_{\mathrm{MCOMB}}(c)$) can be computed in a similar way. The overall distance evaluated for the c-th ($1 \le c \le C$) music genre is defined as the sum of each individual distance [20]:

$$d(c) = d_{MMFCC}(c) + d_{MOSC}(c) + d_{MNASE}(c) + d_{MCOMB}(c) \quad (34)$$

Thus, the subject code s that denotes the identified music genre is determined by finding the music class that has the minimum overall distance:

$$s = \arg\min_{1\le c\le C} d(c) \quad (35)$$

III. EXPERIMENTAL RESULTS

A. Datasets

The dataset used in the ISMIR2004 Music Genre Classification Contest [21] will be employed for performance comparison. This dataset consists of 1458 music tracks in which 729 music tracks are used for training and the other 729 tracks for testing. The audio file format is 44.1 kHz, 128 kbps, 16-bit, stereo MP3 files. In this study, each stereo MP3 audio file was first converted into a 44.1 kHz, 16-bit, mono audio file before classification. These music tracks are classified into six classes: Classical, Electronic, Jazz/Blue, Metal/Punk, Rock/Pop, and World. In summary, the music tracks used for training/testing include 320/320 tracks of Classical, 115/114 tracks of Electronic, 26/26 tracks of Jazz/Blue, 45/45 tracks of Metal/Punk, 101/102 tracks of Rock/Pop, and 122/122 tracks of World music genre. Since the music tracks per class are not equally distributed, the overall classification accuracy is defined as follows:

$$CA = \sum_{1\le c\le C} P_c \times CA_c, \quad (36)$$

where P_c is the probability of appearance of the c-th music genre, CA_c is the classification accuracy for the c-th music genre.

B. Classification Results

Table II compares the classification accuracy of different modulation spectral feature vectors derived from modulation spectral analysis of MFCC, OSC, and NASE: MMFCC, MOSC, MNASE, and their combination MCOMB using

different number of prototype vectors and LDA as the classifier. The results indicated that the best result is obtained when all feature vectors (MMFCC, MOSC, and MNASE) as well as the concatenated feature vector MCOMB are integrated with a classification accuracy is 89.99% when the PCA threshold $\alpha_{PCA} = 0.99$.

Table III shows the comparison with the results from the ISMIR2004 Music Genre Classification Contest as well as other approaches [8, 9, 22] with the same experimental setup. From this table, we can see that our proposed approach performs the best and achieves higher classification accuracy (89.99%) than the winner of the contest with a classification accuracy of 84.07%.

TABLE II. CLASSIFICATION ACCURACY (%) OF DIFFERENT MODULATION SPECTRAL FEATURES USING DIFFERENT NUMBER OF PROTOTYPE VECTORS (K) AND LDA AS THE CLASSIFIER.

α_{PCA}	Feature Set	*K*							
		2	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0.98	MMFCC	82.44	83.95	82.58	84.77	83.68	84.36	84.50	84.09
	MOSC	80.93	82.99	81.07	82.99	82.99	82.58	83.13	83.13
	MNASE	81.34	82.30	82.72	83.40	82.85	83.68	83.54	82.72
	MCOMB	87.93	87.79	88.61	88.20	87.93	88.61	88.34	87.79
	MMFCC+MOSC+ MNASE+MCOMB	88.89	88.89	89.30	89.03	89.30	**89.71**	89.16	89.44
0.99	MMFCC	83.68	83.54	83.68	85.19	84.22	85.60	85.46	84.09
	MOSC	83.26	83.68	83.54	82.72	83.13	84.36	84.50	85.46
	MNASE	81.89	82.72	83.40	84.22	84.09	83.95	84.50	83.81
	MCOMB	88.20	87.24	88.48	88.48	87.79	88.20	88.75	87.79
	MMFCC+MOSC+ MNASE+MCOMB	88.34	88.61	89.03	89.16	89.03	89.03	89.85	**89.99**

TABLE III. COMPARISON WITH THE RESULTS FROM THE *ISMIR2004 MUSIC GENRE CLASSIFICATION CONTEST* AND APPROACHES WITH THE SAME EXPERIMENTAL SETUP (50:50 TRAINING/TESTING SET SPLIT).

References	***CA***
Our proposed approach (MMFCC+MOSC+MNASE+MCOMB)	89.99%
C. H. Lee *et al.* [9]	86.83%
Y. Song *et al.* [22]	84.77%
T. Lidy & A. Rauber [8]	79.70%
E. Pampalk (winner)	84.07%
K. West (2nd rank)	78.33%
G. Tzanetakis (3rd rank)	71.33%
T. Lidy & A. Rauber (4th rank)	70.37%
D. Ellis & B. Whitman (5th rank)	64.00%

IV. CONCLUSIONS

A new set of modulation spectrum feature sets derived from long-term modulation spectral analysis of MFCC, OSC, and NASE features is proposed for music genre classification. Experiments conducted on ISMIR 2004 music dataset have shown that our proposed approach can achieve higher classification accuracy than other approaches with the same experimental setup.

REFERENCES

[1] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals", *IEEE Trans. on Speech and Audio Processing*, vol. 10, no. 3, pp. 293-302, July 2002.

[2] T Li, M. Ogihara, and Q. Li, "A Comparative study on content-based music genre classification", in *Proc. ACM Conf. on Research and Development in Information Retrieval*, 2003, pp. 282-289.

[3] D. N. Jiang, L. Lu, H. J. Zhang, J. H. Tao, and L. H. Cai, "Music type classification by spectral contrast feature", in *Proc. IEEE Int. Conf. on Multimedia & Expo*, vol. 1, 2002, pp. 113-116.

[4] K. West and S. Cox, "Features and classifiers for the automatic classification of musical audio signals", in *Proc. Int. Conf. on Music Information Retrieval*, 2004.

[5] M. F. McKinney and J. Breebaart, "Features for audio and music classification", in *Proc. 4th Int. Conf. on Music Information Retrieval*, 2003, pp. 151-158.

[6] J. J. Aucouturier and F. Pachet, "Representing music genres: a state of the art", *Journal of New Music Research*, vol. 32, no. 1, pp. 83-93, 2003.

[7] A. Meng, P. Ahrendt, J. Larsen, and L. K. Hansen, "Temporal feature integration for music genre classification", *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1654-1664, July 2007.

[8] T. Lidy and A. Rauber, "Evaluation of feature extractors and psycho-acoustic transformations for music genre classification", in *Proc. 6th Int. Conf. on Music Information Retrieval*, 2005, pp. 34-41.

[9] C. H. Lee, J. L. Shih, K. M. Yu, and H. S. Lin, "Automatic Music Genre Classification Based on Modulation Spectral Analysis of Spectral and Cepstral Features", *IEEE Trans. on Multimedia*, vol. 11, no. 4, pp. 670-682, June 2009.

[10] C. C. Lin, S. H. Chen, T. K. Truong, and Y. Chang, "Audio classification and categorization based on wavelets and support vector machine", *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 5, pp. 644-651, Sep. 2005.

[11] H. G. Kim, N. Moreau, and T. Sikora, "Audio classification based on MPEG-7 spectral basis representation", *IEEE Trans. On Circuits and Systems for Video Technology*, vol. 14, no. 5, pp. 716-725, May 2004.

[12] F. Mörchen, A. Ultsch, M. Thies and I. Löhken, "Modeling timbre distance with temporal statistics from polyphonic music", *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 81-90, Jan. 2006.

[13] T. Lambrou, P. Kudumakis, R. Speller, M. Sandler, and A. Linney, "Classification of audio signals using statistical features on time and wavelet transform domains", in *Proc.IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 6, 1998, pp. 3621-3624.

[14] J. Bergatra, N. Casagrande, D. Erhan, D. Eck, and B. Kégl, "Aggregate features and Adaboost for music classification", *Machine Learning*, vol. 65, no. 2-3, pp. 473-484, Jun. 2006.

[15] C. Xu, N. C. Maddage, and X. Shao, "Automatic music classification and summarization", *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 3, pp. 441-450, May 2005.

[16] S. Sukittanon, L. E. Atlas, and J. W. Pitton, "Modulation-scale analysis for content identification", *IEEE Trans. on Signal Processing*, vol. 52, no. 10, pp. 3023-3035, Oct. 2004.

[17] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, 1993.

[18] R. Duda, P. Hart, and D. Stork, *Pattern Classification*, New York: Wiley, 2000.

[19] K. Fukunaga, *Statistical Pattern Recognition*, Academic Press, 1990.

[20] J. Kittler, M. Hatef, R. Duin, and J. Matas, "On combining classifiers", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 226-239, Mar. 1998.

[21] http://ismir2004.ismir.net/ISMIR_Contest.html.

[22] Y. Song and C. Zhang, "Content-based information fusion for semi-supervised music genre classification", *IEEE Trans. on Multimedia*, vol. 10, no. 1, pp. 145-152, Jan. 2008.

Date: Thu, 25 Aug 2011 13:34:59 +0000
From: PC Chair <cisp-bmei@dhu.edu.cn>
To: chlee@chu.edu.tw
Subject: CISP'11-BMEI'11 P2543 Acceptance Notification

Dear Chang-Hsing Lee,

Paper ID : P2543
Paper Title : Music Genre Classification Using Modulation Spectral Features and Multiple Prototype Vectors Representation

(All Chinese characters in this email are intended for authors from China's mainland only. 请浏览会议网站上的中文注册和终稿上传信息。)

Congratulations! We are pleased to inform you that your above paper has been accepted for presentation at **the 4th International Conference on Image and Signal Processing** (**CISP'11**) to be held from 15-17 October 2011, in Shanghai, China. After you complete the requirements below, your paper will appear in conference proceedings and will be indexed by both EI Compendex and ISTP, as well as the IEEE Xplore (IEEE Conference Record Number for CISP'11: 18205; IEEE Conference Record Number for BMEI'11: 18206. CISP IEEE Catalog Number: CFP1194D-CDR, ISBN: 978-1-4244-9305-0; BMEI IEEE Catalog Number: CFP1193D-CDR, ISBN: 978-1-4244-9350-0. CISP-BMEI 2008-2010 papers have already been indexed in EI Compendex). Substantially extended versions of best papers will be considered for publication in a CISP'11-BMEI'11 special issue of the Computers and Electrical Engineering journal (SCI-indexed). Only registered papers will be considered for the SCI journal special issue and only the selected authors will be notified by 30 October 2011.

The conference will feature world-renowned keynote speakers: Thanos Stouraitis, President-Elect of IEEE Circuits and Systems Society; Seong-Whan Lee, Hyundai-Kia Motor Chair Professor; Metin Akay, Fellow of the Institute of Physics (IOP), the American Institute of Medical Biological Engineering (AIMBE) and the American Association for the Advancement of Science (AAAS); Yuan-Ting Zhang, Editor-in-Chief for IEEE Trans. on Information Technology in Biomedicine (all Fellow of the IEEE).

In order for your paper to be included in the proceedings indexed by Ei Compendex/ISTP, it is important that you closely follow each and every instruction below, as **the acceptance is conditional on your accurate and timely reactions** :

1. Revise your paper, appropriately addressing the reviewer comments (at the end of this email, if any) which are intended to help you improve your paper for final publication. If any review comments seem vague, please revise your paper according to your best understanding.

2. Strictly follow the IEEE format requirements; incorrectly formatted papers cannot be included in the proceedings. Please refer to the conference website disable_http://cisp-bmei.dhu.edu.cn/ for detailed formatting instructions and templates. Some of the formatting instructions are given below. Closely follow the instructions at the conference website (Final Submission page) to convert your paper to IEEE Xplore-compliant pdf file using PDF eXpress and upload your final camera-ready full paper as soon as possible, but latest by **8 September 2011.** Please ensure that all formulas, figures and embedded objects in your file are error-free. It is crucial to make sure your pdf file is IEEE Xplore-compliant using PDF eXpress. Otherwise your paper may not be included in the IEEE Xplore or indexed in EI/ISTP. Please submit your final paper, IEEE Copyright Form, Registration Form, and Payment Confirmation by clicking "Upload Final Paper" at the conference submission system. In addition, please click "Edit Submission" at the conference submission system to ensure that all paper information are accurate, including the paper title and all author names, emails, and affiliations. This step is very important, since the same author and paper information will appear in the proceedings and indexing.

3. Please download IEEE Copyright Form at the conference website, complete the form, sign it, and upload a scanned form to us. In the Copyright Form, you will need to enter the conference name. Please note that your paper above is accepted in **CISP'11**, not BMEI'11. If you have more than one paper accepted, each paper may be accepted by a different conference. In addition, some papers were originally submitted to one of the two conferences (i.e., CISP'11 and BMEI'11), but were later transferred to the other conference by the Program Committee for better matches in topics.

4. Each paper must have 1 dedicated registration with full payment received by **8 September 2011** for the paper to be included in the proceedings. The registration fee is US$400 or RMB 2600 for each paper of maximum 5 pages. These payments must also be received by **8 September 2011** for the paper to be included in the proceedings.

5. You may pay with a credit card through the secure link provided by Paypal.com available at the conference website registration page.

You may also pay by telegraphic transfer to the following bank account:

For authors outside of China's mainland:	
Details of Beneficiary's Bank:	BANK OF CHINA SHANGHAI CHANG NING SUB-BRANCH
Address of Beneficiary's Bank:	2067 YAN'AN ROAD (WEST) SHANGHAI
Swift Code:	BKCHCNBJ300
Beneficiary's Name:	DONG HUA UNIVERSITY
Beneficiary's A/C No:	044175-8300-04360818091001

For authors within China's mainland (内地作者银行信息):	
收款单位：	东华大学
开户行	工行上海市松江支行
账号	1001739619000026626
地址	上海市松江区人民北路2999号东华大学信息科学与技术学院
邮编:	201620

您也可以从邮局汇款：
汇款地址：上海市松江区人民北路2999号东华大学信息科学与技术学院
收款人：刘肖燕
邮政编码：201620

请注意，如果通过银行或者邮局汇款，请您在银行转账或者邮局汇款单据的备注栏填写 **CISP'11** 和所录用的论文编号P****（未填则无法确认该论文已交费）。汇出注册费后，请把转账或者汇款单据扫描后上传到最终稿系统。如果没有收到扫描的单据，我们将视为没有收到注册费。

6. We would greatly appreciate it if you could complete the Registration Form (download at conference registration website) and upload it to us latest by **8 September 2011,** together with a scanned copy of the bank transfer (showing your Paper_IDs) if you pay by bank transfer or a copy of your credit-card payment confirmation. This is very important to help us sort out which payment is for which paper.

If any of the above requirements are not met by the deadline, your paper cannot be included in the conference proceedings or the conference program. Your kind cooperation will be greatly appreciated.

This notification serves as our Acceptance Letter. If you require a hardcopy of our Acceptance Letter or have any queries, please send an email to the Conference Secretariat CISP-BMEI@dhu.edu.cn.

Thank you for choosing CISP-BMEI conferences to present your research results and we look forward to seeing you in October 2011, Shanghai, China. We also hope that you will submit your excellent work to future CISP-BMEI conferences (further details will be announced later)

Yours sincerely,

Program Chairs, **CISP'11**

disable_http://cisp-bmei.dhu.edu.cn/

Some points to note when you format your paper:

1. Do not use double-line-spacing (with spacing between lines wide enough to fit another line). Use single-line-spacing (there should be no spaces between lines).

2. Make sure that the margins at the 4 sides are not too wide or too narrow.

3. For color figures, please make sure that the figures are legible when they are printed in black-and-white (printed proceedings will be in black-and-white).

4. In your reference list, do not add things like [J] and [C]

5. Avoid using undefined acronyms, i.e., if you wish to use an acronym, you must first define it in your paper, e.g., fuzzy neural network (FNN).

6. Please do a thorough spelling check, e.g., by using the spelling tool in Word.

Comments from Reviewer 1 :

Written english is good and authors have refered many references. The research content is appropriate to the conference.

行政院國家科學委員會補助國內專家學者出席國際學術會議報告

101 年 1 月 6 日

附件三

報告人姓名	李建興	服務機構及職稱	中華大學資訊工程系
時間 會議 地點	2010/11/4~2010/11/6 日本-福岡	本會核定 補助文號	計劃編號： NSC 99-2221-E-216-048-
會議名稱	（中文） （英文）The Fifth International Conference on Broadband and Wireless Computing, Communication (International Workshop on Intelligent Sensors and Smart Environments, ISSE)		
發表論文題目	（中文） （英文）A 3D Model Retrieval System Based On The Cylindrical Projection Descriptor		

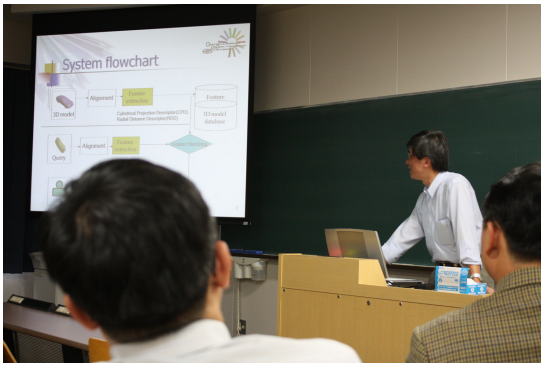
一、參加會議經過

此次會議地點位於日本九州北部的福岡，我是在會議前一天來到福岡。會議地點位於福岡西北方的福岡工業大學，必須搭 JR 線火車前往，還好學校在車站附近，沒有找尋上的困擾。此次會議同時有研討會 (BWCCA 和 3PGCIC)合辦，因此有相當多的 workshops 同時進行，但 session rooms 又分布於不同的建築物，感覺上人數比較稀疏。我所屬的 session(International Workshop on Intelligent Sensors and Smart Environments, ISSE)是在會議的第二天中午，這個 workshop 是由雲林科技大學張傳育教授所主持，雖然聆聽的人數並不多，但發問討論卻極為踴躍情。

會場-福岡工業大學

會場報到處

研討會發表論文



研討會發表論文

晚宴的地點是在福岡市區的 Hotel Centraza Hakata，採自助式方式以方便學者間的交流，參加的學者人數不少。晚宴中除了介紹工作人員、頒發論文獎及介紹下屆主辦單位地點之例行流程外，並穿插了日本傳統技藝表演。

晚宴表演

晚宴會場

二、與會心得

1. 會議中與各國學者作深切的學術交流，獲益良多。
2. 國內方面出國留學的學生日漸減少，政府與學校單位鼓勵碩博士生出國參加會議是值得稱許的。

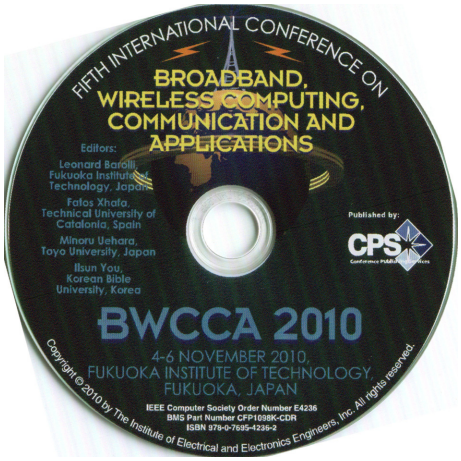
三、考察參觀活動(無是項活動者省略)

四、建議

1. 建議台灣多爭取舉辦國際學術研討會，除了可以和各國學者作廣泛之學術交流，並能促進觀光產業之發展。
2. 國內舉辦國際性研討會時，可於議程中安排半日或一日遊，讓外國學者增加認識台灣的機會，或於晚宴時安排有台灣特色的表演，如此對推展觀光也許有一些幫助。
3. 相關資訊的提供，包括會場地圖，搭車方式，飯店位置等，算是充足的，值得學習。

五、攜回資料名稱及內容

BWCCA'2010 論文光碟。



六、其他

1. 日本的 JR 火車真準時。日本人民有禮貌，街道乾淨。想要發展服務業，提昇觀光產業的台灣，這些是值得學習的。
2. 非常感謝國科會之補助得以參加該研討會。

A 3D Model Retrieval System Based On The Cylindrical Projection Descriptor

Jau-Ling Shih, Chang-Hsing Lee, Chih-Hsun Chou, Hsiang-Yuen Chang

Department of Computer Science and Information Engineering,
Chung Hua University, Hsinchu, Taiwan, R.O.C
E-mail: sjl@chu.edu.tw

ABSTRACT

In recent years, the demand for a content-based 3D model retrieval system becomes an important issue. In this paper, the cylindrical projection descriptor (CPD) will be proposed for 3D model retrieval. To derive better retrieval results, the CPD will be combined with the radial distance descriptor (RDD). The experiments are conducted on the Princeton Shape Benchmark (PSB) database. Experiment results show that our proposed method is superior to others.

1. INTRODUCTION

Recent development in advanced techniques for modeling, digitizing and visualizing 3D models has made 3D models as plentiful as images and video. Therefore, it is necessary to design a 3D model retrieval system which enables the users to efficiently and effectively search interested 3D models. The primary challenge to a content-based 3D model retrieval system is how to extract the most representative features to discriminate the shapes of various 3D models [1].

Vranic et al. applied Fourier transform to the sphere with spherical harmonics to generate embedded multi-resolution 3D shape features [2]. To be rotation invariant, pose normalization must be conducted prior to feature extraction. Therefore, Funkhouser et al. proposed a modified rotation invariant shape descriptor based on the spherical harmonics in which no pose normalization is needed [3].

Some features to represent the 3D models are based on the histograms of geometric statistics. Ankerst et al. tried to search similar 3D models using shape histograms which characterize the area of intersections of a 3D model with a collection of concentric shells and sectors [4]. The MPEG-7 shape spectrum descriptor (SSD) [5] calculates the histogram of the curvatures of all points on the 3D surface. SSD represents the distribution of geometric characteristics and is robust to tessellation of 3D polygonal models. Osada et al. [6] proposed five features, A3, D1, D2, D3, and D4, to represent 3D models by the probability distributions of geometric properties computed from a set of randomly selected points located on the surface of the model. For instance, D2, the best feature among these five features, is the distribution of distances between two random points. However, these features are invariant to tessellation of 3D polygonal models. Thus, Shih et al. [7] proposed grid D2 (GD2) to improve D2. A 3D model is first decomposed into a voxel grid. The distribution of distances between any two randomly selected valid grids is measured to represent a 3D model.

The 3D models also can be described by its 2D silhouettes from different views. Users can find similar 3D models by 2D shape features. Super and Lu [8] exploit 2D silhouette contours for 3D object recognition. Curvature and contour scale space are extracted to represent each silhouette. Chen et al. [9] proposed the LightField descriptor (LFD) to represent 3D models. The LFD is computed from 10 silhouettes. Each silhouette is represented by a 2D binary image. The Zernike moments and Fourier descriptors are employed to describe each binary image. In fact, 2D silhouettes represented by binary images can not describe the altitude information of the 3D model from different views. Shih et al. [10] proposed the elevation descriptor (ED) to represent the altitude information of a 3D model from six views. However, LFD and ED represent only the exterior shape of 3D model without capturing the interior shape information.

Kuo and Cheng [11] proposed a 3D shape retrieval system based on the principal plane analysis. First, by projecting the 3D model onto its principal plane, a 3D model can be transformed into a 2D binary image. The feature vectors are then extracted from the binary shape image. However, using only one 2D binary image can not represent a complex 3D model well. Therefore, Shih et al. [12] proposed the principal plane descriptor (PPD) to describe a 3D model with three 2D binary images by projecting it on the principal, second and third planes. The proper feature vectors can be extracted from three binary images to do 3D model retrieval.

978-0-7695-4236-2/10 $26.00 © 2010 IEEE
DOI 10.1109/BWCCA.2010.130

Novotni and Klein proposed a 3D shape retrieval method using 3D Zernike moments, which is naturally an extension of spherical harmonics based descriptors [13]. Ricard et al. [14] presented a 3D shape descriptor, the 3D Angular Radial Transform (3D-ART) for 3D model retrieval. First, the 3D models are represented in spherical coordinates. Next, a Principal Components Analysis (PCA) is applied to align the 3D models along the z-axis. Then, the 3D extension of MPEG-7's ART [15] is applied to extract feature vectors.

Mademlis *et al.* [16] decomposed 3D models into meaningful parts and an attributed graph was constructed based on the connectivity of the parts. Then, the 3D Distance Field Descriptor (3D-DFD) was computed and associated to the corresponding graph nodes for partial and global 3D model retrieval.

Papadakis *et al.* [17] proposed two shape descriptors for 3D model retrieval. The 3D model was first aligned by continuous PCA (CPCA) or normal PCA (NPCA). In CPCA, the traditional one, the principal component is analyzed based on the covariance matrix computed from the coordinate vectors of the vertices, whereas in NPCA the covariance matrix is computed from the unit normal vectors of the mesh surfaces. The spherical harmonics was then applied on the filled 3D model to extract two feature vectors from the CPCA and NPCA aligned models separately. Vranic and Saupe proposed a modified PCA which used the corresponding triangle areas as weighting factors for covariance matrix computation [18]. The directions of 20 vertices on dodecahedron and the distances computed from the center point to the farthest intersections were used as features to index similar 3D models.

Zarpalas *et al.* [19] proposed a 3D model retrieval method using 240 (12×20) 2D gray-level projection images, which are obtained by projecting a 3D model onto the 240 planes rendered from the 12 vertices of 20 icosahedrons with different radii. Features were extracted from these gray-level images and combined to improve the performance. Another 3D model retrieval system used 20 depth images rendered from the 20 vertices of a dodecahedron [20]. The depth information of a pixel in each depth image was encoded as a 5-level character. Each row (depth line) in the depth image is then represented as a sequence of depth information. Dynamic programming was then used to compute the distance between two depth line descriptors.

In this paper, the cylindrical projection descriptor (CPD) will be proposed for 3D model retrieval. To derive better retrieval results, the CPD will be combined with the radial distance descriptor (RDD) [21]. The rest of the paper is organized as follows. In Section 2, the proposed 3D model retrieval method will be described. In Section 3, gives the experimental results to show the effectiveness of the proposed features. Finally, conclusions are given in Section 4.

2. THE PROPOSED 3D MODEL RETRIEVAL METHOD

In this study, two descriptors, including the radial distance descriptor (RDD) [21] and the cylindrical projection descriptor (CPD) are used for 3D model retrieval. Before extracting the feature vectors, the 3D model is aligned according to the principal plane [12].

2.1 Radial Distance Descriptor(RDD)

The main steps for computing the radial distance descriptor [21] are described as follows:

(1) 3D model is aligned by it's the principal plane [12]. The principal plane is defined as the symmetric plane on which the sum of distance of all points projected is minimal.

(2) The bounding cube is then decomposed into a voxel grid of size 100×100×100 (see Fig. 1). A voxel located at coordinates (x, y, z) will be defined as an opaque voxel, notated as $Voxel(x, y, z) = 1$, if there is a mesh located within this voxel; otherwise, the voxel is defined as a transparent voxel, notated as $Voxel(x, y, z) = 0$. To normalize for translation and scale, the object's mass center, is moved to the point (0, 0, 0) and the average distance from non-zero voxels to the mass center is scaled to 25.

(3) Six projection planes (see Fig. 1), which describe the radial distance from the 3D model surface to the mass center (see Fig. 2), are derived to represent a 3D model. Each projection plane is represented by a gray level image in which the gray value denotes the distance from an opaque voxel to the mass center (see Fig. 3). Let the six projection planes be notated as I_k, $k = 1, 2, \ldots, 6$. Then ,the gray value of each pixel on these images is defined as follows:

$$I_1(x,z) = \max_{1\le y\le 50}(\mathrm{R}(x,y,z)\mathrm{Voxel}(x,y,z)),$$
$$\text{for} -50 \le x,z \le 50,$$
$$I_2(x,y) = \max_{1\le z\le 50}(\mathrm{R}(x,y,z)\mathrm{Voxel}(x,y,z)),$$
$$\text{for} -50 \le x,y \le 50,$$
$$I_3(y,z) = \max_{1\le x\le 50}(\mathrm{R}(x,y,z)\mathrm{Voxel}(x,y,z)),$$
$$\text{for} -50 \le y,z \le 50,$$
$$I_4(x,z) = \max_{-50\le y\le -1}(\mathrm{R}(x,y,z)\mathrm{Voxel}(x,y,z)),$$
$$\text{for} -50 \le x,z \le 50,$$
$$I_5(x,y) = \max_{-50\le z\le -1}(\mathrm{R}(x,y,z)\mathrm{Voxel}(x,y,z)),$$
$$\text{for} -50 \le x,y \le 50,$$
$$I_6(y,z) = \max_{-50\le x\le -1}(\mathrm{R}(x,y,z)\mathrm{Voxel}(x,y,z)),$$
$$\text{for} -50 \le y,z \le 50,$$

where $\mathrm{R}(x,y,z) = \sqrt{x^2+y^2+z^2}$.

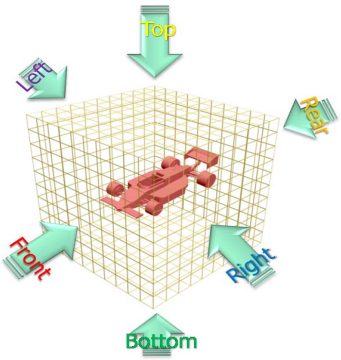


Fig. 1 The six views of 3D racing car model.

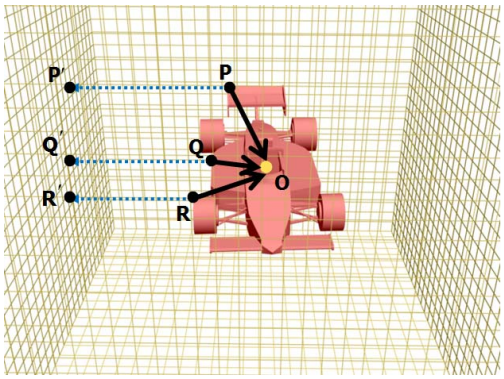
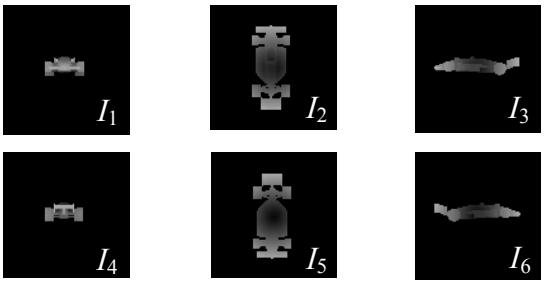


Fig. 2 The $\overline{PO}$, $\overline{QO}$, and $\overline{RO}$ represent the radial distance from the 3D model surface to the mass center O.



(a) (b) (c)

Fig. 3 3D racing car model and its six gray-level projection planes. (a) The front plane I_1 and the rear plane I_4. (b) The top plane I_2 and the bottom plane I_5. (c) The right plane I_3 and the left plane I_6.

(4) The MPEG7's angular radial transformation (ART) [15] is used to extract the feature vector from each projection plane. The ART descriptor consists of the magnitudes of all complex ART coefficients. In the MPEG-7 standard, the suggested ART descriptor consists of 35 coefficients, $\left|f_k^{\mathrm{ART}}(n,m)\right|$, for $0 \le n \le 2$ and $0 \le m \le 11$, excluding $n = 0$ and $m = 0$. In summary, the radial distance descriptor (RDD) is defined as:

$$\mathbf{rdd} = [(\mathbf{rrd}_1)^{\mathrm{T}}, (\mathbf{rrd}_2)^{\mathrm{T}}, ..., (\mathbf{rrd}_6)^{\mathrm{T}}]^{\mathrm{T}},$$

where $\mathbf{rdd}_k$, $1 \le k \le 6$, is the ART feature vector extracted from the k-th projection plane:

$$\begin{aligned}\mathbf{rdd}_k &= [\mathbf{rrd}_k(1), \mathbf{rrd}_k(2), ..., \mathbf{rrd}_k(35),]^{\mathrm{T}} \\ &= [|f_k(0,1)|, ..., |f_k(0,11)|, |f_k(1,0)|, \\ &\quad ..., |f_k(1,11)|, |f_k(2,0)|, ..., |f_k(2,11)|]^{\mathrm{T}}.\end{aligned}$$

2.2 The Cylindrical Projection Descriptor (CPD)

The main steps for computing the cylindrical projection descriptor (CPD) are described as follows:

(1) 3D model is aligned by it's the principal plane [12] as Sec 2.1.1.

(2) A cylindrical projection can unfold a portion of the surface of a sphere into a flat plane. (see Fig. 4). As shown in Fig. 5, and 6, the three gray-level images F_1, F_2, F_3, can be obtained by mapping the $\overline{PO}$ value on the flat planes by three directions: **x**, **y**, and **z**.

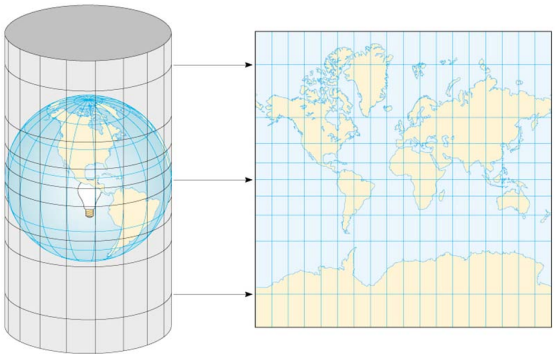


Fig. 4 The cylindrical projection.

(3) 2D-FFT (Fast Fourier Transform) is used to extract the feature vector from three 256×128 projection images (see Fig. 6). The FFT descriptor consists of the magnitudes of the first 32×32 FFT coefficients. In this paper, the FFT descriptor consists of 1024 coefficients, $\left|f_k^{\mathrm{FFT}}(u,v)\right|$, for $0 \le u \le 31$ and $0 \le v \le 31$. In summary, the cylindrical projection descriptor (CPD) is defined as:

$$\mathbf{cpd} = [(\mathbf{cpd}_1)^{\mathrm{T}}, (\mathbf{cpd}_2)^{\mathrm{T}}, (\mathbf{cpd}_3)^{\mathrm{T}}]^{\mathrm{T}}$$

where $\mathbf{cpd}_k$, $1 \le k \le 3$, is the feature vector extracted from the k-th projection.

$$\begin{aligned}\mathbf{cpd}_k &= [\mathbf{cpd}_k(1), \mathbf{cpd}_k(2), ..., \mathbf{cpd}_k(1024),]^{\mathrm{T}} \\ &= [|f_k(0,0)|, ..., |f_k(0,31)|, |f_k(1,0)|, \\ &\quad ..., |f_k(1,31)|, |f_k(31,0)|, ..., |f_k(31,31)|]^{\mathrm{T}}.\end{aligned}$$

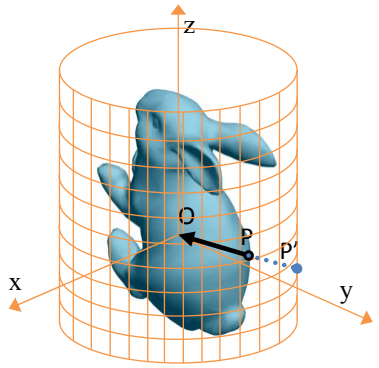


Fig. 5 The cylindrical projection descriptor. $\overline{PO}$ represent the distance from the 3D model surface to the mass.

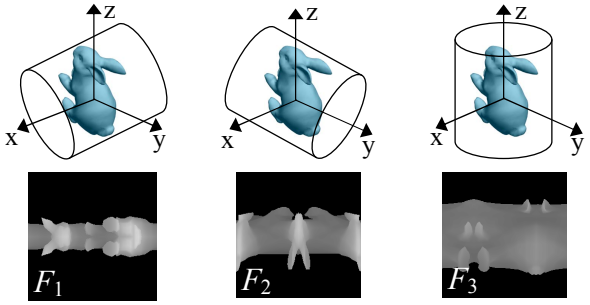


Fig. 6 The three gray-level image F_1, F_2, and F_3, are obtained by the cylindrial projection.

2.3 Distance Computation

Let $\mathbf{rdd} = [(\mathbf{rdd}_1)^{\mathrm{T}}, (\mathbf{rdd}_2)^{\mathrm{T}}, \ldots, (\mathbf{rdd}_6)^{\mathrm{T}}]^{\mathrm{T}}$ and $\mathbf{rdd}^b = [(\mathbf{rdd}_1^b)^{\mathrm{T}}, (\mathbf{rdd}_2^b)^{\mathrm{T}}, \ldots, (\mathbf{rdd}_6^b)^{\mathrm{T}}]^{\mathrm{T}}$ denote the RDD of a query model and the *b*-th matching model in the database, respectively. The distance between the query model and the *b*-th matching model is defined as follows:

$$\mathrm{Dis}_{\mathrm{RDD}}^b = \frac{1}{N_{\mathrm{RDD}}} \sum_{k=1}^{6} \left\| \mathbf{rdd}_k - \mathbf{rdd}_k^b \right\|_1 = \frac{1}{N_{\mathrm{RDD}}} \sum_{k=1}^{6} \sum_{i=1}^{36} \left| \mathrm{rdd}_k(i) - \mathrm{rdd}_k^b(i) \right|$$

where $N_{\mathrm{RDD}} = 6 \times 36$. CPD is defined as:

Let $\mathbf{cpd} = [(\mathbf{cpd}_1)^{\mathrm{T}}, (\mathbf{cpd}_2)^{\mathrm{T}}, (\mathbf{cpd}_3)^{\mathrm{T}}]^{\mathrm{T}}$ and $\mathbf{cpd} = [(\mathbf{cpd}_1^b)^{\mathrm{T}}, (\mathbf{cpd}_2^b)^{\mathrm{T}}, (\mathbf{cpd}_3^b)^{\mathrm{T}}]^{\mathrm{T}}$ denote the CPD of a query model and the *b*-th matching model in the database, respectively. The distance between the query model and the *b*-th matching model is defined as follows:

$$\mathrm{Dis}_{\mathrm{CPD}}^b = \frac{1}{N_{\mathrm{CPD}}} \sum_{k=1}^{3} \left\| \mathbf{cpd}_k - \mathbf{cpd}_k^b \right\|_1 = \frac{1}{N_{\mathrm{CPD}}} \sum_{k=1}^{3} \sum_{i=1}^{1024} \left| \mathrm{cpd}_k(i) - \mathrm{cpd}_k^b(i) \right|$$

where $N_{\mathrm{CPD}} = 3 \times 1024$. Finall we use three kinds of similarity measure methods to combine RDD and CPD:

1) $\mathrm{Sim}_1^b = \dfrac{1}{\mathrm{Dis}_{\mathrm{RDD}}^b + \mathrm{Dis}_{\mathrm{CPD}}^b}$

2) Use the Borda Count Algorithm [34] to combine the RDD and CPD:

$$\mathrm{Sim}_2^b = \frac{1}{\mathrm{Rank}_{\mathrm{RDD}}^b + \mathrm{Rank}_{\mathrm{CPD}}^b}$$

where. $Rank_{\mathrm{RDD}}^b$ and $Rank_{\mathrm{CPD}}^b$ are the retrieval rank values of the *b*-th matching model for the RDD and CPD, respectively.

3. EXPERIMENTAL RESULTS

To demonstrate the effectiveness of the proposed method for different 3D models, some experiments have been conducted on the Princeton Shape Benchmark (PSB) database [23]. The PSB database contains 1814 models (161 classes) which are divided into 907 training models (90 classes) and 907 test models (92 classes). Note that in this database the number of models is different for each class. Since the number of models in each class is different in the PSB database, the *recall* value (Re_i^j) for the *j*-th query model in the *i*-th class is defined as follows:

$$Re_i^j = N_i^j / N_i,$$

where N_i^j denotes in the retrieval list the number of models labeled as class *i* and N_i is the total number of models in class *i*. The average recall values is defined as follows:

$$Re = \frac{1}{T_S} \sum_{i=1}^{92} \sum_{j=1}^{T_i} Re_i^j$$

where $T_s = T_1 + T_2 + \ldots + T_{92}$. The Discounted Cumulative Gain (DCG) [28], will also be employed to compare the performance of different approaches. DCG at the *k*-th rank is recursively defined as follows:

$$\mathrm{DCG}_k = \begin{cases} \mathrm{DCG}_{k-1} + \dfrac{L_k}{\log_2(k)}, & k \geq 2 \\ L_1, & k = 1 \end{cases},$$

where L_k=1 if the *k*-th retrieval model and the query one belong to the same class; otherwise, L_k=0. The overall DCG score for a query model *q* is defined as $\mathrm{DCG}_{k_{\max}}$, where k_{max} is the total number of models in the database. DCG is clear that if the top-ranked models and the query one are of the same class, $\mathrm{DCG}_{k_{\max}}$ will be larger than the retrieval result with similar models appearing in the bottom of the retrieval list.

In our experimental, each model in database is presented as a query one. Table 1 compares the

retrieval results of the proposed method with other descriptor. It also shows that the combination of RDD and CPD outperforms other descriptors in terms of the average recall value and DCG. The combination of RDD and CPD using the second similarly measure, Sim_2, has the best recall and DCG values. Moreover, we compare the retrieval performance of our proposed method with another state-of-the-art descriptors in Table 2. We can also see that the proposed method outperforms these descriptors in terms of DCG.

4. CONCLUSION

With the development of computer graphics and virtual realities, the demand for a content-based 3D retrieval system becomes urgent. In this study, two features, the radial distance descriptor (RDD) and the cylindrical projection distance (CPD) are combined for 3D model retrieval. The experiments have been conducted on the Princeton Shape Benchmark (PSB) database. Experiment results show that the proposed methods are superior to others.

5. ACKNOWLEDGEMENT

This research was supported in part by the National Science Council, R.O.C. under Contract NSC 98-2221-E-216-039.

Table 1. Comparison of the proposed and other descriptors on the PSB database in terms of the *recall* value(%) and DCG(%). N_L denotes the number of retrieval models.

Method		*Re* (N_L=T_i)	*Re* (N_L=4T_i)	DCG
RDD		41.71	62.05	71.60
CPD		36.91	55.59	67.59
RDD+CPD	Sim_1	**43.53**	**62.27**	**72.05**
	Sim_2	42.75	61.75	71.15
ED[22]		35.48	56.03	67.04
AED [24]		38.61	60.29	70.29
DED[22]		36.19	55.87	66.92
CED[22]		37.32	57.80	68.04
PPD [12]		34.23	55.35	65.86
SH [3]		27.06	41.02	58.35
SSD [5]		15.87	26.64	48.07
GD2 [7]		28.30	47.61	60.91

Table 2. Comparison of the proposed method and other descriptors on the *PSB* database in terms of DCG(%). (Note that the approaches marked with * are implemented by Akgul et al. and originally appeared in [28])

Method		DCG	Method	DCG
RDD+CPD	Sim_1	**72.05**	DSR [27]*	66.50
EGI [25]		43.80	DBF [28]	65.90
CRSF [17]		66.80	DSR+DBF [28]	70.20
LF [9]		64.30	SWD [29]*	65.40
SH-GEDT [26]		58.40	SIL [27]*	59.70
DBI [27]*		66.30	3DHT [30]*	57.70
RISH [11]*		58.40	CAH [31]*	43.30
SHIST [13]*		54.50	REXT [32]*	60.10
			AVC [33]	60.20

6. REFERENCES

[1] J.W.H. Tangelder and R.C. Veltkamp, "A survey of content based 3D shape retrieval methods", Shape Modeling Applications, pp. 145-156, 2004.

[2] D.V. Vranic, D. Saupe, and J. Richter, "Tools for 3D-object retrieval: Karhunen- Loeve transform and spherical harmonics", Proceedings of IEEE Workshop on Multimedia Signal Processing, pp. 293-298, 2001.

[3] T. Funkhouser, P. Min, M. Kazhdan, J. Chen, A. Halderman, D. Dobkin, and D. Jacobs, "A search engine for 3D models", ACM Trans, Graphics 22, pp. 83-105, 2003.

[4] R. Osada, T. Funkhouser, B. Chazelle, and D. Dobkin, "Shape Distributions", ACM Trans. on Graphics, pp. 807-832, 2002.

[5] S. Manjunath, P. Salembier, and T. Sikora, "Introduction to MPEG-7 Multimedia Content Descriptor Interface", John Wiley & Sons Ltd., 2002.

[6] M. Ankerst, G. Kastenmuller, H.P. Kriegel, and T. Seidl, "3D shape histograms for similarity search and classification in spatial databases", Proceedings of 6th International Symposium on Spatial Databases (SSD'99), pp. 207-226, 1999.

[7] J.L. Shih, C.H. Lee, and J.T. Wang, "3D Object Retrieval System Based on Grid D2", Electronics Letters, pp. 23-24, 2005.

[8] B.J. Super and H. Lu, "Evaluation of a hypothesizer for silhouette-based 3-D object recognition", Pattern Recognition, pp. 69-78, 2003.

[9] D.Y. Chen, X.P. Tian, Y.T. Shen, and M. Ouhyoung, "On visual similarity based 3D model retrieval", Computer Graphics Forum, pp. 223-232, 2003.

[10] J.L. Shih, C.H. Lee, and J.T. Wang, "A New 3D Model Retrieval Approach Based on Elevation

Descriptor", Pattern Recognition, pp. 283-295, 2007.
[11] C.T. Kuo and S.C. Cheng, "3D model retrieval using principal plane analysis and dynamic programming", Pattern Recognition, pp. 742-755, 2007.
[12] J.L. Shih and W.C. Wang, "A 3D Model Retrieval Approach based on The Principal Plane Descriptor", Proceedings of The Second International Conference on Innovative Computing, Information and Control (ICICIC), pp. 59-62, 2007.
[13] M. Ankerst, G. Kastenmuller, H.P. Kriegel, and T. Seidl, "3D shape histograms for similarity search and classification in spatial databases", Proceedings of 6th International Symposium on Spatial Databases (SSD'99), pp. 207-226, 1999.
[14] J. Ricard, D. Coeurjolly and A. Baskurt, "Generalizations of angular radial transform for 2D and 3D shape retrieval", Pattern Recognition Letters, pp. 2174-2186, 2005.
[15] MPEG Video Group, "MPEG-7 Visual part of experimentation Model Version 9.0", 2001.
[16] A. Mademlis, P. Daras, A. Axenopoulos, D. Tzovaras, and M. G. Strintzis, "Combining Topological and Geometrical Features for Global and Partial 3D Shape Retrieval", IEEE Tran. on Multimedia, pp. 819-831, 2008.
[17] Panagiotis Papadakisa, Ioannis Pratikakisa, Stavros Perantonisa, Theoharis Theoharis, "Efficient 3D shape matching and retrieval using a concrete radialized spherical projection representation", Pattern Recognition, pp. 2437-2452, 2007.
[18] D. V. Vranic and D. Saupe, "3D Model Retrieval", Proceedings of the Spring Conference on Computer Graphics and its Applications (SCCG2000), pp. 89-93, 2000.
[19] Dimitrios Zarpalas, Petros Daras, Apostolos Axenopoulos, Dimitrios Tzovaras, and Michael G. Strintzis, "3D Model Search and Retrieval Using the Spherical Trace Transform," EURASIP Journal on Advances in Signal Processing, 2007.
[20] Mohamed Chaouch, Anne Verroust-Blondet, " A New Descriptor for 2D Depth Image Indexing and 3D Model Retrieval", IEEE International Conference on Image Processing, pp. 373-376, 2007.
[21] J.L. Shih, C.H. Lee and C.H Chuang, "A 3D Model Retrieval System Based On The Derivative Radial Distance", Proceedings of The 22th IPPR Conference On Computer Vision, Graphics and Image Processing (CVGIP) 2009.
[22] J.L. Shih, T.Y Huang, and Y.C. Wang, "A 3D Model Retrieval System Using the Derivative Elevation and 3D-ART", Proceedings of the IEEE Asia-Pacific Services Computing Conference, (APSCC), pp. 739-744, 2008.
[23] P. Shilane, P. Min, M. Kazhdan, T. Funkhouser, "The Princeton shape benchmark", Proceedings of Shape Modeling Applications, pp. 167-178, 2004.
[24] J. L. Shih and H. Y. Chen, "A 3D model retrieval approach using the interior and exterior 3D shape information", Multimedia Tools Applicaion., vol. 43, no. 1, pp. 45-62, May 2009.
[25] B. K. P. Horn, "Extended Gaussian images", in Proceedings of IEEE, vol. 72, no. 12, pp. 1671-1686, Dec. 1984.
[26] M. Kazhdan, T. Funkhouser, and S. Rusinkiewicz, "Rotation invariant spherical harmonic representation of 3D shape descriptors", in Proceedings of Eurographics/ACM SIGGRAPH Symposium on Geometry processing, pp. 156-164, 2003.
[27] D. V. Vranic, "3D model retrieval", Ph.D. Dissertation, University of Leipzig, Department of Computer Science, 2004.
[28] C. B. Akgul, B. Sankur, Y. Yemez, and F. Schmitt, "3D model retrieval using probability density-based shape descriptors", IEEE Transaction on Pattern Analysis and Machine Intelligence, vol. 31, no. 6, pp. 1117-1133, June, 2009.
[29] H. Laga, H. Takahashi, and M. Nakajima, "Spherical wavelet descriptors for content-based 3D model retrieval," in Proceedings of IEEE International Conference on Shape Modeling and Application (SMI'06), 2006.
[30] T. Zaharia and F. J. Preteux, "Shape-based retrieval of 3D mesh models", in Proceedings of the IEEE International Conference on Multimedia and Expo, vol. 1, pp. 437-440, 2002.
[31] E. Paquet and M. Rioux, "Nefertiti: A Query by Content Software for Three-Dimensional Models Databases Management", in Proceedings of International Conference on Recent Advances in 3D Digital Imaging and Modeling, pp. 345-352, 1997.
[32] D. V. Vranic, "An Improvement of Rotation Invariant 3D Shape Descriptor Based on Functions on Concentric Spheres", in Proceedings of IEEE International Conference on Image Processing, pp. 757-760, Sept. 2003.
[33] T. F. Ansary, M. Daoudi, and J.-P. Vandeborre, "3D Model Retrieval Based on Adaptive Views Clustering", LNCS 3687, pp. 473–483, 2005.
[34] M. Jovic, Y. Hatakeyana, F. Dong, and K. Hirota, "Image Retrieval Based on Similarity Score Fusion from Feature Similarity Ranking Lists", LNAI 4223, pp. 461-470, 2006.

日期: Wed, 23 Jun 2010 03:30:27 +0100
寄件者: ISSE-2010 <isse2010@easychair.org>
收件者: Chang-Hsing Lee <chlee@chu.edu.tw>
主旨: ISSE-2010 notification for paper 17

Dear Prof. Chang-Hsing Lee

We are pleased to inform you that your paper:
A 3D Model Retrieval System Based On The Cylindrical Projection Descriptor
has been accepted for presentation at ISSE-2010.

We would like to kindly remind you the following important issues:

Please follow EXACTLY the online Submission Guidelines (http://isse2010.yuntech.edu.tw/), provided by the Conference Publishing Services of the IEEE Computer Society and us, in the preparation of your camera-ready copies. Please upload your camera-ready copies via the online submission system (http://www.easychair.org/conferences/?conf=isse2010) by July 23, 2010. In the submission system, you can use the item [Submit a new version] to upload your camera-ready copies.

As a prerequisite of having your papers included in the proceedings, conference registration is due by July 30, 2010.

Congratulations on this fine achievement! We are looking forward to seeing you in Fukuoka in November 2010.

Sincerely,
Chien-Cheng Lee, PC chair
ISSE-2010

國科會補助計畫衍生研發成果推廣資料表

日期:2012/01/04

國科會補助計畫	計畫名稱: 調變頻譜分析於音樂曲風及樂器音色之自動分類辨識之研究
	計畫主持人: 李建興
	計畫編號: 99-2221-E-216-048- 學門領域: 圖形辨識

無研發成果推廣資料

99 年度專題研究計畫研究成果彙整表

計畫主持人：李建興	計畫編號：99-2221-E-216-048-
計畫名稱：調變頻譜分析於音樂曲風及樂器音色之自動分類辨識之研究	

成果項目			量化			單位	備註（質化說明：如數個計畫共同成果、成果列為該期刊之封面故事...等）
			實際已達成數（被接受或已發表）	預期總達成數(含實際已達成數)	本計畫實際貢獻百分比		
國內	論文著作	期刊論文	1	0	100%	篇	
		研究報告/技術報告	1	1	100%		
		研討會論文	0	0	100%		
		專書	0	0	100%		
	專利	申請中件數	0	0	100%	件	
		已獲得件數	0	0	100%		
	技術移轉	件數	0	0	100%	件	
		權利金	0	0	100%	千元	
	參與計畫人力（本國籍）	碩士生	2	2	100%	人次	
		博士生	1	1	100%		
		博士後研究員	0	0	100%		
		專任助理	0	0	100%		
國外	論文著作	期刊論文	0	0	100%	篇	
		研究報告/技術報告	0	0	100%		
		研討會論文	1	0	100%		
		專書	0	0	100%	章/本	
	專利	申請中件數	0	0	100%	件	
		已獲得件數	0	0	100%		
	技術移轉	件數	0	0	100%	件	
		權利金	0	0	100%	千元	
	參與計畫人力（外國籍）	碩士生	0	0	100%	人次	
		博士生	0	0	100%		
		博士後研究員	0	0	100%		
		專任助理	0	0	100%		

其他成果 (無法以量化表達之成果如辦理學術活動、獲得獎項、重要國際合作、研究成果國際影響力及其他協助產業技術發展之具體效益事項等，請以文字敘述填列。)	無

	成果項目	量化	名稱或內容性質簡述
科教處計畫加填項目	測驗工具(含質性與量性)	0	
	課程/模組	0	
	電腦及網路系統或工具	0	
	教材	0	
	舉辦之活動/競賽	0	
	研討會/工作坊	0	
	電子報、網站	0	
	計畫成果推廣之參與（閱聽）人數	0	

國科會補助專題研究計畫成果報告自評表

請就研究內容與原計畫相符程度、達成預期目標情況、研究成果之學術或應用價值（簡要敘述成果所代表之意義、價值、影響或進一步發展之可能性）、是否適合在學術期刊發表或申請專利、主要發現或其他有關價值等，作一綜合評估。

1. 請就研究內容與原計畫相符程度、達成預期目標情況作一綜合評估

■達成目標

□未達成目標（請說明，以 100 字為限）

□實驗失敗

□因故實驗中斷

□其他原因

說明：

2. 研究成果在學術期刊發表或申請專利等情形：

論文：■已發表 □未發表之文稿 □撰寫中 □無

專利：□已獲得 □申請中 ■無

技轉：□已技轉 □洽談中 ■無

其他：(以 100 字為限)

目前已發表之論文如下：

1. C. H. Lee et al., ’’ Automatic Music Genre Classification Using Modulation Spectral Features and Nonparametric Discriminant Analysis’’, J. of Inf. Tech. and Appl., Vol. 5, No. 2, June 2011, pp. 75-82.

2. C. H. Lee et al., ’’ Music Genre Classification Using Modulation Spectral Features and Multiple Prototype Vectors Representation’’, in Proc. CISP’11, 2011, pp. 2762-2766.

3. 請依學術成就、技術創新、社會影響等方面，評估研究成果之學術或應用價值（簡要敘述成果所代表之意義、價值、影響或進一步發展之可能性）(以 500 字為限)

本計劃完成音樂曲風之自動分類系統，能夠根據音樂的性質事先將音樂曲目分類為不同的曲風類型，有效率的管理龐大的音樂資料庫，此外也可做為音樂推薦系統使用，當使用者在選取一首喜愛的音樂時，可以將曲風相似之音樂曲目推薦給使用者，減少使用者搜尋性質相似之音樂所花的時間。