第二章 高斯混合模型(Gaussian Mixture Model)的背景建立

2.1 高斯混合模型簡介

高斯混合模型(GMM)是在背景濾除(Background Subtraction)研究上一種常用來建立背景影像模型的方法，原因是影像的像素值(pixel value) 是不會固定的，不固定的原因分成 2 大類：

(1) 移動所造成的改變：包括實際物體的移動，如被風吹動的樹葉或草皮、濺起漣漪的湖面、漂動的雲、走動的人…等，或是因為攝影機搖晃所造成像素值的變化。

(2) 亮度的改變：在靜態的影像中，即使沒有移動的物體，一樣可能因受外界環境的影響產生亮度變化，如太陽位置的變化、陰影的遮蔽、室內電燈的開關、CRT 螢幕的掃描、日光燈的閃爍…等影響。

這些改變造成像素的值在原來的值附近做小幅的變動，所以非常適合用高斯分佈去模型化背景的顏色分佈；但在很多情況下，顏色的分佈不是只在一個值附近做變動，而是在某幾個值做變動，如閃爍的湖面、閃爍的電腦螢幕、或是隨光線移動所造成陰影的改變等情況，下圖 2-1 可以清楚的看出這些情況的顏色分佈，所以採用多個高斯的分佈來模型化背景是較適合的方式。

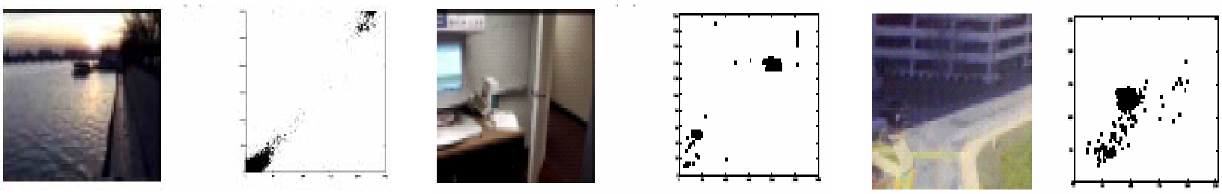

圖 2-1 常見的真實環境及其顏色(R-G)分佈圖

以數學的觀點來看，對任一具有多類別的樣本(Pattern)而言，高斯混合模型具有極佳的近似能力，與傳統的單一高斯分佈(Single Gaussian Mixture)及向量量化(Vector Quantization)兩種模型比較，單一高斯分佈模型，僅能用一個平均值向量來代表一堆樣本在向量空間的中心位置，用共變異矩陣來近似這些樣本在向量空間中所分佈的形狀，其效果當然不好。而向量量化的模型，是用幾個重要的位置來代表整個向量空間，但模型本身並沒有把這些樣本在空間中的分佈大小、形狀描述出來，因此此種方法也不理想。而高斯混合模型使用多個高斯來代表特徵向量的分佈，以數學的觀點來看，它不但精準地紀錄樣本的各種類別、在向量空間中的位置，也能描述出這些類別在空間中的大小及形狀，因此，高斯混合模型適合描述特徵向量在顏色空間的分佈。

在採用高斯混合模型時，有一點要注意，假設我們所求取的特徵向量的每一個維度在統計上是互相獨立(Statically Independent)的關係，即(R，G，B)三個顏色分佈是各自獨立的，此假設在顏色學的觀點上是合理的，所以全共變異矩陣(Full Covariance Matrix)是不需要的，對角共變異矩陣(Diagonal Covariance Matrix)的高斯分佈的線性組合，就具有描述特徵向量維度間的相關能力；做此假設的另一個原因，是可以降低計算時的複雜度，因此在本論文中，高斯混合模型的共變異矩陣皆是對角矩陣。

2.2 模型描述

一個高斯混合模型具有三個參數，分別是混合加權值(mixture weights)、平均值向量(mean vector)以及共變異矩陣(covariance matrix)，將這些參數集合起來並賦予新的符號，如下所示：

$$\lambda = \{w_i, \mu_i, \sum\nolimits_i \}, i = 1,2,..., M \tag{2.1}$$

其中 W_i 表示混合加權值，μ_i 表示平均向量，$\sum_i$ 表示共變異矩陣，

而 M 則是高斯分佈的個數，對每一個影像像素而言，都可以用 λ 來表示像素的模型。若我們的資料 $X_N = \{X_1, X_2, ..., X_n\}$ 在 D 維空間中分佈，其高斯混合模型的相似度表示如下：

$$p(x_N \mid \lambda) = \sum_{i=1}^{M} w_i g_i(x_N) \tag{2.2}$$

$$g_i(x) = \frac{1}{(2\pi)^{D/2} |\Sigma_i|^{1/2}} e^D \qquad D = -\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i) \tag{2.3}$$

其中 $g_i(x)$ 為第 i 個高斯分佈的密度函數，而混合加權值也必須滿足 $\sum_{i=1}^{M} w_i = 1$ 的條件。我們可以將高斯混合模型的架構用圖 2-2 來表示之。

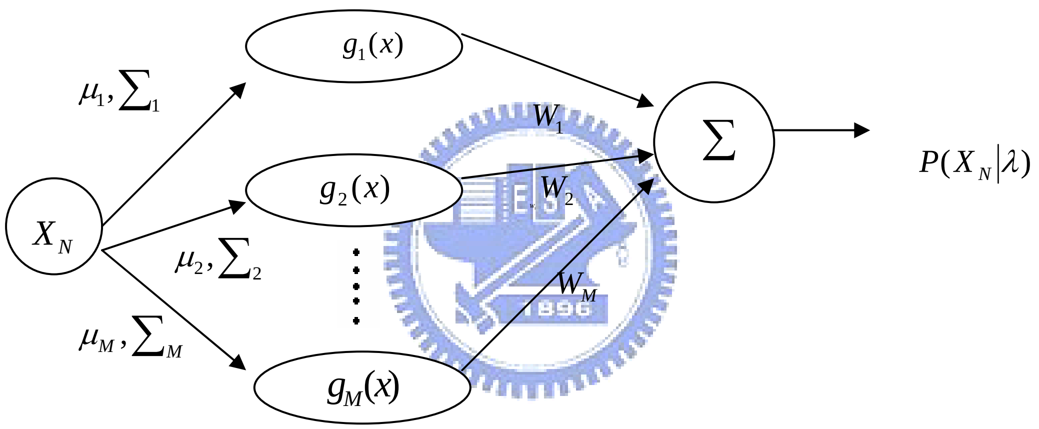

圖 2-2 高斯混合模型架構圖

從下面的圖 2-3 可以看出，左圖是某像素強度的分佈曲線，大概看出它可以用 3 個高斯分佈的組合來近似它，右圖則是近似後的分佈曲線。

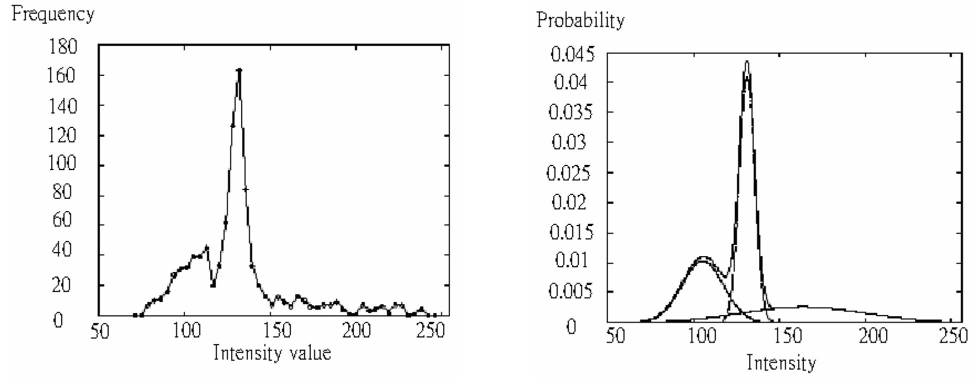

圖 2-3 強度的統計分佈及 GMM 近似的分佈

2.3 模型參數的初始化

我們如果希望快速地、精確地找出高斯混合模型的最佳參數，使得系統有最佳的表現，則在尋找最佳參數之前我們必須對參數做初使化的動作。向量量化(VQ)是一項運用非常廣泛的技術，它能將一堆特徵向量的資料，濃縮成幾個具代表性的類別(class)或群集(cluster)，所以這裡我們先採用 VQ 的技術，將我們得到的影像像素值，做初步的分群，得到高斯混合模型參數的初始化值(包括群的個數、群的中心)，以利於後面做參數的最佳化。VQ 的方法有很多種，我們採用 K 平均值分類法(K-means Cluster)，其流程如圖 2-4 所示，詳細的步驟說明如下：

0、收集資料：

經過一段時間的收集，獲得 N 個欲做訓練的特徵向量。

1、初始化：

假設一開始的群數是 K，並隨機地取 K 個向量當成每群的中心點。

2、以新的群中心來分群：

其他(N-K)個向量對這 K 個群中心做距離測量，以距離做為分群的依據，每個向量被分類到距離最短的中心。

3、更新群中心：

接著對每一群算出新的向量平均值，以此做為新的群中心。

4、判斷分群是否收斂：

將新的群中心與舊的群中心作比較，如果不再有變動，表示已收斂，則做步驟 5；反之，則重複步驟 2、3。

5、判斷是否該合併群：

如果這 K 群中，任一 2 群距離太近(可以合併成一群)，或是某一群的向量點只有一個，表示群數須減少，則群數減一(K<=K-1)，並回到步驟 1 重新分群；反之，則做步驟六。

6、得到初始化參數：

將最後分群的個數、群的中心、群的變異數以及每一群的資料個數當作高斯混合模型的初始參數(M、μ、$\sum$ 、w)。

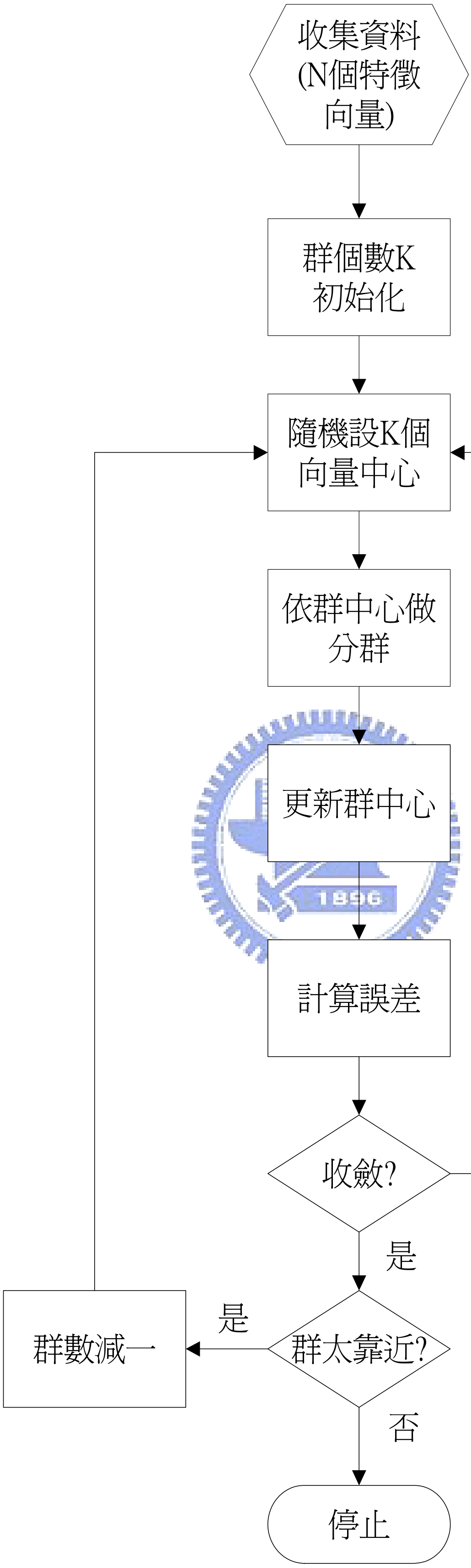

圖 2-4 K 平均值分類法(K-means Cluster)流程圖

2.4 期望值最大演算法(Expectation Maximization ,EM)

我們在做背景模型訓練時，最終的目的是估測最佳的高斯混合模型參數λ，所謂的『最佳』指的是，影像像素值真正的分佈，與模型參數λ估測出來的分佈有最大的相似度，估測最佳參數的方法有很多，但最受歡迎、最適合的方法是『最佳相似性估測法』(Maximum Likelihood Estimation ,MLE)。

在 2.2 節高斯密度函數的假設下，當$x = x_i$時，其機率密度為$P(x_i|\lambda)$，如果x_i,i=1~n 之間是互相獨立的事件，則發生$X = \{x_1, x_2, ..., x_n\}$的機率密度之相似函數(likelihood function)可以表示成：

$$P(X|\lambda) = \prod_{i=1}^{n} P(x_i|\lambda) \tag{2.4}$$

由於 X 是確定的，因此 MLE 主要就是找出使得高斯混合模型的相似函數值為最大時的參數λ'，也就是$\lambda' = \arg\max_{\lambda} P(X|\lambda)$，但是(2.4)式對$\lambda$而言是一個非線性的方程式，無法直接最大化相似函數，所以我們採用期望值最大演算法(Expectation Maximization Algorithm)，利用疊代的方式找出 MLE 的估測參數λ'。

EM 演算法的基本做法是先由之前 K 平均值分類法找出的初始化參數λ，再利用 EM 估計出新的參數$\bar{\lambda}$，使得滿足$P(X|\bar{\lambda}) \geq P(X|\lambda)$，令$\lambda = \bar{\lambda}$重新疊代估計新的$\bar{\lambda}$，直到$P(X|\lambda)$收斂或是達到某個門檻值才停止。EM 演算法主要分成 2 個部分，與 likelihood 函數有關的 E-Step，以及更新參數方程式的 M-Step。

2.4.1 E-Step

目的是測試我們所求的 likelihood 函數值，是否達到我們的要求，若符合要求，EM 演算法就停止，反之就繼續執行 EM 演算法。這裡為了數學推導的方便，假設我們的模型是由三個高斯分佈函數所構成，則其密度函數可表示成：

$$P(x) = w_1 g(x;\mu_1,\Sigma_1) + w_2 g(x;\mu_2,\Sigma_2) + w_3 g(x;\mu_3,\Sigma_3) \tag{2.5}$$

其中共變異矩陣部分Σ_j，因為 2.1 節有提過每個維度彼此獨立，所以只剩對角有值，P(x) 的參數 λ =[$w_1, w_2, w_3, \mu_1, \mu_2, \mu_3, \Sigma_1, \Sigma_2, \Sigma_3$]，參數個數為 (1+1+1+d+d+d+d+d+d)=3+6d 個，依前述 MLE 原則，求出 likelihood 的最大值：

$$\begin{aligned} E(\lambda) &= \ln\left(\prod_{i=1}^{n} P(x_i)\right) = \sum_{i=1}^{n} \ln P(x_i) \\ &= \sum_{i=1}^{n} \ln[w_1 g(x_i;\mu_1,\Sigma_1) + w_2 g(x_i;\mu_2,\Sigma_2) + w_3 g(x_i;\mu_3,\Sigma_3)] \end{aligned} \tag{2.6}$$

爲簡化討論，再引進另一個數學符號稱事後機率(post probability):

$$\begin{aligned} \beta_j(x) &= p(j|x) = \frac{p(j \cap x)}{p(x)} = \frac{p(j)p(x|j)}{p(x)} \\ &= \frac{p(j)p(x|j)}{p(1)p(x|1) + p(2)p(x|2) + p(3)p(x|3)} \\ &= \frac{w_j g(x;\mu_i,\Sigma_i)}{w_1 g(x;\mu_1,\Sigma_1) + w_2 g(x;\mu_2,\Sigma_2) + w_3 g(x;\mu_3,\Sigma_3)} \end{aligned} \tag{2.7}$$

2.4.2 M-Step

主要目的是為了要找到使 likelihood 函數最大化的參數，因此我們分別對 w_i 、μ_i 、Σ_i 做偏微分，再做後續的運算，於是我們便可以得到所求的參數，接著返回 E-Step 繼續做。

假設初始參數是λ_{old}，我們希望找出新的λ值，滿足$E(\lambda)>E(\lambda_{old})$，因為根據$\ln(\frac{a}{b})=\ln(a)-\ln(b)$，$E(\lambda)-E(\lambda_{old})$可以延伸成下式：

$$E(\lambda)-E(\lambda_{old})=\sum_{i=1}^{n}\ln\left[\frac{w_1 g(x_i;\mu_1,\Sigma_1)+w_2 g(x_i;\mu_2,\Sigma_2)+w_3 g(x_i;\mu_3,\Sigma_3)}{w_{1,old} g(x_i;\mu_{1,old},\Sigma_{1,old})+w_{2,old} g(x_i;\mu_{2,old},\Sigma_{2,old})+w_{3,old} g(x_i;\mu_{3,old},\Sigma_{3,old})}\right]$$

$$=\sum_{i=1}^{n}\ln\left[\frac{w_1 g(x_i,\mu_1,\Sigma_1)}{D(\lambda_{old})}\frac{\beta_1(x_i)}{\beta_1(x_i)}+\frac{w_2 g(x_i,\mu_2,\Sigma_2)}{D(\lambda_{old})}\frac{\beta_2(x_i)}{\beta_2(x_i)}+\frac{w_3 g(x_i,\mu_3,\Sigma_3)}{D(\lambda_{old})}\frac{\beta_3(x_i)}{\beta_3(x_i)}\right]$$

$$\geq\sum_{i=1}^{n}\left[\beta_1(x_i)\ln\frac{w_1 g(x_i;\mu_1,\Sigma_1)}{D(\lambda_{old})\beta_1(x_i)}+\beta_2(x_i)\ln\frac{w_2 g(x_i;\mu_2,\Sigma_2)}{D(\lambda_{old})\beta_2(x_i)}+\beta_3(x_i)\ln\frac{w_3 g(x_i;\mu_3,\Sigma_3)}{D(\lambda_{old})\beta_3(x_i)}\right]$$

$$=Q(\lambda) \qquad (2.8)$$

上式中，因為 ln(x)是一個凸函數(Convex Function)，滿足下列不等式：

$$\ln[\alpha x_1+(1-\alpha)x_2]\geq\alpha\ln(x_1)+(1-\alpha)\ln(x_2) \qquad (2.9)$$

推廣上式到「傑森不等式」(Jensen Inequality)：

$$\ln\left(\sum_{i=1}^{n}\alpha_i x_i\right)\geq\sum_{i=1}^{n}\alpha_i\ln(x_i),\ \sum_{i=1}^{n}\alpha_i=1 \qquad (2.10)$$

因為$\sum_{j=1}^{3}\beta_j(x_i)=1$，所以可以將傑森不等式套用在 2.8 式，最後得到下式：

$$E(\lambda)\geq E(\lambda_{old})+Q(\lambda) \qquad (2.11)$$

只要$Q(\lambda)>0$，必滿足$E(\lambda)>E(\lambda_{old})$，但我們通常希望$E(\lambda)$越大越好，最直接的方式就是找出使得$Q(\lambda)$最大的$\lambda$值，那$E(\lambda)$也會跟著變大，見圖 2-5。

$Q(\lambda)$是λ的函數，將一些與λ不相關的部分併入常數項，並重新整理$Q(\lambda)$成下式：

$$Q(\lambda)=\sum_{i=1}^{n}\sum_{j=1}^{3}\beta_j(x_i)[\ln w_j+\ln g(x_i;\mu_j,\Sigma_j)]+c1$$

$$=\sum_{i=1}^{n}\sum_{j=1}^{3}\beta_j(x_i)\left\{\ln w_j+\ln\left[\frac{1}{(2\pi)^{d/2}[\det\Sigma_j]^{1/2}}\exp\left(-\frac{(x_i-\mu_j)\Sigma_j^{-1}(x_i-\mu_j)^T}{2}\right)\right]\right\}+c1$$

對 μ_j 偏微分，$\partial_{\mu_j} Q = 0 \Rightarrow \mu_j = \dfrac{\sum_{i=1}^{n} \beta_j(x_i) x_i}{\sum_{i=1}^{n} \beta_j(x_i)}$ (2.12)

對 Σ_j 偏微分，$\partial_{\Sigma_j} Q = 0 \Rightarrow \Sigma_j = \dfrac{\sum_{i=1}^{n} \beta_j(x_i)(x_i - \mu_j)(x_i - \mu_j)^T}{\sum_{i=1}^{n} \beta_j(x_i)}$ (2.13)

欲得到最佳之 w_j 值，須將 w_j 的總合為 1 的條件加入，引進 Lagrange Multiplier，並定義新的目標函數(object function)為：

$$E_{new}(\lambda) = E(\lambda) + \alpha(w_1 + w_2 + w_3 - 1) \tag{2.14}$$

將 E_{new} 對 3 個 weighting 做偏微分，可得到下面 3 個方程式：

$$\frac{\partial E_{new}}{\partial w_j} = -\frac{1}{w_j}\sum_{i=1}^{n} \beta_j(x_i) + \alpha = 0,\ j = 1,2,3 \tag{2.15}$$

最後將 2.15 的 3 個式子相加，可得到：

$$(w_1 + w_2 + w_3)\alpha = -\sum_{i=1}^{n} [\beta_1(x_i) + \beta_2(x_i) + \beta_3(x_i)]$$

$$\Rightarrow \alpha = -\sum_{i=1}^{n} 1 = -n$$

$$\Rightarrow w_j = \frac{1}{n}\sum_{i=1}^{n} \beta_j(x_i),\ j = 1,2,3 \tag{2.16}$$

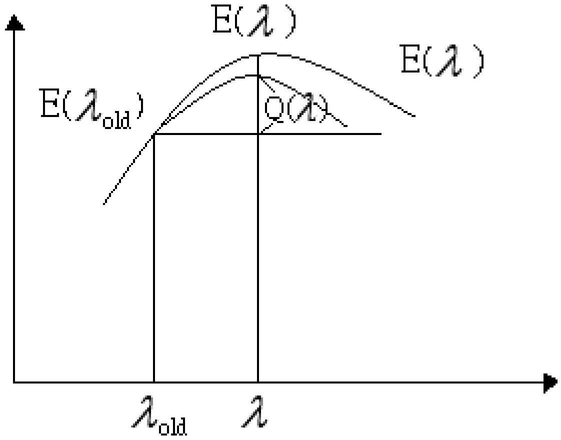

圖 2-5 Likelihood function $E(\lambda)$ 最大化的示意圖

2.5 GMM 建立的架構

綜合前面各小節的說明，GMM 建立的整個流程如圖 2-6 所示，先將 N 個準備拿來訓練模型的資料點，經過 K-means Clustering 後得到初始的參數，再由 EM 演算法得到的三個方程式，

$$\mu_j = \frac{\sum_{i=1}^{n} \beta_j(x_i) x_i}{\sum_{i=1}^{n} \beta_j(x_i)} \qquad \Sigma_j = \frac{\sum_{i=1}^{n} \beta_j(x_i)(x_i - \mu_j)(x_i - \mu_j)^T}{\sum_{i=1}^{n} \beta_j(x_i)} \qquad w_j = \frac{1}{n} \sum_{i=1}^{n} \beta_j(x_i)$$

進行參數的更新，並計算新的相似函數的值，如此不斷的疊代，不斷地更新模型的參數，直到相似函數的值已經沒什麼變動，或是疊代的次數超過某個門檻值，才停止疊代。

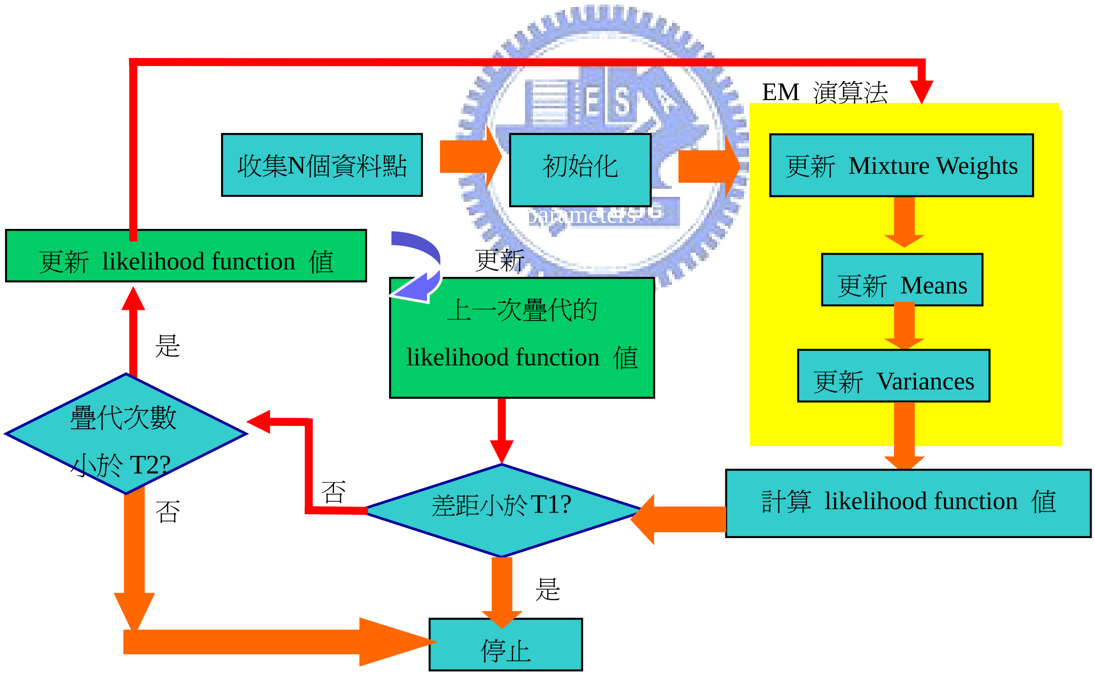

圖 2-6 高斯混合模型建立的架構