

# A RELEVANCE FEEDBACK LEARNING MODEL FOR REGION BASED IMAGE ANNOTATION AND RETRIEVAL SYSTEM

Student: Chia-Pin Cheng

Advisor: Dr.Hao-Ren Ke, Dr. Wei-Pang Yang

Institute of Computer and Information Science

Nation Chiao Tung University

## Abstract

Conventional keyword-based image retrieval systems take a large amount of human labor to annotate images; furthermore, the annotations given by human are subjective to annotators. This thesis proposes an algorithm to automatically annotate images. This algorithm exploits three techniques to modify the co-occurrence model proposed by [Mori99].

1. Segment an image into regions, each of which corresponds to an object. The regions identified by region-based segmentation are more consistent with human cognition than those identified by block-based segmentation.
2. According to visual features (color and shape), map an identified region into three most similar clusters to obtain its associated semantic concept.
3. Strengthen the weight of a region located in the center of an image, because the central object is usually the focus of an image.

The semantic concepts derived by the above algorithm may not be the same as the real semantic concepts of the underlying images, because the former concepts depend on the low-level visual features. To ameliorate this problem, we propose a relevance-feedback model to learn the long-term and short-term interests of users.

The experiments show that the proposed algorithm outperforms the traditional co-occurrence model about 19.5%; furthermore, after five times of relevance feedback, the mean average precision improves from 46% to 82.7%.

Keywords: keyword-based image retrieval, co-occurrence model, relevance feedback.



# 回饋式影像自動標註與檢索系統

研究生：鄭佳彬

指導教授：柯皓仁博士，楊維邦博士

## 國立交通大學資訊科學所

### 摘要

以傳統方式建構關鍵字式影像檢索系統時，必須花費大量的人力與時間為影像進行標註，然而標註的內容往往受到標註人員主觀性的影響。本論文提出一套自動影像標註方法，利用影像切割取得影像中的物件，透過共同出現模式，採用下列三項技術協助影像進行標註：1.利用區域式影像切割，將影像切割成幾個跟人們視覺上比較吻合的物件；2.將所取出之物件對映到最接近的前三群進行正規劃取得更適合物件的語意概念；3.加強位於影像中央物件之語意概念所佔的權重。本論文亦透過使用者相關回饋的方法，提供使用者更為精確的檢索結果。由實驗中得知，相較於傳統共同出現模式，本系統平均準確率提升了 19.45%，經過五次的相關回饋後，系統的平均準確度可以由原本的 46%提升到 82.7%。

關鍵字：關鍵字式檢索、共同出現模式、相關回饋。

## 致謝

兩年前，剛升上研究所，一個人來到了原本我十分陌生的交大，當時的我一度適應不良這邊的生活步調，曾經一度萌生想放棄念研究所的信念。十分幸運的，當時有了實驗室的老師與學長們的鼓勵，讓我一顆不安的心逐漸平靜了下來，也開始了我研究所的求學路程。在這兩年的時間內，非常感謝柯皓仁老師以及楊維邦老師兩位指導教授給予我在研究過程中的指導，並且很有耐心的為我的畢業論文進行修訂與指導，讓我的論文得以順利完成。

謝謝實驗室的鄭培成與葉鎮源兩位學長，在我研究遇到困難時，對我的指導與幫忙，使我的研究得以順利進行，並且我要謝謝實驗室的所有學長、學弟妹跟同學們讓我這兩年的生活有了許多美好的回憶。

最後我要謝謝我的父母與家人，沒有你們的細心栽培與鼓勵也沒有今天的我，所以在這邊獻上我最誠摯的感謝。



**May 27, 2004**

# 目錄

英文摘要.....	I
中文摘要.....	III
致謝.....	IV
目錄.....	V
表目錄.....	VI
圖目錄.....	VII
第一章 簡介.....	1
第一節 影像檢索系統.....	1
第二節 研究動機與目的.....	2
第三節 論文架構.....	3
第二章 相關研究工作.....	4
第一節 內容式影像檢索相關研究.....	5
第二節 自動化影像標註相關研究.....	7
第三節 提升語意式影像檢索方法效能之相關研究.....	17
第三章 回饋式影像自動標註與檢索系統 (MRCOM).....	24
第一節 系統架構.....	24
第二節 影像特徵擷取.....	26
第三節 訓練模組.....	30
第四節 自動化標註模組.....	35
第五節 使用者查詢與回饋模組.....	39
第六節 更新訓練資料集中每一群之KPV.....	45
第四章 實驗與討論.....	46
第一節 實驗資料與分類.....	46
第二節 評估方法.....	46
第三節 MRCOM與整體影像內容標註方法之效能評估.....	48
第四節 自動化標註方法效能評估.....	54
第五節 使用者回饋機制效能評估.....	64
第五章 結論與未來研究方向.....	69
第一節 結論.....	69
第二節 未來研究方向.....	70
參考文獻.....	72

## 表目錄

表格 1：顏色量化對照表[Ravishankar98].....	27
表格 2：群K的關鍵字機率向量.....	35
表格 3：IBAM與MRCOM之MAP比較.....	51
表格 4：風景類之MAP.....	55
表格 5：交通類之MAP.....	60
表格 6：動物類之MAP.....	61
表格 7：植物類之MAP.....	61
表格 8：飾品類之MAP.....	62
表格 9：建築類之MAP.....	62
表格 10：六大類別之整體MAP.....	63



## 圖目錄

圖 1：相關研究工作發展.....	4
圖 2：顏色長條圖範例影像.....	6
圖 3：範例影像圖 2 之長條圖.....	6
圖 4：鏈碼[Freeman74].....	7
圖 5：Block Based切割[Lim99].....	8
圖 6：Region Based切割[Duygulu02].....	8
圖 7：共同出現模組流程圖 1[Mori99].....	11
圖 8：共同出現模組流程圖 2[Mori99].....	11
圖 9：轉換模組[Duygulu 02].....	13
圖 10：語意網路[Lu00].....	17
圖 11：語意式相關回饋流程圖[Lu00].....	20
圖 12：概念相似矩陣[Zhou02].....	22
圖 13：MRCOM系統架構圖.....	25
圖 14：影像作完切割之結果.....	26
圖 15：訓練模組流程圖.....	31
圖 16：影像切割流程圖.....	32
圖 17：K-Means Clustering流程圖.....	34
圖 18：判別位於中央之區塊圖.....	37
圖 19：影像標註表示法示意圖.....	38
圖 20：群與關鍵字關連性對照表 (CKAM).....	39
圖 21：使用者查詢與回饋模組.....	40
圖 22：修改使用者所指定影像之KPV流程圖.....	42
圖 23：修改CKAM流程圖.....	43
圖 24：IBAM與MRCOM檢索關鍵字「河流」之 11-point Interpolated Measure Graph.....	49
圖 25 IBAM與MRCOM檢索關鍵字「車」之 11-point Interpolated Measure Graph.....	50
圖 26：IBAM與MRCOM檢索關鍵字「鹿」之 11-point Interpolated Measure Graph.....	50
圖 27：IBAM與MRCOM檢索關鍵字「向日葵」之 11-point Interpolated Measure Graph.....	50
圖 28：IBAM與MRCOM檢索關鍵字「戒指」之 11-point Interpolated Measure Graph.....	51
圖 29：IBAM與MRCOM檢索關鍵字「建築物」之 11-point Interpolated Measure Graph.....	51
圖 30：以IBAM檢索「車」回傳之前 20 張影像.....	52

圖 31：以MRCOM檢索「車」回傳之前 20 張影像 .....	53
圖 32：風景類之 11-point Interpolated Measure Graph .....	55
圖 33：以瀑布為查詢條件，BCOM執行結果 .....	57
圖 34：以瀑布為查詢條件，RCOM執行結果 .....	57
圖 35：以瀑布為查詢條件，MRCOM執行結果 .....	58
圖 36：瀑布影像 .....	59
圖 37：將圖 34 作block影像切割所得結果 .....	59
圖 38：將圖 34 作region影像切割所得結果 .....	59
圖 39：交通類之 11-point Interpolated Measure Graph .....	60
圖 40：動物類之 11-point Interpolated Measure Graph .....	61
圖 41：植物類之 11-point Interpolated Measure Graph .....	61
圖 42：飾品類之 11-point Interpolated Measure Graph .....	62
圖 43：建築類之 11-point Interpolated Measure Graph .....	62
圖 44：六大類別整體之 11-point Interpolated Measure Graph .....	63
圖 45：老虎第一次檢索結果 .....	64
圖 46：老虎第一次相關回饋後檢索結果 .....	65
圖 47：老虎第二次相關回饋後檢索結果 .....	66
圖 48：老虎第二次相關回饋後檢索結果 .....	67
圖 49：六大類別相關回饋後之 11-point Interpolated Measure Graph .....	68
圖 50：六大類別相關回饋後之MAP .....	68





# 第一章 簡介

## 第一節 影像檢索系統

隨著資訊技術的進步、網際網路的蓬勃發展，人們常常利用電腦連結上網，找尋他們想要的資料。一般使用者經常利用網站所提供的搜尋引擎查詢想要找尋的相關網站。在這個過程中，人們必須提供想找尋資料的相關資訊，最常見的就是由使用者提供關鍵字，接著搜尋引擎再去檢索與這些關鍵字相關的文件，並且送回結果給使用者。搜尋引擎在進行關鍵字的檢索方面有著很不錯的效率，但是隨著數位科技的進步，人們想找尋的文件並不侷限於文字相關的文件，而進一步衍生出想要尋找影像或者是影片相關的多媒體文件。本論文主要探討的是如何讓使用者可以找到他們想要的影像資料。

針對於這個問題，一般最直覺的想法就是人工替所有影像資料一一註解，以達到詮釋影像的目的。再利用搜尋引擎對於文字強大的檢索能力去尋找相關的影像。利用這樣的方式，有兩個主要的問題：第一，就是人工註解大量的影像資料時，必須花費大量的人力與時間。第二，所謂一張影像勝過千言萬語，要充分描述影像裡面的內容是很不容易的，並且由不同的人來看同一張影像所得到的觀感不盡相同，因此隨著幫影像作註解的人不一樣，所作出來的註解可能就不太一樣，也就是人們下註解時主觀性的問題。

影像本身包含了大量的資訊，除了人工添加註解外，還可以利用影像的一些特徵去比對整個影像資料庫中所有的影像，挑出與使用者欲查詢之影像的特徵較為相近的影像，並將結果回傳給使用者，這種的檢索方式統稱為內容式影像檢索(Content based Image Retrieval, 簡稱 CBIR)。最常被人們用來描述的影像特徵有顏色、形狀與紋理等。

## 第二節 研究動機與目的

當使用內容式影像檢索系統時，系統往往會要求使用者提供一張影像當作查詢的範例，甚至有些系統提供了簡易的繪圖工具，讓使用者描繪所欲檢索的影像輪廓 (Picasso System[Bimbo97])，接著系統將所提供的範例進行特徵的擷取，再去跟影像資料庫中的所有影像作特徵相似度的比對，找出相似度較高的影像傳回給使用者。但是這樣的檢索方式對使用者而言是相當的不方便，因為當使用者無法提供查詢範例或是系統所提供的繪圖工具很難去描繪要查詢的影像時，就很難利用內容式影像檢索系統去檢索使用者想要找尋的影像。另一方面，電腦對於顏色與紋理等特徵的判斷分析，對人們而言是比較難以理解，所以當使用內容式影像檢索系統時，系統常常回傳一些令人無法理解的影像。由於現階段各大搜尋引擎所提供的搜尋方法大多是要求使用者給予關鍵字以表達他們所要找尋的資料，且這種查詢的方法對人們而言是比較親切而方便的，所以提供一套讓使用者利用關鍵字表達他們所要找尋影像資料的語意概念 (Semantic Concept)，進而去作影像的檢索是很有意義的。但是目前各大搜尋引擎針對利用關鍵字去搜尋相關影像處理的方法，普遍是將出現在網頁上影像週遭的文字當作是與影像相關的文字，並以一般檢索網頁的方式對這些文字進行檢索，可是這些文字有時候並非與影像有絕對的關係，導致於檢索出來的影像有時候讓使用者覺得無法理解。

本論文主要研究如何幫影像作整體概念上的詮釋，並且提供使用者語意式 (Semantic Based) 的檢索方式。如前所述，若是利用人工的方式來幫影像作詮釋將會產生耗時耗力與人們主觀性的等等問題，所以本論文利用機器學習 (Machine Learning) 與圖形識別 (Pattern Recognition) 的方法來協助建立影像註解。本論文利用影像處理 (Image Processing) 的方法切割影像，所得之區域 (Region)則視為存在於影像中的物件 (Object)，並且利用這些已取得之物件搭配人工給予的文字形成一個訓練資料集 (Training Set)。對於未加入註解的影像，

由先前已經預先訓練好的資料裡面，學習出每一張新的影像中所代表的語意概念，由這些語意概念再進而求得代表此影像之關鍵字。如此一來使用者將可以利用他們所熟悉的關鍵字檢索來查詢影像。由於利用機器學習與圖形識別並不一定可以達到十分準確的效果，所以本論文也利用使用者相關回饋 (User Relevance Feedback) 來提升系統的準確度。在本論文的系統中，將提供了多樣式的檢索方式，有範例式檢索、關鍵字式檢索與物件式檢索。

### 第三節 論文架構

本論文的第二章將介紹影像檢索與影像標註的相關研究；在第三章則介紹本論文所提出的影像標註方法與使用者相關回饋機制；第四章透過實驗的評估驗證本論文所提出方法之可行性；最後在第五章總結本論文，並且探討未來研究的方向。



## 第二章 相關研究工作

本章介紹與本論文相關的研究工作。對於本論文中，自動化影像標註與檢索的相關研究主要分為以下三方面：

- 內容式影像檢索：[Pinheiro00][Zhou01][Cinque01][Zhang02][Ko02]
- 自動影像標註：[Mori99][Lim99][Duygulu02][Barnard03][Blei03]
- 回饋模組：[Rui99][Lu00][Zhou02]

圖 1 是依照年份與相關技術所整理的相關研究發展，粗體部分為本論文所參考的相關方法。

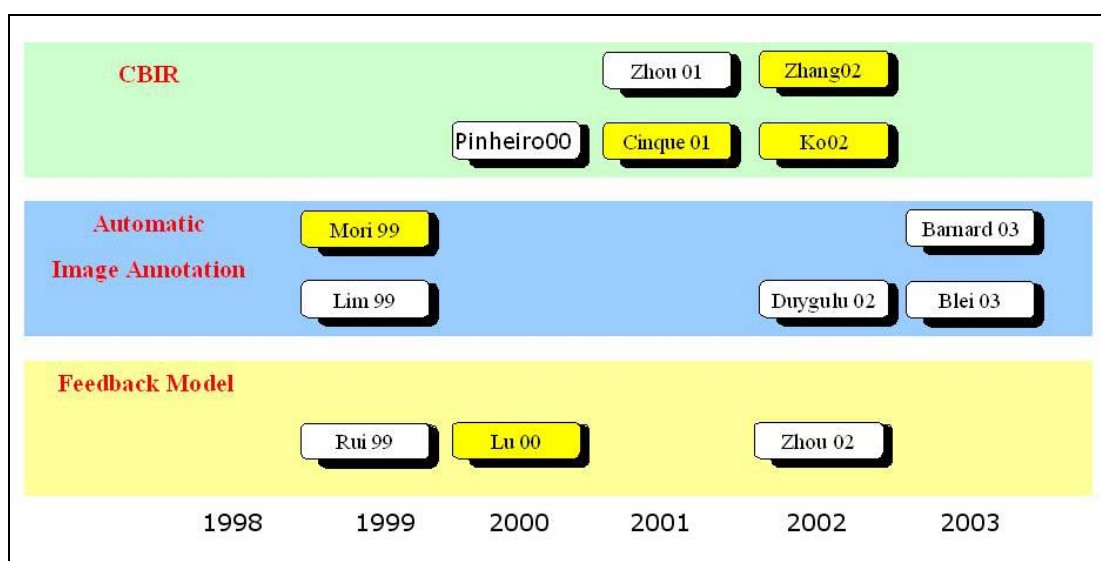


圖 1：相關研究工作發展

本論文主要是研究如何自動化地幫助影像建立標註與提供使用者利用關鍵字作檢索，所以相關研究主要把重點放在如何做影像的檢索、如何建立影像標註與如何利用使用者相關回饋機制來提升系統效能。

本論文是從影像的低階特徵去推論的影像語意概念，所以首先在第一節先介紹內容式影像檢索中利用影像內容特徵進行相似度比較的相關研究；接著第二節

介紹自動影像標註方法的相關研究。因為推論出來影像的語意概念可能與人們實際看法有所差異或是標註錯誤，所以在第三節介紹利用使用者相關回饋的機制進行學習，並且提升系統效能的相關方法。

## 第一節 內容式影像檢索相關研究

近年來有許多影像檢索相關研究提出利用影像內容進行影像檢索，也就是在第一章中所提到的內容式影像檢索 (CBIR)，並且有許多的相關系統也已實作出來，其中 IBM 的 QBIC (Query by Image Content) [Flickner95]、Visual Retrievalware、Virage Search Engine[Gupta97]屬於商業性質的系統；其它如 Photobook[Pentland96]、Candid[Kelly95]、Chabot[Qgle95]與哥倫比亞大學所開發的 Visual Seek[Smith96]等屬於研究性質的系統等。內容式影像檢索系統係利用影像的低階特徵，如顏色、形狀、結構與紋理等來進行影像之間的相似度量測。茲整理主要用來表示影像的特徵如下：



### 2.1.1 顏色表示方式

一般在擷取影像的顏色特徵時，會先將影像轉換到特定的顏色空間(Color Space)，比較常用的表示法有 RGB color space、CIE-Lab color space [Paschos01]、HSV color space[Paschos01]、與 YIQ color space 等。在將影像轉換到特定的顏色空間後，可以用長方條的統計圖(Histogram)[Swain91][Funt95]、顏色在空間上的分布情況 (Spatial Layout) [Hu02]、或顏色的變化程度 (Color Moment) [Mostafa02]等作為影像在該顏色空間的特徵表示法。其中，長條圖表示法是將顏色空間分割成不同的區間 (Bins)，如圖 2；接著再統計影像中的顏色在每一個區間 (Bin) 中各佔多少比例，並表示成長方條的統計圖，如圖 3。



圖 2：顏色長條圖範例影像

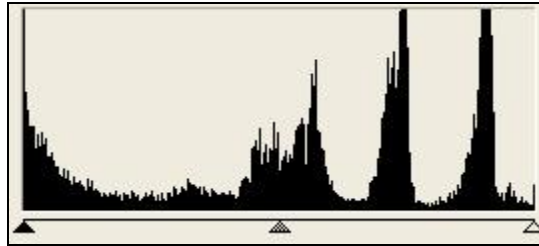


圖 3：範例影像圖 2 之長條圖

### 2.1.2 形狀表示方式



在有些系統中，會將影像切割成數個區域 (Region)，並以區域的形狀來代表區域的特徵。現行研究中，較常用來作為影像形狀特徵的表示法有鏈碼 (Chain Code) [Freeman74]、多邊形逼近 (Polygonal Approximations) [Pinheiro00]、曲率 (Curvature) [Pinheiro00]、富立葉描述器 (Fourier Descriptor) [Zhang02][Tello95]、變化描述器 (Moment Descriptor) [Zhang02]、面積、周長、所在位置[Ko02]等等特徵。其中，鏈碼是一種用來紀錄形狀邊緣方向性變化的表示方式。如圖 4 為一區塊之形狀邊緣，S 表示鏈碼的起始點，八個方向性各有其代表數字，若採順時針方向作紀錄，則此圖形鏈碼的表示法為 00060000555443322。

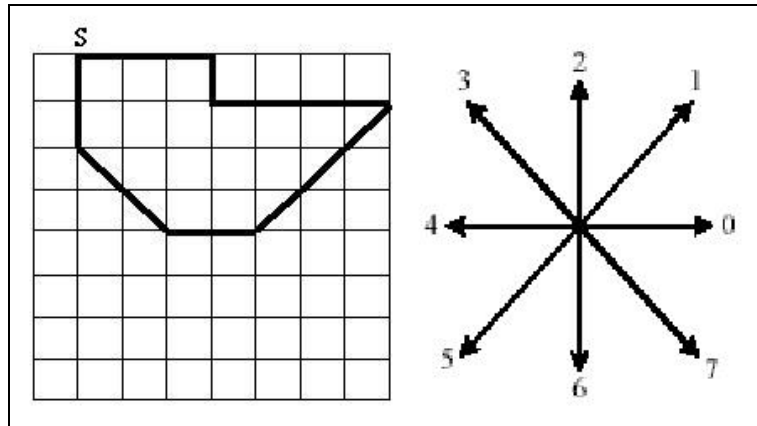


圖 4：鏈碼[Freeman74]

### 2.1.3 紋理表示方式

在影像特徵表示法中，會以紀錄紋理來表示整張影像的結構資訊。常被用到的紋理特徵表示法有加波過濾器 (Gabor Filter) [Paschos01]、多向性過濾描述器 (Multi-orientation Filter Descriptor) [Malik90]、二次變化矩陣 (Second Moment Matrix) [Forstner94][Garding96]、亂度 (Entropy) [Yoo02]等。在 IBM 的 QBIC 系統[Flickner95]中，則以對比 (Contrast)、粗糙度 (Coarseness) 與方向性 (Directionality) 等作為紋理的特徵表示法。

由於內容式影像檢索系統是透過影像低階特徵進行相似度的比對，其執行結果常常跟人們的語意概念有差距，所以也有蠻多的研究在探討如何將影像與人們的語意概念做結合的方法。在本論文中，將建構一套自動化影像標註系統，提供一個將影像與人們語意概念作結合的方法。

## 第二節 自動化影像標註相關研究

針對影像自動標註的相關研究，主要分為兩方面介紹：

1. 自動化影像標註的相關方法：取出影像特徵，針對特徵加以作註解。
2. 影像標註的表示法：建構影像之語意概念之表示方法。

## 2.2.1 自動化影像標註相關方法

為一張影像建立標註時，最主要的就是要取得影像中的內容特徵。由於影像的內容是由存在於影像之中的物件 (Object) 所組成的，所以當進行自動影像標註之前，必須先確認有哪些物件存在於影像之中。目前大部分的做法是利用影像切割找出影像中的物件，接著再由這些物件去計算影像的標註值。切割影像尋找物件的方法主要有兩個方向，一個是 Block Based[Lim99][Mori99]，一個是 Region Based[Carson99]。

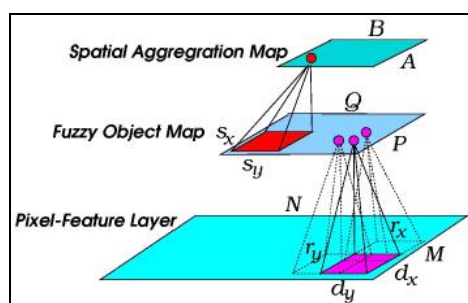


圖 5：Block Based 切割[Lim99]

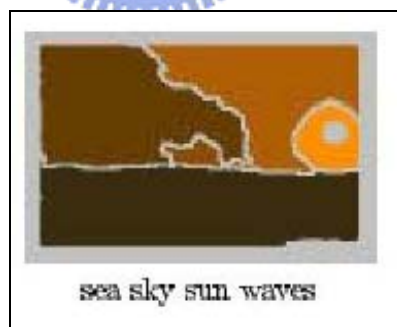


圖 6：Region Based 切割[Duygulu02]

其中，Block Based 在實作上較 Region Based 更為簡單，因為 Block Based 只是單純地將影像切割成數個等大小的矩形 (如圖 5)，由一個或多個矩形所組成的區域則被認定為一個物件。但是 Block Based 切割出來的物件與人們所認定的物件往往有所差異。比如在影像中的一隻老虎，利用 Block Based 切割的話有可



能把老虎分割到數個區塊中，跟人們所認定的一隻完整的老虎無法吻合。在 [Lim99] 中提出了如圖 5 的三層式對映架構，最下面一層像素特徵層 (Pixel-Feature Layer) 代表著影像的最原始的像素特徵。藉由把影像切割成一個一個矩形的 Blocks (Blocks 之間可以重疊)，切出來的每一個 Block 稱為一個視覺化標記 (Visual Token)。再利用如 SOM [Kohonen84] 與 K-means [McQueen67] 等分群的方法或用 EM 演算法 (Expectation Maximization Algorithm) [Bilmes98] 進行推論，把將這些視覺化標記對映到中間的模糊物件對映圖 (Fuzzy Object Map)。在物件層中的每一個點則代表一個物件，由於物件存在於影像之中會有位置關係的對應，所以最後會再將所有的物件對映到最上層的空間聚集對映圖 (Spatial Aggregation Map)，利用此三層架構即可以進行影像的分類。

Region Based 的作法則將影像切割成比較吻合人們所認定的物件 (如圖 6)。但是這種作法的難度較高，因為雖然有很多影像切割技術被提出來，但是很難認定何種切割方法是最好的。比如在影像中有一個人，那到底是要將整個人視為一個物件而切割出來，還是把人的頭、手、腳、身體等分別切開視為獨立的物件呢？所以 Region Based 在實作上有其困難性，但是所切割出來的物件卻是與人們的感知是比較吻合的。

在將影像切割成數個 Regions 或 Blocks 後，接下來則是對這些 Regions 或 Blocks 進行標註，主要的方法有：

1. 共同出現模式 (Co-occurrence Model)。
2. 翻譯模式 (Translation Model)。
3. 跨媒體相關模式 (Cross-Media Relevance Model)。

#### 2.2.1.1 共同出現模式 (Co-occurrence Model)

[Mori99]提出 Co-occurrence Model，計算出現在相同類別中的區塊與關鍵字的頻率，進行機率的統計。做法為將影像資料切割成  $N*N$  個等大小的 Blocks，把每一個 Block 視為是影像中的一個物件，藉由將文字與 Blocks 結合並且分群的方法，學習出每一個 Block 的語意概念。

[Mori99]主要的流程如圖 7 與圖 8 所示，其步驟說明如下：

1. 將用來學習的影像預先給予關鍵字。
2. 將影像切割成數個等份的矩形。
3. 每一個矩形 Block 繼承了整張影像的所有關鍵字。
4. 將所有的 Block 利用向量量化 (Vector Quantization) 的方法進行分群 (Clustering)。
5. 統計在每一個群 (Cluster) 中各個字出現的頻率，並且計算每一個字出現的可能性。

對於未標註的影像，同樣將它切成數等份，如圖 8，擷取每一個 Blocks 的特徵並且尋找與它們最為相近的群。結合找到的所有群的關鍵字出現的可能性，並且判定出現在影像中機率較大的關鍵字。

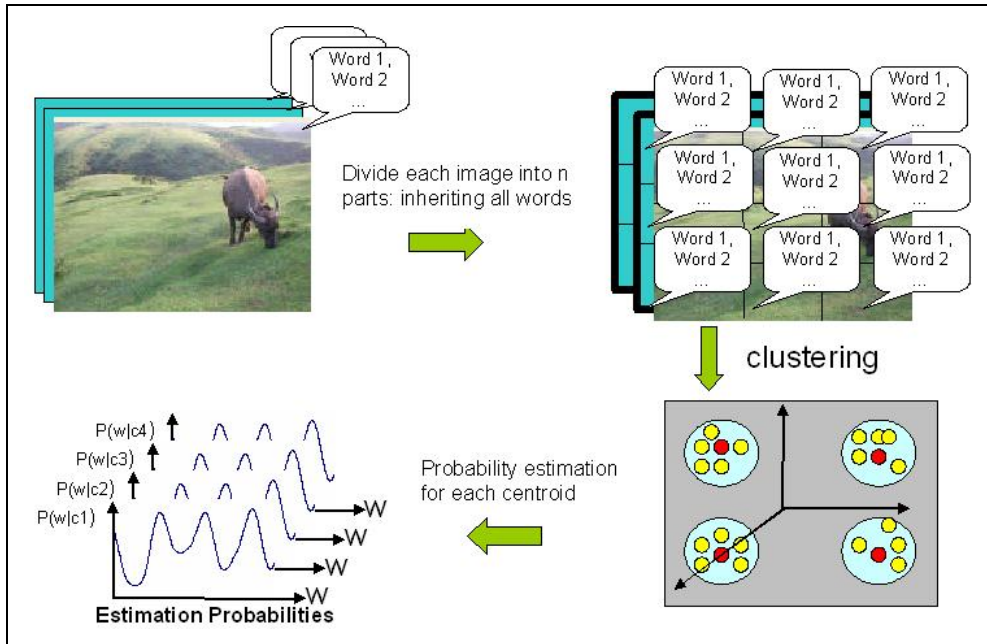


圖 7：共同出現模組流程圖 1[Mori99]

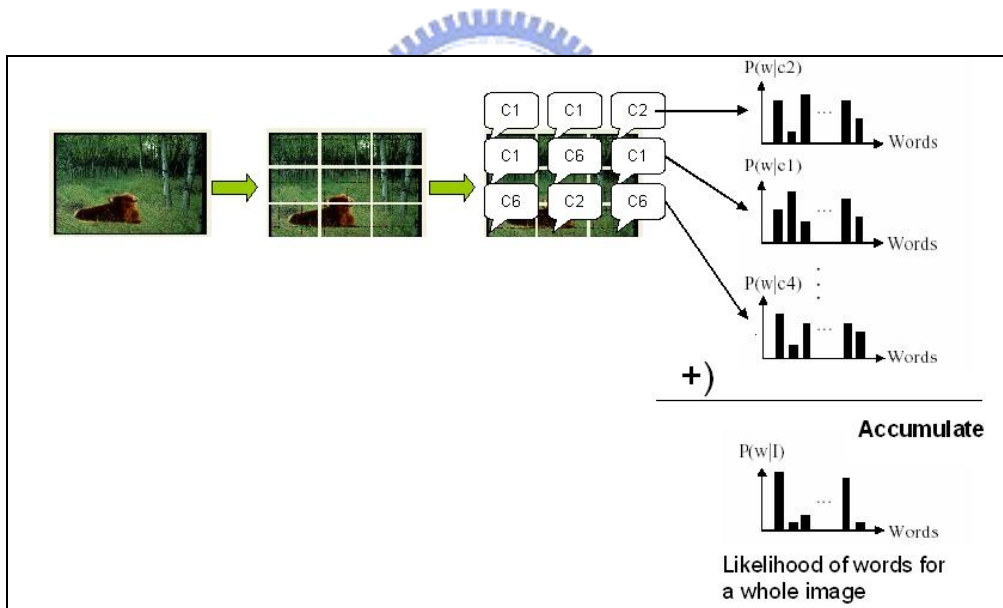
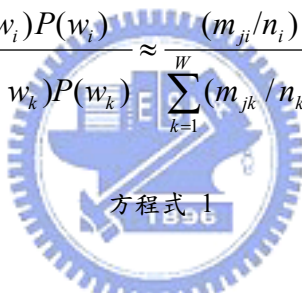


圖 8：共同出現模組流程圖 2[Mori99]

這個方法最主要的構想就是透過相似的物件彼此會有較多相同的關鍵字來降低雜訊。比如說：有一張影像 $I_1$ 中有兩個關鍵字分別為「天空」與「山」，在 $I_1$ 中被切出來且應該屬於天空的部份將同時繼承 $I_1$ 的關鍵字「天空」和「山」，但「山」配置給天空的部份其實是不適當的。然而在另外一張影像 $I_2$ 之中，有兩個

關鍵字「天空」和「河流」，在統合這兩張的資訊後，同屬於天空的那部份應該會被分在同一群裡面，在這一群中的關鍵字，總共有了兩個「天空」、一個「山」與一個「河流」。如此一來可以很容易地發現，「天空」這個關鍵字在這一群之中扮演著比較重要的角色，相對的「山」跟「河流」會隨著越來越多應該屬於天空的樣本 (Pattern) 聚集在一起，而降低它們的重要性。

在 Training Data 依序切割並且完成向量量化與分群後，必須去計算出在每一群中所有相對應的關鍵字發生的可能性 (Likelihood)。假如每一個群的群中心分別為  $c_j$  ( $j=1, 2, \dots, C$ )，每一個可能出現的關鍵字分別為  $w_i$  ( $i=1, 2, \dots, W$ )，則在任一群  $c_j$  中， $w_i$  出現的可能性  $P(w_i | c_j)$  如方程式 1 所示：

$$P(w_i | c_j) = \frac{P(c_j | w_i)P(w_i)}{\sum_{k=1}^W P(c_j | w_k)P(w_k)} \approx \frac{(m_{ji}/n_i) \times (n_i/N)}{\sum_{k=1}^W (m_{jk}/n_k) \times (n_k/N)} = \frac{m_{ji}}{\sum_{k=1}^W m_{jk}} = \frac{m_{ji}}{M_j}$$


方程式 1

其中， $m_{ji}$  表示在群  $c_j$  中  $w_i$  出現次數， $M_j$  表示在群  $c_j$  中所有關鍵字的出現次數， $n_i$  表示在所有資料中的關鍵字  $w_i$  的出現次數， $N$  表示出現在影像資料中的所有關鍵字的數量。最後結果就是藉由統計的方法，計算得到某個關鍵字  $w_i$  出現在某一群  $c_j$  中的相對頻率，以代表在某一群  $c_j$  下關鍵字  $w_i$  可能發生的機率。

### 2.2.1.2 翻譯模式 (Translation Model)

在[Duygulu02]中提出了一種翻譯模組 (Translation Model) 的做法，將翻譯模組視為一種辭典 (Lexicon)，利用機器翻譯的 (Machine Translation) 將某一種語言翻譯到另一種語言。在此即把影像中的區塊當作一種語言的文字，透過機器轉換的動作翻譯成標註的文字。然而機器轉換是將離散的物件 (某種語言中的文

字) 轉換到離散的物件 (另一種語言中的文字), 但是影像中 Region 的特徵並非屬於了離散空間, 所以透過將 Region 的特徵作向量量化進行分群, 則可將其轉換成一種離散的物件。透過分群的動作, 每一個 Region 可以對應到與本身最為相似的群中心, 並將自己標示成特定的 Blob。也就是說透過分群的方法, 屬於同一群的 Regions 會有相同的 Blob 當做自己的標籤。

[Duygulu02] 預測在某一張影像  $I$  中, 某一個 Blob  $b$  所對應在影像  $I$  中的文字意義 (如圖 9)。如方程式 2, 主要是要讓所有影像中的 Blobs 對應到最合適的文字, 所以要盡可能地使方程式 2 計算出來的值達到最大, 也就是讓所有 Blobs 搭配到的文字之機率在整體上達到最大的可能性。其中,  $p(a_{nj} = i)$  表示在影像  $n$  的第  $i$  個 Blob 被轉換到在影像  $n$  中第  $j$  個文字的機率;  $t(w|b)$  則代表給定 Blob  $b$ ,  $b$  可能被轉換成文字  $w$  的機率。在 [Duygulu02] 中利用 EM 演算法 (Expectation Maximization Algorithm) 進行非監督 (Unsupervised) 的學習模式, 如圖 9 所示, 學習的過程中主要是去進行 Blobs 跟 Words 之間發生機率的推論找出最有可能的結果。

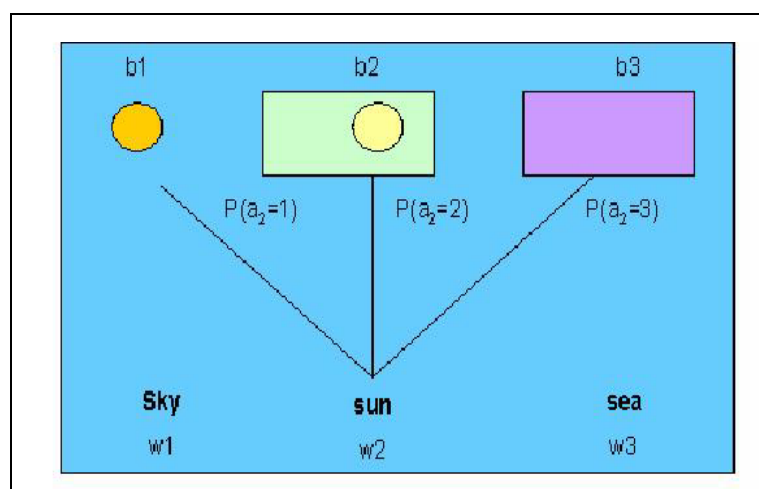


圖 9：轉換模組[Duygulu 02]

$$p(w|b) = \prod_{n=1}^N \prod_{j=1}^{M_n} \sum_{i=1}^{L_n} p(a_{nj} = i) p(w = w_{nj} | b = b_{ni})$$

方程式 2

### 2.2.1.3 跨媒體相關模式 (Cross-Media Relevance Model)

在[Jeon03]中指出，Regions 與字之間並非是一對一 (one to one) 的對映關係，因為有一些字可能是多個 Regions 共同出現時，才有可能出現的。比如說可能在有山、溪流、瀑布、草原等風格的區塊共同出現時，風景這個關鍵字才有可能屬於這張影像。[Jeon03]主要的概念就是利用所有共同出現在同一張影像中的 Regions 來推論每一個關鍵字出現的可能性。方法如下：

假設一個訓練集  $T$  包含了已經被註解好的影像資料。在  $T$  中的任何一張影像資料  $J$  表示成： $J = \{b_1, \dots, b_m; w_1, \dots, w_n\}$ 。其中， $\{b_1, \dots, b_m\}$  代表的是影像  $J$  中的 Regions 所對應到的 Blobs； $\{w_1, \dots, w_n\}$  代表的則是在影像  $J$  中所標註好的關鍵字。對於尚未被標註的影像  $I$ ，將  $I$  切成數個 Regions，並且找到他們所對應的 Blobs，即  $I = \{b_1, \dots, b_m\}$ ，接著計算  $I$  在有這幾個 Blobs 的情況之下，哪些關鍵字最能夠反映出影像中確切的內容。對於一個關鍵字  $w$ ，其出現在  $I$  中的可能性可表示成方程式 3：

$$P(w|I) \approx P(w|b_1, \dots, b_m)$$

方程式 3

亦即，將關鍵字  $w$  發生在影像  $I$  中的可能性趨近於當影像  $I$  有  $\{b_1, \dots, b_m\}$  這些 Blobs 出現的行況下，關鍵字  $w$  出現的可能性。條件機率  $P(w | b_1, \dots, b_m)$  又可轉換為方程式 4：

$$P(w | b_1, \dots, b_m) = \frac{P(w, b_1, \dots, b_m)}{P(b_1, \dots, b_m)}$$

方程式 4

$P(w, b_1, \dots, b_m)$  可以由訓練資料集  $T$  經統計的方法求得每一個 Blob 出現的機率。所以在此主要的問題就是去求得  $P(w, b_1, \dots, b_m)$ 。在此可以利用在訓練資料裡的每一張影像中所包含的 Blobs 與字的一些統計資料來推論  $P(w, b_1, \dots, b_m)$  的值，如方程式 5：



$$P(w, b_1, \dots, b_m) = \sum_{J \in T} P(J) P(w, b_1, \dots, b_m | J)$$

方程式 5

其中  $J$  表在訓練資料集裡面的某一張影像， $P(J)$  表此影像出現的機率，通常對於在訓練資料集裡的所有影像  $J$ ， $P(J)$  都是一致的。在這一個方法中，假設了  $(w, b_1, \dots, b_m)$  之間為獨立不相互影響的事件，所以方程式 5 可改寫為如方程式 6：

$$P(w, b_1, \dots, b_m) = \sum_{J \in T} P(J) P(w | J) \prod_{i=1}^m P(b_i | J)$$

方程式 6

上面的  $P(w | J)$  與  $P(b_i | J)$  可由訓練資料集經方程式 7 與方程式 8 的計算得之：

$$P(b|J) = (1 - \beta_J) \frac{\#(b,J)}{|J|} + \beta_J \frac{\#(b,T)}{|T|}$$

方程式 7

$$P(w|J) = (1 - \alpha_J) \frac{\#(w,J)}{|J|} + \alpha_J \frac{\#(w,T)}{|T|}$$

方程式 8

其中， $\beta_J$  與  $\alpha_J$  分別代表特定權重常數， $\#(b,J)$  表示在影像  $J$  中 Blob  $b$  出現的次數， $\#(b,T)$  表示 Blob  $b$  在訓練資料集  $T$  中出現的次數； $\#(w,J)$  表示在影像  $J$  中關鍵字  $w$  出現的次數（通常值為 0 或 1，因為相同的字很少會重複出現在一張影像中）， $\#(w,T)$  表示關鍵字  $w$  在整個訓練資料集  $T$  中出現的次數； $|J|$  表示在影像  $J$  中所有關鍵字與 Blob 的出現次數， $|T|$  表示訓練資料集  $T$  中影像的張數。

### 2.2.2 影像標註的表示法



本論文主要是研究如何自動化地幫助影像加入標註，所以接下來將討論有關於影像標註的表示法。目前在各相關研究中，針對幫助影像建立標註後的表示方法大致可以分為三大類，這三大類分別為：固定式標註模式 (Fixed Annotation Model) [Jeon03]、語意網路模式 (Semantic Network Model) [Lu00]、機率式標註向量模式 (Probability Annotation Model) [Jeon03]，以下是這三種模組的大致介紹。

#### 1. 固定式標註模式 (Fixed Annotation Model)

採用這種方式來表示影像的註解，就是當認定某張影像跟某關鍵字有強烈的相關性時，便將此關鍵字配置給此影像，屬於一種絕對的表示方式。如在影像  $I$



中，對它的標註方式為  $(I | \text{天空, 太陽, 水})$ ，意義就是說影像  $I$  只與這三個關鍵字相關，而對其餘關鍵字則無相關性。

## 2. 語意網路模式 (Semantic Network Model)

這個模式的主要做法是採用建立影像與關鍵字所形成的語意網路來代表此影像所代表的語意概念。如圖 10，在影像  $I$  中，若包含與關鍵字  $i$  相關的資訊，則在語意網路中會建立起一個連結，連結上有一個權重的標記，用來指出影像  $I$  與關鍵字  $w$  間的關聯程度，權重的值越高表示二者的關連性越大。

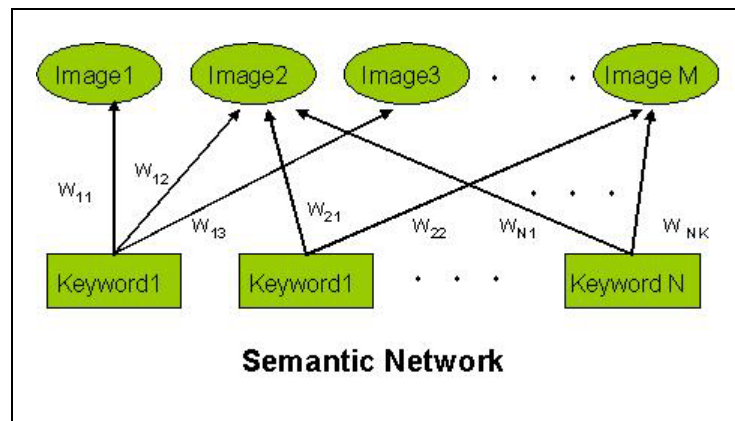


圖 10：語意網路[Lu00]

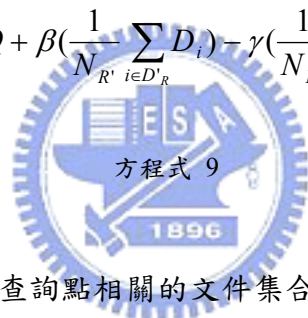
## 3. 機率式註解向量模式 (Probability Annotation Model)

在這個表示法中，是以機率向量來代表影像中所代表的語意概念，向量中每一個維度的值對應到某一關鍵字在此影像出現的可能性，每一個關鍵字相對於此影像的關聯程度介於 0~1 之間。在套用這種表示方式之前，必須把所有可能的關鍵字收集起來形成一套字典，然後將字典中所有的關鍵字，對每一張影像建立起一個機率向量表示式，再依照每個維度所對應的關鍵字為影像建立機率的值。

## 第三節 提升語意式影像檢索方法效能之相關研究

在一些影像檢索系統中提到了利用使用者的相關回饋 (User Relevance Feedback) 的機制，來改善系統的效能。最常見的做法就是在一次檢索後，由使用者提供與檢索目標相關與不相關的樣本 (Samples)，系統經由分析比對這些樣本，進行學習，再重新作檢索。如此，每次的互動後，系統回傳的結果將可以越來越趨近於使用者所想要找尋的目標。在[Lu00]中提到，相關回饋的研究，大致可以分為兩個方向：(1) 修正查詢點 (Query Point)；(2) 重新給予各特徵權重值 (Re-weighting)。修正查詢點的想法就是利用與使用者互動，使查詢點移向使用者所提供較好的樣本點並且遠離不好的樣本點，最後逐漸形成一個理想的查詢點 (Ideal Query Point)。最常被用來修正查詢點的形式就是 Rocchio's Formula，表示法如方程式 9：

$$Q' = \alpha Q + \beta \left( \frac{1}{N'_R} \sum_{i \in D'_R} D_i \right) - \gamma \left( \frac{1}{N'_N} \sum_{i \in D'_N} D_i \right)$$



其中， $D'_R$  表示與理想查詢點相關的文件集合， $D'_N$  表示與理想查詢點不相關的文件集合； $\alpha$ ， $\beta$ ， $\gamma$  分別為合適的常數； $N'_R$  與  $N'_N$  分別代表  $D'_R$  與  $D'_N$  的個數。

在重新給予各特徵權重值的方面，[Rui97]提到利用標準差法 (Standard Deviation Method) 來達到重新給予各特徵權重值的目的。由於每一張影像都可以表示成一個  $N$  個維度的特徵向量，所以把一張影像看作是  $N$  維空間中的一個點。假如由使用者所提供認為是合適的影像轉化在  $N$  維空間中的點，它們在維度  $j$  中有著相當大的變異度 (Variance)，就可以把  $j$  維度為跟我們想要查詢的理想點關連性很低，所以在下一次作檢索的時候，系統就可以將維度  $j$  的權重降低。

以下茲簡述整合語意式查詢與整合使用者相關回饋的相關研究。

### 2.3.1 語意式相關回饋 (Semantic Based Relevance Feedback)

在[Lu00]中提出了一種利用相關回饋的機制整合影像語意網路與影像低階特徵的方法，並且實作出 *iFind* Retrieval System。在[Lu00]中，把使用者回饋的相關資訊作兩部分的處理：(1) 語意式相關回饋 (Semantic Based Relevance Feedback)；(2) 利用相關回饋的機制整合影像低階特徵的方法。

語意式相關回饋的處理過程如下：

1. 剛開始的時候，把每一個語意網路上連結的權重都設為 1，也就是代表每一個字都有相同的重要性。
2. 收集使用者所給予的查詢目標與提供之正例與反例資料。
3. 確認在使用者查詢的目標中是否有關鍵字是目前資料庫內所沒有的。如果有的話，把這一個關鍵字加入到資料庫之中，但是不建立任何的連結到語意網路之中。
4. 針對所有使用者提供的正例影像  $I$ ，確認在使用者查詢目標中是否有任何關鍵字  $k$  與  $I$  在語意網路中並無連結；如果有的話，在語意網路中建立起  $I$  與  $k$  的連結，並給予起始權重 1；若是  $k$  與  $I$  在語意網路中原本就有連結存在的話，將其連結的權重加 1。
5. 針對所有使用者所提供的反例影像  $I'$ ，確認在使用者查詢目標中是否有任何關鍵字  $k'$  與  $I'$  在語意網路中存在連結；若有的話，將此連結的權重更改為原本權重的四分之一。若連結的權重小於 1，則將連結移除。

語意式相關回饋流程圖：

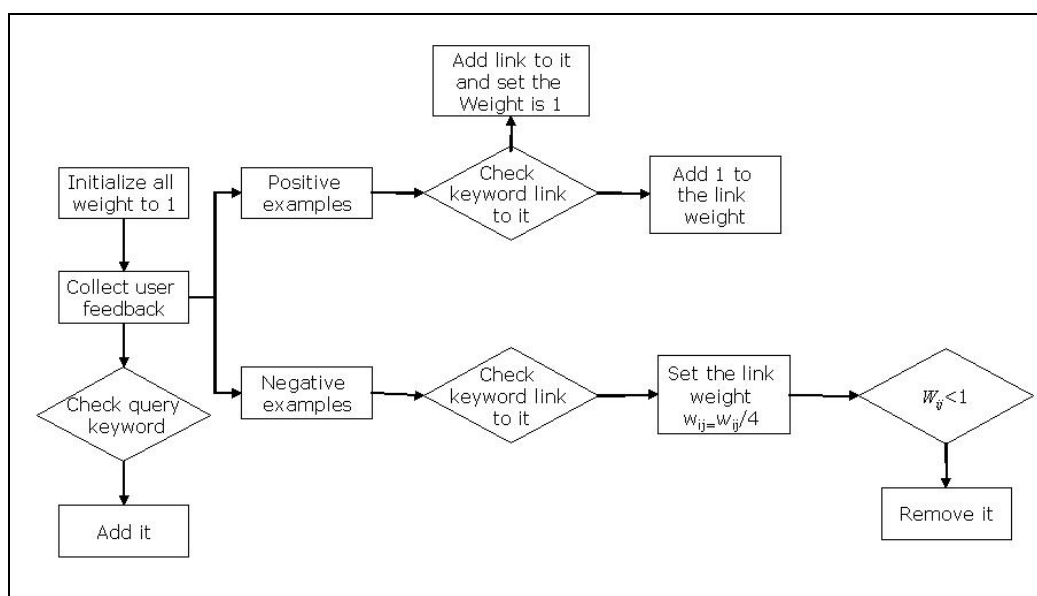


圖 11：語意式相關回饋流程圖[Lu00]

[Lu00]提到了如果某一個關鍵字  $k$  跟影像資料庫中大部分的影像之間都有連結存在的時候，可以利用方程式 10 計算相關因子來調整這個問題的發生，其中  $r_k$  代表著相對應的相關因子， $w_k$  表示原本影像與關鍵字間連結的權重， $M$  代表在影像資料庫中影像的數量， $d_k$  則是代表著有多少連結存在於所有的影像  $I$  與關鍵字  $k$  之間。

$$r_k = w_k \left( \log_2 \frac{M}{d_k} + 1 \right)$$

方程式 10

當使用者下達查詢時，系統會依據查詢中出現的關鍵字與影像之間的相關程度進行排名回傳給使用者。當使用者不滿意系統所給的結果時，系統可藉由使用者所提供之相關與不相關的影像範例整合影像低階特徵進行重新學習。[Lu00] 主要是將 Rocchio's formula 進行修改成如方程式 11；

$$G_j = \log(1 + \pi_j)D_j + \beta \left\{ \frac{1}{N_R} \sum_{k \in N_R} \left[ \left( 1 + \frac{l_1}{A_1} \right) S_{jk} \right] \right\} - \gamma \left\{ \frac{1}{N_N} \sum_{k \in N_N} \left[ \left( 1 + \frac{l_2}{A_2} \right) S_{jk} \right] \right\}$$

方程式 11

其中， $\pi_j = \alpha^M \sum_{i=1}^M r_{ji}$ ； $M$  代表與影像  $j$  有連結存在的查詢關鍵字數量； $r_{ji}$  代表的是在影像  $j$  與關鍵字  $i$  的相關因子； $\alpha$  為一個常數； $D_j$  是利用 [Rui99] 中做相關回饋計算所得的分數； $\beta$  與  $\gamma$  分別代表正例與反例計算結果所佔的權重； $N_R$  與  $N_N$  分別代表使用者所回傳的正例與反例的個數； $l_1$  為在影像  $j$  中與存在使用者回傳的正例影像中所共同擁有且不重複的關鍵字個數； $l_2$  為在影像  $j$  中與存在使用者回傳的反例影像中所共同擁有且互斥的關鍵字個數； $A_1$  與  $A_2$  分別代表正例與反例中互斥的關鍵字個數。利用  $l_1$ ， $l_2$ ， $A_1$  與  $A_2$  這四個參數可以判斷影像  $j$  與使用者所想要找的目標在文字意義上的吻合程度， $S_{jk}$  代表的是影像  $j$  與影像  $k$  之間利用影像低階特徵所計算而得的相似度。

### 2.3.2 經由相關回饋方式學習關鍵字之間的語意關連性

當進行影像標註的時候，通常在同一張影像中各個標註的關鍵字是被視為互相獨立沒有關係的。例如對於在河流中進行泛舟的影像，可能會給予「河流」、「泛舟」、「天空」與「人」等關鍵字，但其實這些關鍵字都是互相獨立的個體，它們之間沒有任何相關性可以被歸納。所以利用關鍵字進行影像檢索的時候，不像在作文字文件檢索時可以有 TF/IDF 這樣的工具可以使用。但是，當利用關鍵字進

行影像檢索的同時，除了以這些關鍵字為標註的影像之外，有一些與關鍵字相關的影像也應該被找出來。舉個例子，假使以「汽車」為關鍵字進行影像檢索，如果只針對跟「汽車」有關的影像文件進行檢索，對於一些以「Toyota」與「Ford」等關鍵字進行標註的汽車影像，將很難被檢索出來。在[Zhou02]中提出了利用使用者進行影像檢索的同時，依據使用者所提供的相關回饋資訊去建構出一個文字之間的概念相似矩陣（Concept Similarity Matrix），如圖 12：

	$W_1$	$W_2$	...	$W_{n-1}$	$W_n$
$W_1$	$S_{1,1}$	$S_{1,2}$	...	$S_{1,n-1}$	$S_{1,n}$
$W_2$	$S_{2,1}$	$S_{2,2}$	...	$S_{2,n-1}$	$S_{2,n}$
...	...	...	...	...	...
$W_{n-1}$	$S_{n-1,1}$	$S_{n-1,2}$	...	$S_{n-1,n-1}$	$S_{n-1,n}$
$W_n$	$S_{n,1}$	$S_{n,2}$	...	$S_{n,n-1}$	$S_{n,n}$

圖 12：概念相似矩陣[Zhou02]

其中， $W_n$  代表關鍵字  $n$ ； $S_{ij}$  代表關鍵字  $i$  與關鍵字  $j$  之間的相關程度。建構的過程可以利用方程式 12 計算使用者所回傳的資訊而進行修改。

$$S_{ij} = S_{ij'} + \max(f_i, f_j) \times (\min(f_i, f_j) - c_{ij})$$

方程式 12

其中  $S_{ij'}$  代表在上一回合中關鍵字  $i$  與關鍵字  $j$  之間的相關程度； $f_i$  與  $f_j$  分別代表在這一回合中關鍵字  $i$  與關鍵字  $j$  個別出現的頻率； $c_{ij}$  則是代表關鍵字  $i$  與關鍵字  $j$  共同出現在同一張影像下的頻率。透過概念相似矩陣，在使用者進行檢

索的時候，可以利用如 Hopfield Network[Hopfield82]或 Heuristic Clique Detection Algorithm[Zhou02]等方法幫助關鍵字進行語意式分群 (Semantic Grouping)，而得到關鍵字中的語意類別 (Semantic Classes)，進而可以讓系統找出更吻合使用者所想要的資料。



### 第三章 回饋式影像自動標註與檢索系統 (MRCOM)

本論文提出了一套自動化影像標註與檢索方法，稱為 MRCOM (Modified Region Based Co-occurrence Model)。MROCM 使用 Region Based 切割方式取得影像中的物件，改良[Mori99]中所提之 Co-occurrence Model 進行影像的標註，並且提供了使用者相關回饋的機制，結合使用者的概念與習慣，以提升檢索準確度。利用本系統的運作機制，將可以省去人工對影像進行標註所造成的困擾。此外，系統也提供傳統利用範例影像作檢索的功能。由於有時候以整張影像作為檢索的目標時，並非整張影像中的內容都是使用者所欲找尋的，所以系統提供了影像切割的功能，讓使用者指定影像中的物件，進而做到以物件作為查詢的對象。

首先在第一节描述系統主要架構；第二节將介紹所使用的影像特徵擷取方式；接著第三节到第五節說明系統中主要三個模組的建構方式，分別是訓練模組、標註模組、與使用者檢索模組；第六節將介紹如何透過群與關鍵字對照表來更新在訓練模組中每一個類別所代表的語意概念。

#### 第一節 系統架構

圖 13 為 MRCOM 的系統架構圖，系統共分為三大模組：

1. 訓練模組 (Training Module)：MRCOM 利用影像切割，取得影像中的區域 (Region)，將所取得區域視為影像之中的物件 (Object)。所以在訓練模組中首先會將所有影像作切割取得物件，接著利用 K-Means Clustering 將訓練資料集切割完成所取得的 Regions 進行分群的動作。分群完成後，同屬一群的 Regions 視為相同的 Blobs，代表它們有共同的語意概念 (Semantic Concept)，接著利用[Mori99]中所提出的



Co-occurrence Model 的方法去計算每一種 Blob 所代表的語意概念，得到代表的關鍵字機率向量 (Keyword Probability Vector)。

2. 標註模組 (Annotation Module)：對於尚未進行標註的影像，系統利用相同的影像切割方式取得影像中的區域，將這些區域對應到訓練模組中的所有群中心並且找尋距離最近的前 N 個群與它們之間分別的相似度，再將這些區域所對應到的群之關鍵字機率向量進行整合重新計算，取得最適合此影像的關鍵字機率向量，並且幫助影像進行標註。
3. 檢索與回饋模組 (Query and Feedback Module)：由於自動標註的關鍵字可能會跟使用者的感知有所差異或者是標註錯誤，所以在檢索與回饋模組中，提供使用者透過相關回饋的機制來提升系統效能。由於在訓練資料集所建立起來的每一個群的關鍵字機率向量可能不是很完整，所以透過使用者相關回饋的資訊去維護一個群與關鍵字之間的對照表，利用這個對照表學習得到一個更精確的群與關鍵字之間的相關程度，進而在使用者做下一次檢索的時候提供更佳的準確性。

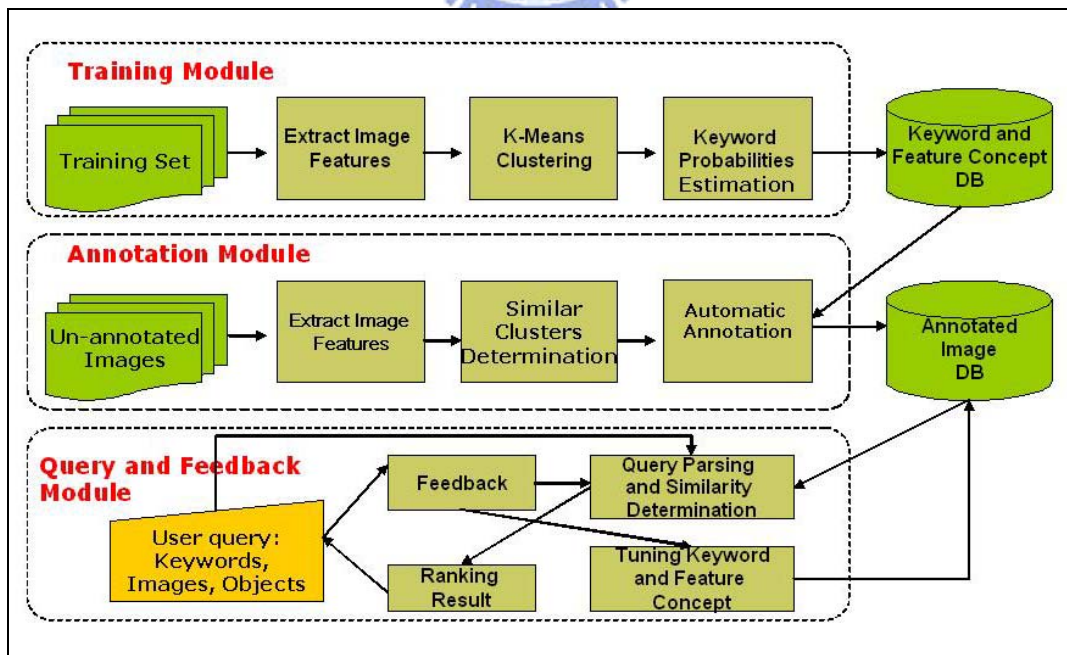


圖 13：MRCOM 系統架構圖

以下在第二節至第四節分別詳細說明訓練、標註、查詢與回饋等三大模組。

## 第二節 影像特徵擷取

本論文採用了兩種的影像特徵當作 Region 的特徵，分別是顏色與形狀。兩個 Regions 之間的相似程度  $sim_{i,j}$  可以用方程式 13 求得，其中  $w_c$  與  $w_s$  分別代表顏色與形狀相似度所佔的權重； $sim\_c_{i,j}$  與  $sim\_s_{i,j}$  則分別代表兩個 Regions  $i$  與  $j$  之間顏色相似度與形狀相似度。

$$sim_{i,j} = w_c \times sim\_c_{i,j} + w_s \times sim\_s_{i,j}$$

方程式 13

如圖 14，當已取得兩個 Regions  $i$  與  $j$  時，它們顏色與形狀相似度的計算過程分別敘述於 3.2.3.1 及 3.2.2.2。

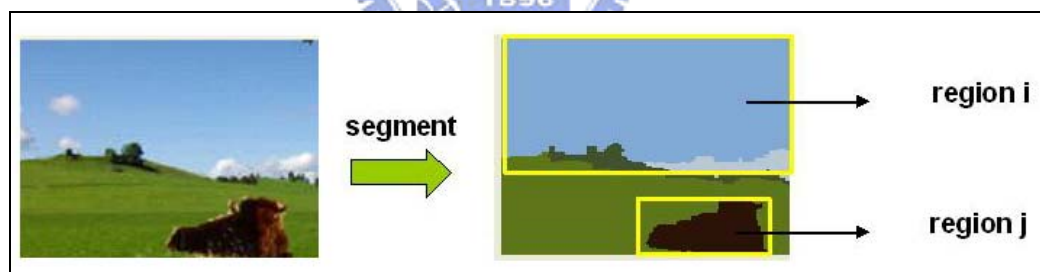


圖 14：影像作完切割之結果

### 3.2.1 顏色特徵擷取與相似度計算

在求取兩個 Regions 間顏色特徵之前，首先求得一個剛好將整個 Region 包圍住的最小矩形 (Basic Rectangle)，將此最小矩形所包圍的影像區域視為一個 Sub-image  $I$ ，並且利用[Ravishankar98]中所採用的顏色對照表，如表格 1，將所有點的顏色量化到 25 種顏色。由顏色對照表將某一點  $p$  對照到某一顏色的方法

是取出  $p$  的 RGB 值分別為  $x$ 、 $y$ 、 $z$ ，利用方程式 14 求得距離最近的顏色，即為  $p$  進行量化後的顏色。其中， $C_{iR}$ 、 $C_{iG}$  與  $C_{iB}$  分別為對照表中顏色  $i$  的 RGB 值。

Color	R	G	B	Color	R	G	B
Black	0	0	0	Plum	146	109	0
Sea Green	0	182	0	Teal	146	182	170
Light Green	0	255	170	Brown	182	0	0
Olive Green	36	73	0	Magenta	182	73	170
Aqua	36	146	170	Yellow Green	182	182	0
Bright Green	36	255	0	Flouro Green	182	255	170
Blue	73	36	170	Red	219	73	0
Green	73	146	0	Rose	219	146	170
Turquoise	73	219	170	Yellow	219	255	0
Dark Red	109	36	0	Pink	255	36	170
Blue Gray	109	109	170	Orange	255	146	0
Lime	109	219	0	White	255	255	255
Lavender	146	0	170				

表格 1：顏色量化對照表[Ravishankar98]

$$C_d = \text{MIN}_{i=1}^{25} \sqrt{(x - C_{iR})^2 + (y - C_{iG})^2 + (z - C_{iB})^2}$$

方程式 14

在進行顏色量化後，利用[Cinque99]中所提出來的顏色特徵取法，萃取出最小矩形中的特徵，分別有：

1. 每一種顏色  $k$  的點占 Sub-image  $I$  的整體比例  $h_I(k)$ 。  $h_I(k)$  可由方程式 15 得之；其中，  $\Lambda_k^I$  表示在 Sub-image  $I$  中顏色為  $k$  的點，  $n$  表 Sub-image  $I$  的寬，  $m$  表 Sub-image  $I$  的高：

$$h_I(k) = \frac{|\Lambda_k^I|}{n \times m}$$

方程式 15

2. 對所有顏色  $k$  的點，在 Sub-image  $I$  中座標分布的相對質心：  
 $b_I(k) = (\bar{x}_k, \bar{y}_k)$ 。其中，  $\bar{x}_k$  與  $\bar{y}_k$  分別可由方程式 16 與方程式 17 求得：

$$\bar{x}_k = \frac{1}{n} \frac{1}{|\Lambda_k^I|} \sum_{(x,y) \in \Lambda_k^I} x$$

方程式 16

$$\bar{y}_k = \frac{1}{m} \frac{1}{|\Lambda_k^I|} \sum_{(x,y) \in \Lambda_k^I} y$$

方程式 17

3. 對所有顏色  $k$  的點，在 Sub-image  $I$  的位置相對於  $b_I(k)$  的標準差  $\sigma_I(k)$ 。  
 $\sigma_I(k)$  可由方程式 18 得之。其中，  $p$  表在 Sub-image  $I$  中顏色為  $k$  的某一個點，  $d(p, b_I(k))$  為點  $p$  與其顏色在 Sub-image  $I$  分布的質心位置  $b_I(k)$  的歐基理得距離 (Euclidean Distance)。

$$\sigma_I(k) = \sqrt{\frac{1}{|\Lambda_k^I|} \sum_{p \in \Lambda_k^I} d(p, b_I(k))^2}$$

方程式 18

對於兩個 Regions  $i$  和  $j$  的顏色相似度，可以由方程式 19[Cinque99]求得。其中， $c$  表所有顏色數量。

$$sim_{-c_{i,j}} = \sum_{k=1}^c \min(h_i(k), h_j(k)) \times \left( \frac{\sqrt{2} - d(b_i(k) - b_j(k))}{\sqrt{2}} + \frac{\min(\sigma_i(k), \sigma_j(k))}{\max(\sigma_i(k), \sigma_j(k))} \right)$$

方程式 19

### 3.2.2 形狀特徵擷取與相似度計算

本論文共取了以下七種特徵當作 Region 的形狀特徵：

1.  $RS$ ：表示 Region 的面積。
2.  $RL$ ：表示 Region 的形狀周長。
3.  $R\_W$ ：表示之前提過最小可以將整個 Region 包圍起來之矩形寬。
4.  $R\_H$ ：表示之前提過最小可以將整個 Region 包圍起來之矩形高。
5.  $E\_mean$ ：表示 Region 邊緣上的每一個點到 Region 中心點（此中心點為 Region 內所有點之相對質心）長度之平均
6.  $E\_std$ ：表示 Region 邊緣上的每一點到 Region 中心點長度之標準差。
7.  $E\_mon1$ ：表示形狀變化的第一種 Moment，可由方程式 20 求之。

其中， $(x', y')$  代表 Region 的中心點座標。若點  $(x, y)$  屬於此 Region，則  $f(x, y)$  為 1，否則為 0。

$$E\_mon1 = \frac{1}{SL} \sqrt{\sum_x^w \sum_y^h \left( f(x, y) \times \left| \sqrt{(x - x')^2 + (y - y')^2} - E\_mean \right| \right)^2}$$

方程式 20

8.  $E\_mon2$ ：類似  $E\_mon1$ ，可由以下方程式 21 求之。

$$E\_mon2 = \frac{1}{SL} \sqrt{\sum_x^w \sum_y^h \left( f(x,y) \times \left| \sqrt{(x-x')^2 + (y-y')^2} - E\_mean \right| \right)^2}$$

方程式 21

接著，將上述 8 種特徵轉換為以下 5 種代表性特徵分別為： $e1 = \frac{SL}{RS}$ ，  
 $e2 = \frac{SL}{R\_W * R\_H}$ ， $e3 = \frac{E\_std}{E\_mean}$ ， $e4 = \frac{E\_mon1}{E\_mean}$ ， $e5 = \frac{E\_mon2}{E\_mean}$ 。則一個 region  
 的形狀特徵向量可以表示成  $(e1, e2, e3, e4, e5)$ 。

當要求得兩 Regions  $i$  與  $j$  之形狀相似度  $sim\_s_{i,j}$  可以由方程式 22 求得。其中， $e_a^i$  表示 Region  $i$  形狀特徵向量的第  $a$  個維度值。

$$sim\_s_{i,j} = \frac{1}{5} \sum_{a=1}^5 \frac{\min(e_a^i, e_a^j)}{\max(e_a^i, e_a^j)}$$

方程式 22

### 第三節 訓練模組

在[Mori99]中利用統計字詞共同出現在相似 Blocks 中的頻率，進而去取得每一個 Block 的語意概念，但是由於其做法只是簡單地將影像切割成數個等大小的矩形，將這些等大小的矩形視為是影像中的一個物件，這種影像切割的做法常常會將人們視覺上認定屬於同一物件的影像區域被切割散落在幾個不同的 Blocks 之間。所以本論文改採用 Region Based 的模式，將影像切割成跟人們視覺上比較吻合的物件，再來進行 Region 的分群動作，將類似的 Regions 歸類為相同的 Blobs。訓練模組主要流程圖如圖 15：

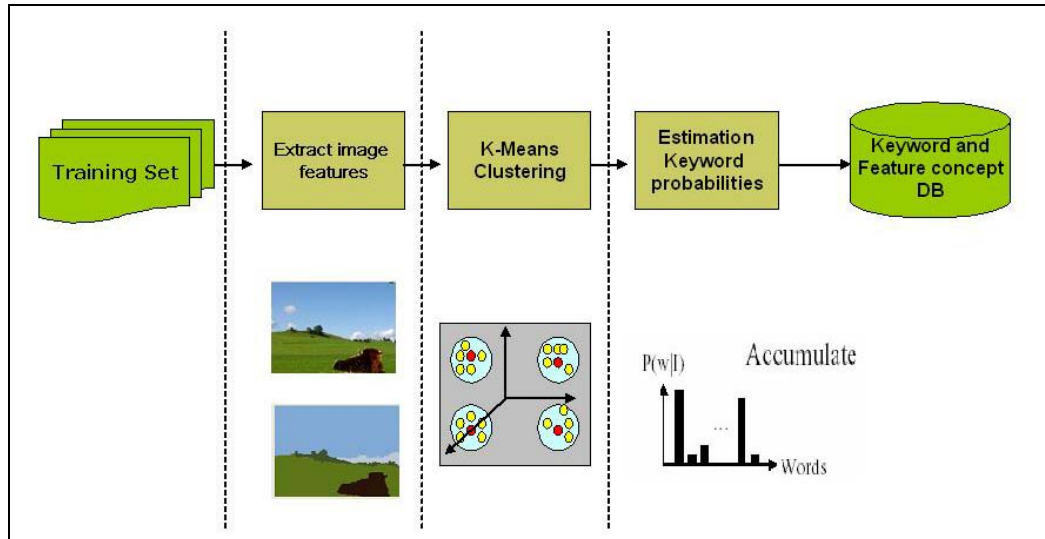


圖 15：訓練模組流程圖

本論文的訓練模組包含三個主要步驟：1. 影像切割，2. K-Means Clustering，3. 推論每一群之中的關鍵字發生機率。主要流程如下：

1. 將所有訓練資料集的影像資料先給予文字標註。
2. 將所有訓練資料集的影像利用 3.3.1 所提之影像切割方法進行切割。
3. 由訓練資料集內某影像  $I$  所切割出來之所有 Regions 將繼承  $I$  的所有標註文字作為其標註。
4. 以第二節中所提之方法取出所有切割完成的 Region 特徵。
5. 以 3.3.2 所提之 K-Means Clustering 演算法將所有 Regions 進行分群。
6. 以 3.3.3 所提之方法計算出每一群中每個關鍵字出現的機率。

### 3.3.1 影像切割

影像切割步驟係採用 Region Growing 的方法持續尋找影像中鄰近且顏色相近的點進行聚合，而得到影像之中的 Regions。

影像切割的流程如圖 16 所示，在切割的過程中，由 Region List ( $RL$ )來記錄目前有哪些 Regions 存在。過程中，會一直尋找在  $RL$  之中顏色最相近且位置相

鄰的兩個 Regions 做結合，一直到整張影像結合成一個 Region 為止。此外並要求影像切割完成後的 Region 個數最多不得超過一個門檻值  $t$ ，所以利用一個 State Record (S) 來做過程中的記錄。當存在於影像之中的 Regions 的個數小於  $t$  時，則將往後的每一次結合時影像的狀態記錄在 S 之中。在影像最後結合成一塊時，計算在 S 中所有狀態與原本影像之間的差異程度（計算每一個點與原始影像的顏色差異），並且計算在 S 中所有相鄰兩個狀態的變化量的梯度 (Gradient)，最後取出與下一狀態梯度為最大的狀態視為最佳影像切割。本論文實作上是將  $t$  設為 30。

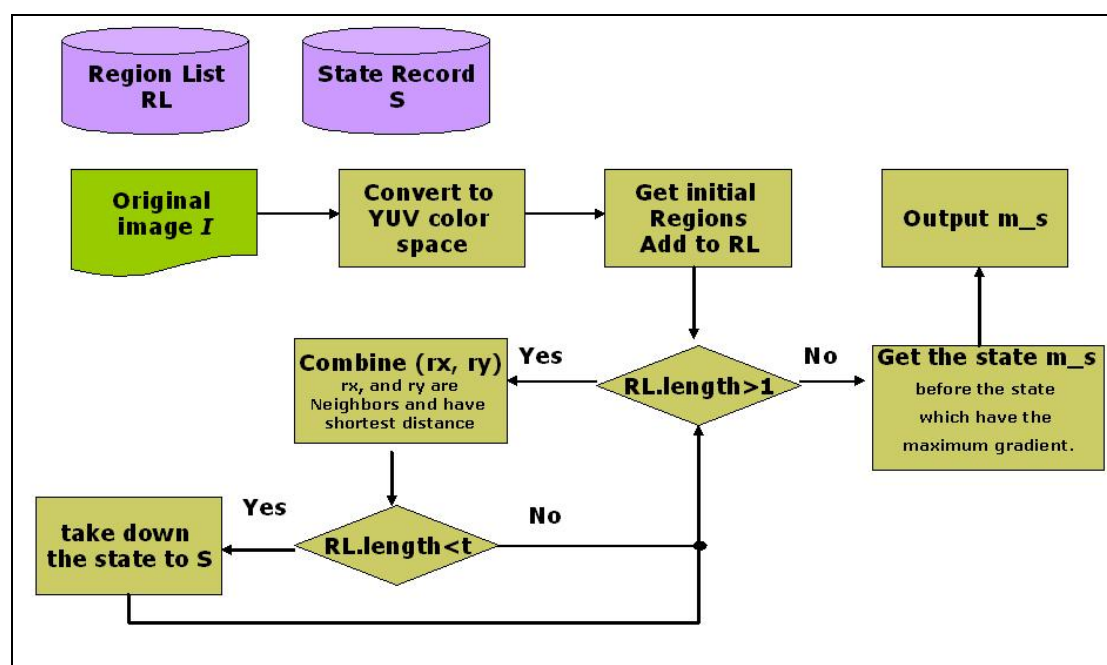


圖 16：影像切割流程圖

影像切割主要的步驟如下：

1. 將輸入的影像  $I$  從 RGB 顏色空間轉換到 YUV 顏色空間，轉換公式如方程式 23。



$$Y = 0.3R + 0.6G + 0.1B$$

$$U = Y - B$$

$$V = Y - R$$

方程式 23

2. 尋找步驟 1. 中轉換到YUV顏色空間的所有點，將位置相鄰且顏色相近的點聚集起來，成為在影像中的起始Regions，並且把起始Regions加入到 $RL$ 之中。判斷任兩點 $n_i$ 與 $n_j$ 在顏色上是否相似是利用計算 $n_i$ 與 $n_j$ 的YUV三個顏色頻道的距離，實作上是採用兩點的YUV值相減後的絕對值必須皆在一個門檻值之內。在這個步驟裡，會把門檻值設的比較嚴苛，主要是預防某一個Region隨意成長。
3. 判斷 $RL$ 中的Regions個數是否大於1，若是則繼續進行步驟4，若不是則跳到步驟6。
4. 選擇 $RL$ 內顏色最為相近且位置相鄰的兩個Regions進行結合，並且重新計算結合後所得新的Region的大小、顏色等相關資訊。其中，每一個Region $r$ 的顏色在結合的過程中是以所有屬於 $r$ 中的點之Y、U、V作平均後表示。所以在結合兩個Regions後，要重新計算出結合後屬於新Region之代表顏色。
5. 判斷目前 $RL$ 中的Region個數是否小於門檻值 $t$ 。若是，則把目前的狀態記錄在S中。回到步驟3。
6. 計算在S之中的狀態與原本影像的變化程度，連續的兩兩狀態取其變化程度的梯度(Gradient)，找出梯度值最大的前一個狀態 $m_s$ 將其視為最佳切割，取得 $m_s$ 並輸出。

### 3.3.2 K-Means Clustering 演算法

將影像  $I$  做完切割後，影像  $I$  中所有的 Regions 會繼承影像  $I$  中所有的標註關鍵字，接著將訓練資料集內所有的 Regions 進行分群 (Clustering)，再利用 [Mori99] 所提出的 Co-occurrence Model 進行每一群中語意概念的推導。本論文選用 K-means Clustering [McQueen67] 作為分群的工具。在分群的過程中，計算任一個 Region 與某一群  $c$  之間的相似程度，可將  $c$  的群中心視為某一特定 Region，接著透過方程式 13 計算 Region 與  $c$  之群中心相似度，流程如圖 17 所示：

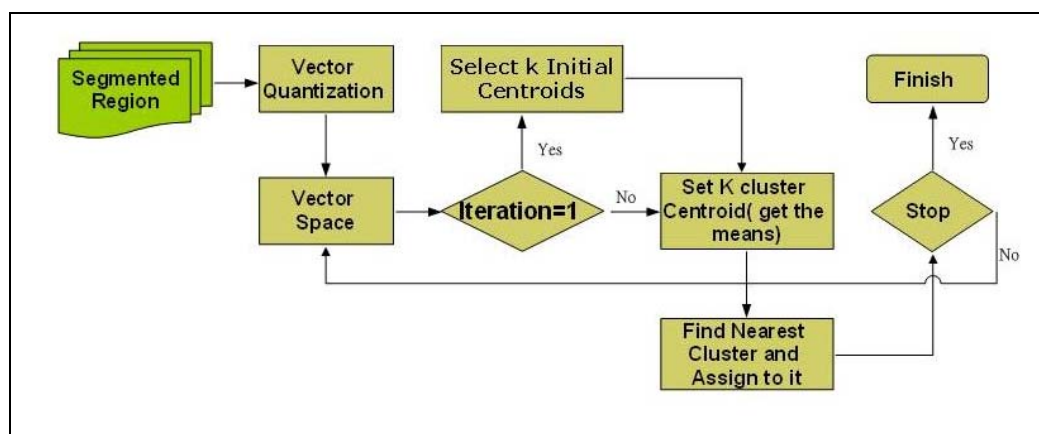


圖 17：K-Means Clustering 流程圖

以下簡略說明 K-means Clustering 演算法的運作流程，其中會訂定一個 Stop Condition 來決定要做到什麼程度。在實作上，我們將 Stop Condition 訂為當所有的 Regions 穩定地被歸類在某一個群為止。

1. 在整個做完向量量化後的向量空間裡任意選擇  $K$  個 Regions 為起始群中心。
2. 在向量空間中的每一個 Region 找到距離它們最接近的群中心，並把該點配置給最接近的群。
3. 判斷是否符合 Stop Condition 的條件，若符合則跳到步驟 5。
4. 求得每一群中所有 Regions 的平均向量當作新的群中心，回到步驟 2。
5. 完成並且輸出。

### 3.3.3 推論每一群中關鍵字出現的機率

本論文採用[Mori99]所提出的 Co-occurrence Model 去求取每一群的關鍵字出現機率，最後每一個群都會有一個關鍵字機率向量 (Keyword Probability Vector)，簡稱為  $KPV$ 。在群  $K$  中關鍵字  $N$  出現的機率  $P_{K,N}$  可以由方程式 24 推導得到，其中  $|w_i|$  表示關鍵字  $i$  出現在某一群中的次數； $M$  表示所有關鍵字個數。

$$P_{K,N} \approx \frac{|w_N|}{\sum_{i=1}^M |w_i|}$$

方程式 24

我們可以將群  $K$  的  $KPV$  表示成如下表格 2 的形式：

	$w_1$	$w_2$	$w_3$	...	$w_{n-2}$	$w_{n-1}$	$w_n$
Cluster $K$	$P_{K,1}$	$P_{K,2}$	$P_{K,3}$	...	$P_{K,n-2}$	$P_{K,n-1}$	$P_{K,n}$

表格 2：群  $K$  的關鍵字機率向量

在完成以上流程後，我們將各群的群中心視為其代表特徵，我們將這些特徵與每一群的  $KPV$  寫入到關鍵字與特徵概念資料庫。

## 第四節 自動化標註模組

對於一張未經標註的影像  $I$ ，將利用以下的自動化標註模組推論影像中的語意概念，給予標註與相關權重。首先，利用訓練模組裡的影像切割方法取出  $I$  的所有 Regions。由於在[Mori99]中採用將 Block 對應到最相似的群中，此種做法太過極端且不適當，因為只找最接近的群可能不足以代表該 Block 的語意概念。所以本論文改採找尋與 Region 最接近的前三群進行正規化 (Normalize)，所得到的

新的  $KPV$  視為此 Region 之  $KPV$ 。一個 Region  $r$  可以表示成

$r = (rid, r_{feature}, rc_1, rs_1, rc_2, rs_2, rc_3, rs_3, KPV_r)$ ，其中  $rid$  表示  $r$  的 Region id， $r_{feature}$  表示  $r$  的低階影像特徵， $rc_i$  表示與  $r$  第  $i$  相近的群， $rs_i$  表示與  $r$  第  $i$  相近的群由方程式 13 所計算出之相似程度。

進行正規化的方法為取得與 Region  $r$  最相似的前三群與對應的相似度值，並且取得此三群的  $KPV$ ，Region  $r$  的  $KPV$  可由方程式 25 求得。

$$KPV_r = \frac{\sum_{j=1}^3 rs_j \times KPV_{rc_j}}{\sum_{i=1}^3 rs_i}$$

方程式 25

其中  $rs_i$  代表 Region  $r$  與第  $i$  個相似群之相似度，分母表  $r$  與前三個相似群之相似度總和， $KPV_{rc_j}$  表示與 Region  $r$  與第  $i$  個相似群之  $KPV$ 。

在求得  $I$  中所有 Region 的  $KPV$  後，利用方程式 26 來推論關鍵字  $w$  出現在  $I$  之中的機率  $P(I_w)$ 。

$$P(I_w) = \frac{1}{N} \sum_{j=1}^N w_{IRj} \times P_{IRj,w}$$

方程式 26

其中， $w_{IRj}$  表示在  $I$  中，Region  $j$  所佔的權重； $N$  表示在影像  $I$  中的 Regions 個數； $P_{IRj,w}$  表示在影像  $I$  第  $j$  個 Region 的關鍵字機率向量中相對於關鍵字  $w$  的值。因為存在於影像中央的物件，在人們的感官中會將它認為是這一張影像中的

主題或是扮演比較重要的角色，所以在此設定了如果 Region  $j$  是位於  $I$  之中央時，則將其權重  $w_{IR_j}$  設為  $1+\alpha$  ( $\alpha \geq 0$ )，其他的 Region 則設為 1。以下是判別位於影像中央的 Region 流程：

1. 將影像切為九個等大小之矩形 (如圖 18)。
2. 計算所有 Region 中的點位於第 5 個矩形範圍內占整個 Region 的比例。
3. 在步驟 2 中比例最大的 Region 即位於中央之 Region。

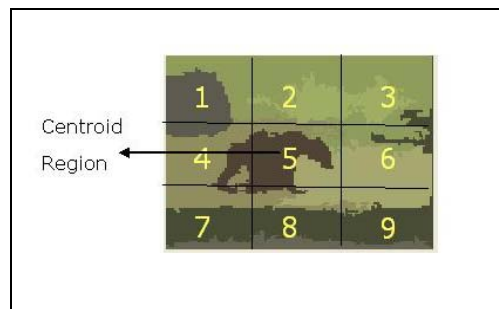


圖 18：判別位於中央之區塊圖

在本論文實作中是將  $\alpha$  設定為 0.3。在套用方程式 28 的時候，也會去記錄在  $I$  中，相對於任一個關鍵字  $w$ ，在哪一個 Region 中， $w$  發生的機率是最大的 (由 KPV 可知)，並記錄其 Region id。這個記錄在此稱之為  $max\_prw$ ，主要是用來紀錄整張影像中相對於關鍵字  $w$  最具代表性的區塊。 $max\_prw$  表示法為  $(image, word, rid)$ ，其中  $image$  表示此影像名稱。若是  $w$  在所有 Regions 的出現機率皆為 0 時，則將其  $max\_prw$  的  $rid$  設為 -1。

對於任一影像  $I$  的標註資訊有：

1. 每一個關鍵字  $w$  出現在  $I$  中的機率。
2. 在  $I$  中，每一個關鍵字  $w$  所對應的  $max\_prw$ 。

所以最後每一張影像的標註表示法如圖 19：

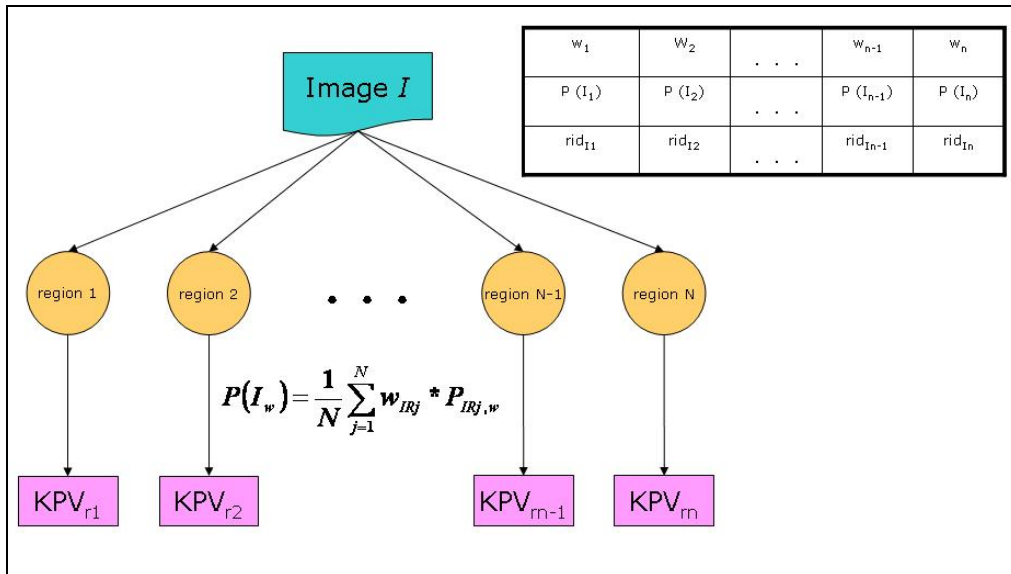


圖 19：影像標註表示法示意圖



## 第五節 使用者查詢與回饋模組

在完成自動化影像標註後，每張影像中都有跟字典裡每一個關鍵字的相關程度權重值。當使用者進行關鍵字影像檢索的時候，使用者可輸入多個關鍵字進行檢索。由於利用機器學習的方法自動化幫助影像加入標註很難達到百分之百的準確度，所以本論文設計了一套利用使用者相關回饋的機制進行學習，提升系統效能。另一方面，由於在訓練模組裡面所採用的資料有可能與欲進行標註的影像有所差異，以致於學習效果有限，所以我們也定義了一個群與關鍵字的關連性對照表 (Cluster-Keyword Association Map) -CKAM。CKAM 主要是利用收集使用者每次檢索所回饋的資訊調整群與關鍵字之間的關連性，以提升系統檢索效能。此外，由於在 CKAM 中群與文字之間的關連性是學習使用者的概念而來，所以在系統使用一段時間後，CKAM 的群與文字之間的關連性資訊將可取代先前在訓練模組所得的每一個群的  $KPV$ 。CKAM 的表示法如下圖 20：


$$\begin{array}{r} \begin{array}{c} w_1 \\ w_2 \\ \dots \\ w_{n-1} \\ w_n \end{array} \\ \left[ \begin{array}{cccccc} c_1 & WCS_{1,1} & WCS_{1,2} & \dots & WCS_{1,n-1} & WCS_{1,n} \\ c_2 & WCS_{2,1} & WCS_{2,2} & \dots & WCS_{2,n-1} & WCS_{2,n} \\ \vdots & \dots & \dots & \dots & \dots & \dots \\ c_{k-1} & WCS_{k-1,1} & WCS_{k-1,2} & \dots & WCS_{k-1,n-1} & WCS_{k-1,n} \\ c_k & WCS_{k,1} & WCS_{k,2} & \dots & WCS_{k,n-1} & WCS_{k,n} \end{array} \right] \end{array}$$

圖 20：群與關鍵字關連性對照表(CKAM)

其中  $wcs_{i,j}$  表示群  $i$  與關鍵字  $j$  之間透過使用者回饋所得的權重值，在剛開始的時候，將表中的每一個元素全部設定為 1。每當使用者對某關鍵字進行檢索時，將透過 CKAM 所回傳的資訊去評估哪幾群的區塊與使用者所想要檢索的關鍵字關聯性比較高，進而可以提供進行影像檢索時用來計算影像與文字關連性的重要資訊。

使用者進行檢索的過程如圖 21：

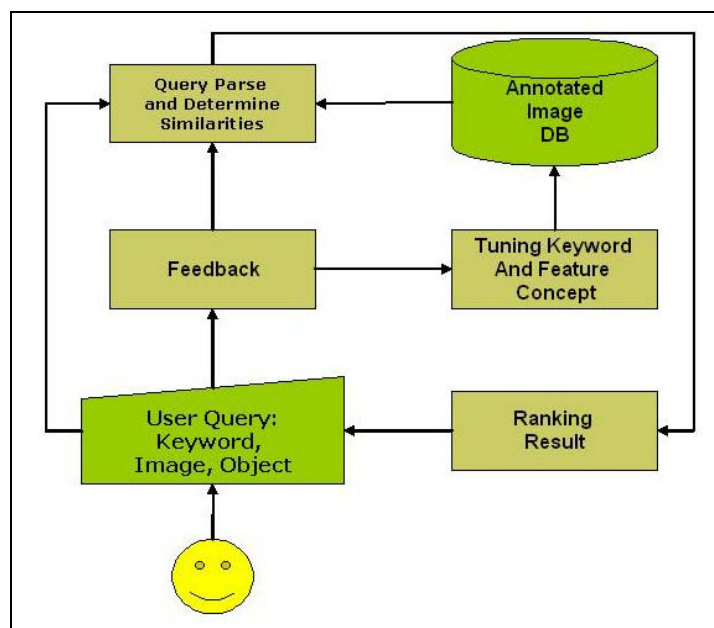


圖 21：使用者查詢與回饋模組

### 3.5.1 使用者檢索



當使用者輸入一個關鍵字集合  $W = \{w_1^i, w_2^i, \dots, w_m^i\}$  進行檢索的時候，系統會將影像資料庫內的所有影像  $I$  利用方程式 27 進行與  $W$  之相關程度的計算：

$$SIM(I, W) = \frac{1}{m} \times \sum_{i=1}^m \left\{ \beta \times P(I_{w_i^i}) + \gamma \times \left[ \frac{1}{N} \sum_{j=1}^N SIM(r_j, w_i^i) \right] \right\}$$

方程式 27

方程式 27 主要分為兩部份：

1. 利用自動標註模組所計算出來的影像與關鍵字之間的關聯性。
2. 透過 CKAM，計算出影像中的區塊與使用者查詢文字的相關程度。



其中， $P(I_{w_i})$ 表示第  $i$  個查詢關鍵字  $w_i$  出現在影像  $I$  中的機率； $SIM(r_j, w_i)$  表示影像  $I$  中的第  $j$  個區塊  $r_j$  與關鍵字  $w_i$  透過 CKAM 所計算而得的相關程度，經由 CKAM 可以知道哪些種類的區塊與此關鍵字是比較相關的； $\beta$  與  $\gamma$  分別表示此二部分所佔權重。 $SIM(r_j, w_i)$  的計算公式如方程式 28：

$$SIM(r_j, w_i) = \frac{1}{\sum_{i=1}^k wcs_{i,k}} \left( \frac{\sum_{y=1}^3 rs_{jy} \times wsc_{i,rc_{jy}}}{\sum_{x=1}^3 rs_{jx}} \right)$$

方程式 28

其中， $wcs_{i,k}$  表示在 CKAM 中，關鍵字  $w_i$  與群  $k$  所對應的權重值； $rs_{jx}$  表示與 Region  $r_j$  第  $x$  個群的相似程度； $wsc_{i,rc_{jy}}$  表示與 Region  $r_j$  第  $y$  個相似的群與關鍵字  $w_i$  在 CKAM 上所對應的權重值。

透過方程式 27 的計算，可以取得所有影像與使用者所下的查詢關鍵字之相似度，透過排序的方法取出相似程度較高的影像回傳給使用者。

### 3.5.2 使用者回饋

由於透過上面的檢索過程所檢索出來的影像，可能有些未必是正確的或者不是使用者所想要的，所以將透過使用者相關回饋的機制來提升系統效能。本論文提出的回饋機制主要分為三個部份：

1. 修改使用者所指定影像之  $KPV$ 。
2. 修改 CKAM。
3. 回傳新的檢索結果給使用者。

### 3.5.2.1 修改使用者所指定影像之 $KPV$

在這個過程中，主要是去收集使用者回傳的所有正例與反例的影像，進而調整這些影像與查詢關鍵字之間的機率，流程如圖 22：

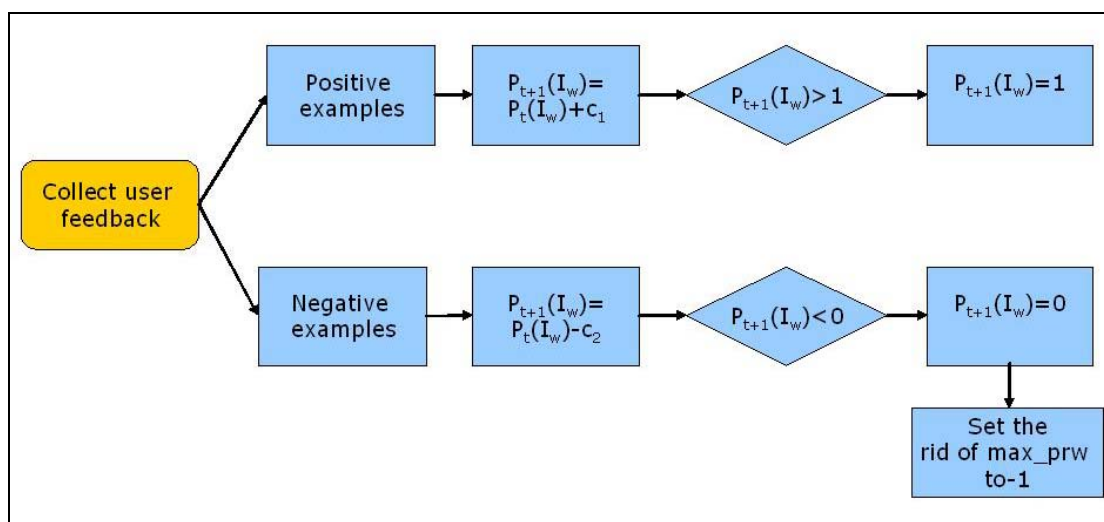


圖 22：修改使用者所指定影像之  $KPV$  流程圖

整個流程分為兩部份：1. 調整正例影像之關鍵字機率；2. 調整反例影像之關鍵字機率。

對於任一個使用者查詢的關鍵字  $w$ ，相對於任一使用者所指定為正例之影像  $I$ ，調整其  $KPV$  中相對於  $w$  的機率。調整的方法是將原本的機率加上  $c_1$ 。由於所有關鍵字出現於影像中的值介於 0 到 1 之間，所以在加上  $c_1$  之後，系統會去檢查其值是否超過 1；若超過 1 則將其歸為 1。

同樣地，對於任一個使用者查詢的關鍵字  $w$ ，相對於任一使用者所指定為負例之影像  $I$ ，調整其  $KPV$  中相對於  $w$  的機率。調整的方法是將原本的機率減掉  $c_2$ 。並且在減完  $c_2$  後，檢查其值是否小於 0；若小於 0 則將其歸為 0，並且將  $I$  中相對於  $w$  之  $max\_prw$  之  $rid$  欄位設定為 -1。

### 3.5.2.2 修改 CKAM

透過修改 CKAM 的動作，可以取得使用者所認定較接近查詢關鍵字的 Region 類型，進而可以在下一次檢索時用來做參考之資訊。主要流程如圖 23：

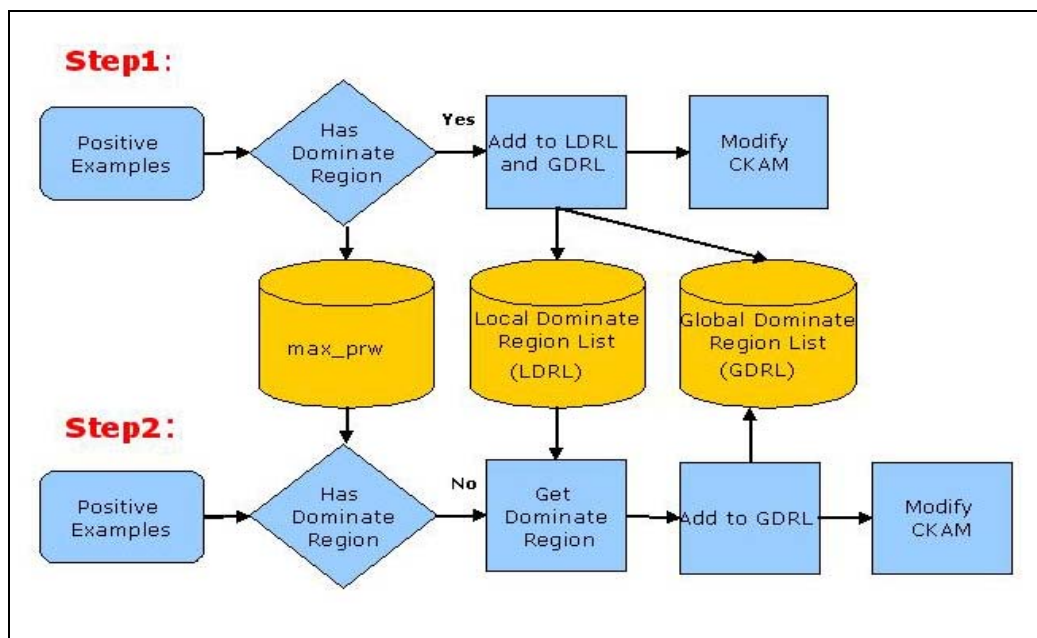


圖 23：修改 CKAM 流程圖

在這個流程中，針對於任一查詢關鍵字集合中的關鍵字  $w$ ，系統將透過收集所有使用者所認定相關的影像，並且取得所有相關影像中與查詢關鍵字相關的  $max\_prw$ ，進而得知影像中相對於關鍵字  $w$  最具代表性的 Region。由於在使用者所指定的影像中，某些影像的  $max\_prw$  之  $rid$  欄位被設定為-1，即影像原本未有任何 Region 相對於  $w$  是比較有代表性的。所以在這個過程中我們將分為兩個步驟處理：

1. 取得所有正例影像中，相對於  $w$  於之  $max\_prw$  的  $rid$  欄位不等於-1 的  $rid$ ，同時將其 Region 的特徵加入到 Local Dominate Region List (LDRL) 與 Global Dominate Region List (GDRL) 中，並調整 CKAM。

2. 對於任一個在正例影像中相對 $w$ 於之 $max\_prw$ 的 $rid$ 欄位等於-1 的影像  $I_p$ ，取得 $I_p$ 中與LDRL的Region Feature有最大相似度的Region  $r$ ，並將 $I_p$ 相對於 $w$ 之 $max\_prw$ 的 $rid$ 設定為 $r$ 之 $rid$ ，並且調整CKAM。

其中，透過任一正例影像  $I$  調整 CKAM 的過程如下：

1. 取得  $I$  中相對於  $w$  的  $max\_prw$  中的  $rid$ 。
2. 取得編號為  $rid$  之 Region 的前三相似類別的群編號 ( $rc_1, rc_2, rc_3$ ) 與對應之相似度 ( $rs_1, rs_2, rs_3$ )。
3. 將此三個相似度值加到 CKAM 中對應的  $wcs_{rc_i,w}$ 。其中， $wcs_{rc_i,w}$  表示在 CKAM 中與 Region  $r$  第  $i$  相似之類別編號相對於關鍵字  $w$  的值。

### 3.5.2.3 回傳新的檢索結果給使用者

在經過上面的動作後，系統將重新計算所有影像與使用者回饋的資訊之相關程度。對於任一影像  $I$  經過使用者回饋的過程後與使用者所欲查詢資料相似度可由以下方程式 29 得知：

$$SIM(I, W) = u \left( \frac{1}{m} \sum_{i=1}^m P(I_{w_i}) \right) + v \left( \max_{j=1 \text{ to } N} \left( \frac{1}{|G|} \sum_{k=1}^{|G|} S_{r_j, G_k} \right) \right)$$

方程式 29

其中， $u$  與  $v$  皆為常數； $m$  表整個查詢關鍵字  $W$  中關鍵字個數； $P(I_{w_i})$  表示關鍵字  $w_i$  在影像  $I$  中出現的機率； $G$  表示在修改 CKAM 步驟中加入到 GDRL 中的 Regions 集合； $|G|$  表示在  $G$  中 Regions 的個數； $S_{r_j, G_k}$  表示存在影像  $I$  中之 Region  $r_j$  與整體  $G$  中第  $k$  個 Region  $G_k$  的相似程度。

## 第六節 更新訓練資料集中每一群之 KPV

透過使用者回饋的過程修改 CKAM 中的資訊，在系統經過一段時間的使用後，CKAM 中的文字與每一群之間的關連性將比較適合使用者的概念，所以我們可以將其資訊用來取代訓練資料集中每一群的 KPV。其中，任一個關鍵字  $w$  出現在群  $K$  中的機率  $P_{K,N}$  可以由方程式 30 進行修改：

$$P_{K,N} = \frac{\left( \frac{wCS_{K,N}}{t_N} \right)}{\left( \sum_{i=1}^n \frac{wCS_{k,i}}{t_i} \right)}$$

方程式 30

其中， $t_i$  表示關鍵字  $w_i$  被檢索過的次數； $wCS_{K,N}$  表示 CKAM 中類別  $C_K$  與關鍵字  $w_N$  所對應的值。



## 第四章 實驗與討論

本章將說明本論文所提出之回饋式影像自動標註與檢索系統的性能與實驗結果。第一節簡介實驗所用的影像資料分類，並說明所設計的實驗；第二節介紹實驗所採用的評估方法。第三節闡述 MRCOM 與將整張影像視為一個物件作標註的方法效能比較與評估；第四節介紹 MRCOM 跟[Mori99]中分別採用 Block Based 作法與 Region Based 效能比較；第五節是對本論文所提出之使用者回饋機制效能進行評估。

### 第一節 實驗資料與分類

在本論文實驗中，總共收集了 1200 張影像。其中，800 張影像用來做為訓練資料集中的影像資料，剩下的 400 張則是用來進行測試。將所有影像分為六大類，每一大類下又包含了一個以上的小類，分類如下：

1. 風景類：山、太陽、河流、瀑布、原野
2. 交通類：車、飛機、船、火車鐵路
3. 動物類：熊、老虎、馬、鹿
4. 植物類：花卉
5. 飾品類：黃金飾品
6. 建築類：建築物

在訓練過程中，針對訓練資料集的 800 張影像分別給予標註，每張影像標註有 1 到 7 個關鍵字。在標註完畢後，過濾掉不常出現的關鍵字，總共取得 64 個關鍵字可在影像自動化標註時使用。

### 第二節 評估方法

本論文採用兩種方法來評估實驗結果，分別是 11-point Interpolated Measure 與 Mean Average Precision (MAP)。

利用 11-point Interpolated Measure 可以看出整體效能的查全率 (Recall) 與準確率 (Precision) 間對照的關係。查全率就是計算所有回傳的答案中，對的答案佔所有資料中真正答案的比例，其計算公式如方程式 31[Marc04]。準確率就是在所有回傳的答案中，正確答案所佔的比例，其計算公式如方程式 32[Marc04]。

$$\text{Recall } r = \frac{\# \text{correct\_found\_answer}}{\# \text{possible\_existing\_answer}}$$

方程式 31

$$\text{Precision } s = \frac{\# \text{correct\_found\_answer}}{\# \text{all\_found\_answer}}$$

方程式 32

舉例而言，假設針對某個查詢條件全部正確的答案共有 10 個。若系統在某次的檢索中，系統回傳了 8 個答案，其中有 3 個是正確的，則在此次的檢索中，查全率為 0.3，準確率為 0.375。

一般而言，準確率會隨著查全率的增加而降低。所以本論文利用計算 MAP (Mean Average Precision) 的方法來做整體比較。MAP 是一種計算整體檢索準確率的比較準則，在此是以各個類別在 11-point Interpolated Measure 中的平均準確度為其 MAP，其計算公式如方程式 33；其中  $P_i$  表示在查全率為  $\frac{i}{10}$  時相對的準確率。

$$\text{MAP} = \frac{1}{11} \sum_{i=0}^{10} P_i$$

方程式 33

### 第三節 MRCOM 與整體影像內容標註方法之效能評估

在內容式影像檢索系統中，大部分是以整張影像中的低階特徵進行影像的檢索，找出整體特徵最為相似的影像。如果以內容式影像檢索的角度進行影像標註，比較直覺的做法就是將一張尚未進行標註的影像  $I$ ，找出在訓練資料集中與  $I$  最為相似的影像，並且將其標註配置給  $I$ 。一般來說，內容式影像檢索系統中利用整體影像特徵進行相似度的比較在風景類方面有不錯的效果，這主要是由於風景類影像的主題就是整張影像的畫面，所以有同樣風格的風景類影像在整張影像的特徵都十分相近。但是以這種方法在找尋以特定物件為主題的影像，效果卻不是十分理想。比如說，一台紅色的車子可能出現在各種場合（如：停車場或是馬路上等...），但是車子可能只佔了整張影像的一小塊區域。此時利用整張影像的特徵進行檢索，車子在影像中所佔的比例如果太小，在相似度計算的時候其重要性很難被突顯出來（如在山中行駛的車子與停在停車場中的車子，由於兩者背景相差甚多，所以很容易被判別為兩者不相似）。相對的，由於本論文的方法是取出所有位於影像中的物件，因此我們預期在不同場合中的車子將可以被尋找出來。

本實驗主要是要印證特定的物件以 MRCOM 方式進行標註將可以得到比用整張影像進行標註更好的效果。因此，本論文設計了一種整張式影像標註的方法稱之為 IBAM (Image Based Annotation Model)，步驟如下：

1. 取出[Cinque99]中所提到的影像特徵。
2. 取得所有訓練資料集中與此影像相似程度最高的前十張影像。
3. 利用方程式 34 計算每一個關鍵字  $w$  與影像  $I$  的相關程度  $SIM(I, w)$ 。

$$SIM(I, w) = \sum_{j=1}^{10} s_j \times f(t_j, w)$$

方程式 34



其中， $t_j$  表示  $I$  與訓練資料集中第  $j$  個相似影像； $s_j$  表示  $I$  與  $t_j$  的相似度；若  $w$  為  $t_j$  標註文字中的關鍵字，則  $f(t_j, w)$  為 1，否則為 0。

在實作 MRCOM 本論文是以 K-Means 為分群的工具，並將 K 設定為 300。其中，用來計算 Regions 之間相似度的方程式 13，在實作時將將方程式 13 中的  $w_c$  設定為 0.8、 $w_s$  設定為 0.4。由於在方程式 13 中顏色相似度的範圍介於 0 到 2 之間，形狀相似度介於 0 到 1 之間，所以方程式 13 所計算出來的相似度值將介於 0 到 2 之間。

我們從六大類中，各取出一個代表性關鍵字來評估 MRCOM 與 IBAM 的效能。以下是各關鍵字在兩系統下執行的 11-point Interpolated Measure Graph：

風景類：以「河流」為代表性關鍵字。

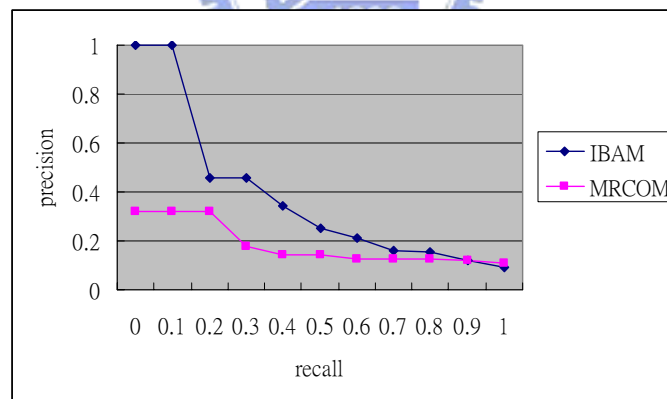


圖 24：IBAM 與 MRCOM 檢索關鍵字「河流」之 11-point Interpolated Measure Graph

交通類：以「車」為代表性關鍵字。

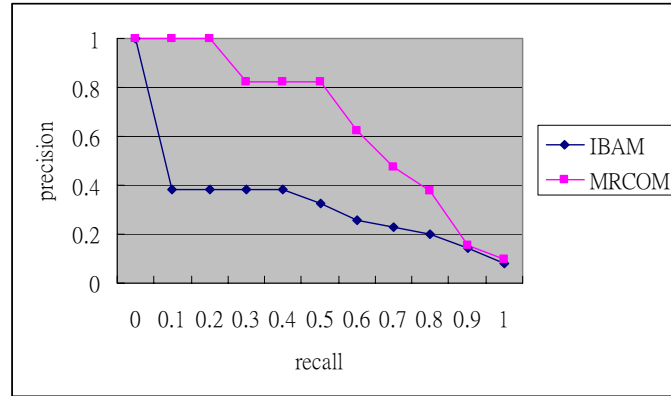


圖 25 IBAM 與 MRCOM 檢索關鍵字「車」之 11-point Interpolated Measure Graph

動物類：以「鹿」為代表性關鍵字。

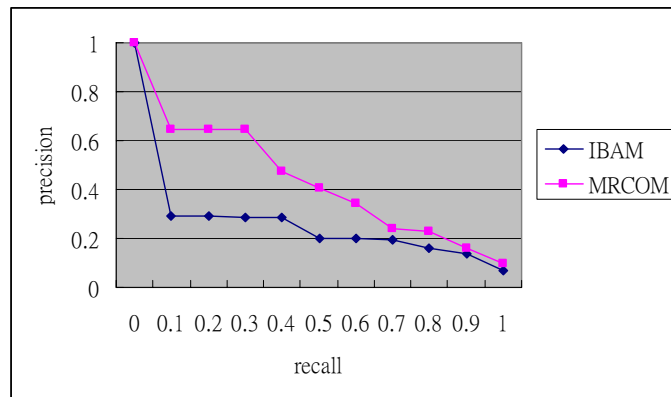


圖 26：IBAM 與 MRCOM 檢索關鍵字「鹿」之 11-point Interpolated Measure Graph

植物類：以「向日葵」為代表性關鍵字。

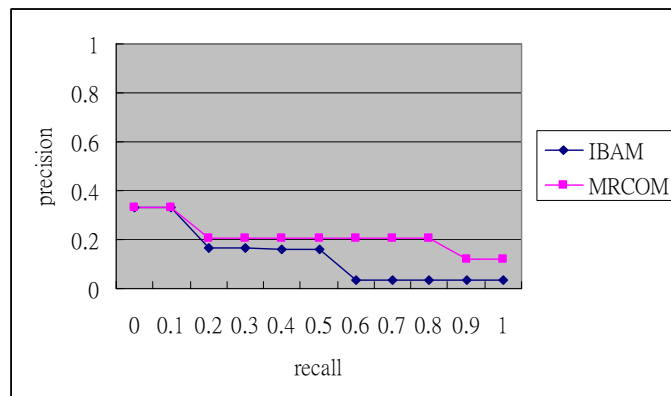


圖 27：IBAM 與 MRCOM 檢索關鍵字「向日葵」之 11-point Interpolated Measure Graph

飾品類：以「戒指」為代表性關鍵字。

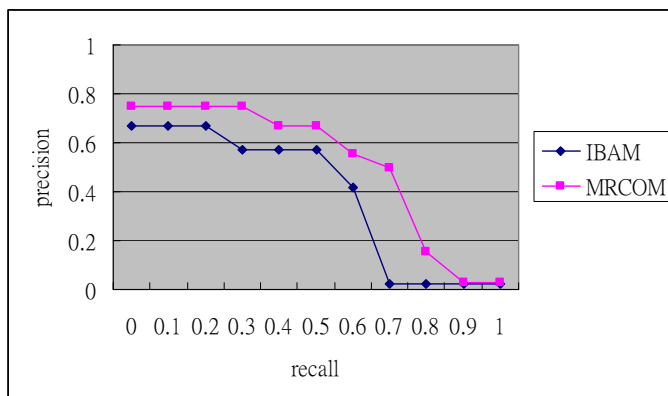


圖 28：IBAM 與 MRCOM 檢索關鍵字「戒指」之 11-point Interpolated Measure Graph

建築類：以「建築物」為代表性關鍵字。

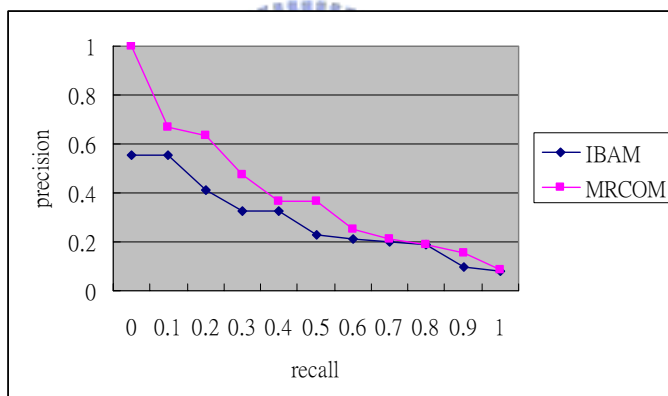


圖 29：IBAM 與 MRCOM 檢索關鍵字「建築物」之 11-point Interpolated Measure Graph

以下是我們整理上述六種查詢條件所得之 MAP：

	河流	車	鹿	向日葵	戒指	建築物
IBAM	38.47%	34.29%	28.30%	13.52%	38.34%	28.87%
MRCOM	18.41%	65.41%	44.34%	21.54%	50.94%	40.02%

表格 3：IBAM 與 MRCOM 之 MAP 比較

由表格 3 得知以「河流」為關鍵字進行檢索時，IBAM 得到了比 MRCOM

較佳的效果。這主要是因為描寫河流景色的影像大都有類似的背景 (如樹林、山或石頭等...), 並且河流中的水大部分有著固定的顏色。相對的, MRCOM 是以取影像中物件的方式判斷影像的語意概念, 但是由於與河流相關的影像中, 河流旁的景色常常會反照在河流的水面上, 所以整個週遭景色都結合在一起, 導致很難去判別哪個部分才是河流。

但是在其它五類檢索中, MRCOM 都獲得比 IBAM 有更佳的效果。以「車」為關鍵字的檢索中, 由於 MRCOM 可以將影像中真正屬於「車」的部分切割出來, 所以很容易得到比較好的效果。IBAM 則因為車子所佔的比例如果太小, 以整體的影像來看, 車子所佔的部分相對的很容易被忽略, 導致以車為主題之影像很容易變成以背景為主題進行標註的結果。比如說, 一張描寫在原野上行駛的汽車被標註成原野的可能性會比標註成汽車的可能性要來的大。

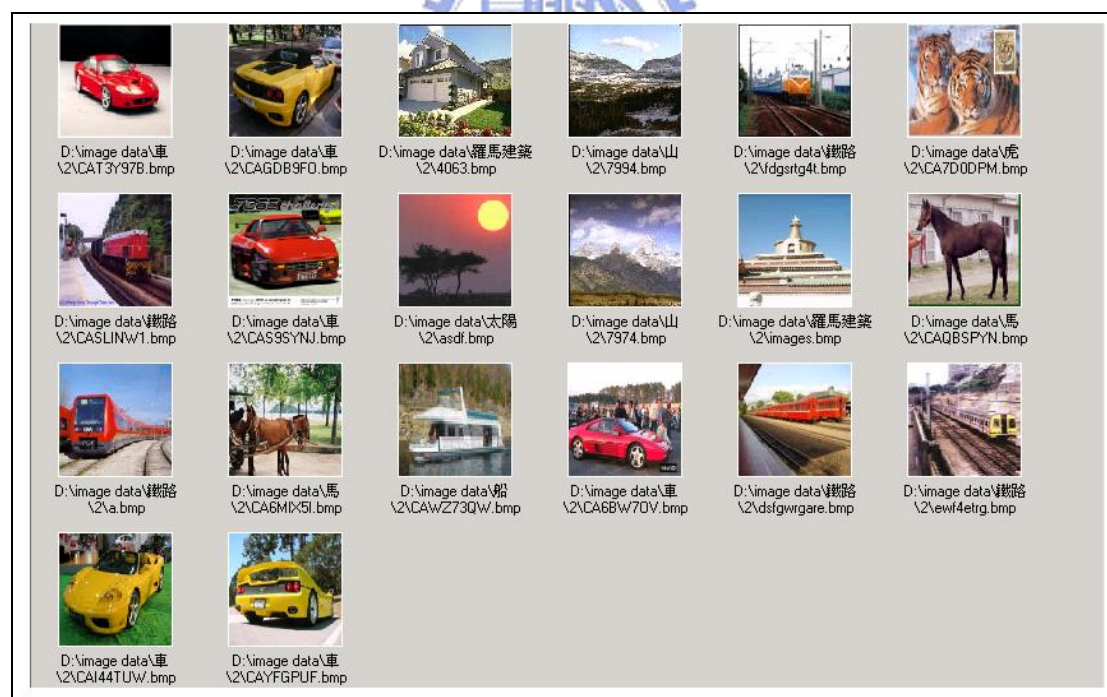


圖 30：以 IBAM 檢索「車」回傳之前 20 張影像

圖 30 是以車為關鍵字利用 IBAM 檢索所得到的前 20 張影像, 可觀察到在這 20 張影像中, 大部分的影像都是以白色為主要的背景顏色, 或是以樹林山脈

為背景所以帶點綠色。由這個觀察，即可驗證之前所推測的以 IBAM 方式進行標註時，背景將扮演很大的影響要件。在圖 30 中，被檢索出來的有關車的影像大部分都是以車為主題，幾乎車子佔了整張影像蠻大的比例，所以在這邊被檢索了出來。相對地，圖 31 是以車為關鍵字利用 MRCOM 檢索所得到的前 20 張影像，與圖 30 相比較，可以看到 MRCOM 所檢索出來的影像不像 IBAM 檢索出來大部分只是具有相同背景的影響而已。在圖 31 中，可以看出有些在影像中所佔整體影像比例不高的車子的影像也被檢索出來了，這便足以說明 MRCOM 可以避免因為背景的影響而模糊了主題的問題。其它四個類別分別也是以物件為主題進行檢索的測試，由表格 3 可知 MRCOM 都扮演著比 IBAM 更加好的效能，這樣的結果與[Carson99]中的實驗有相同的印證。

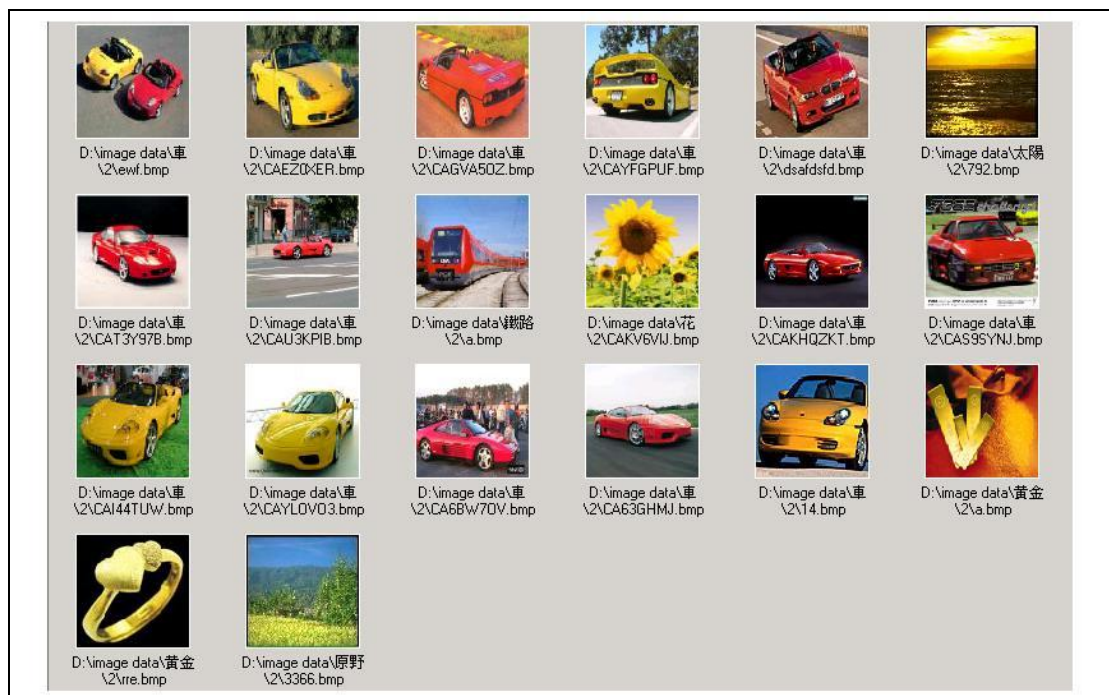


圖 31：以 MRCOM 檢索「車」回傳之前 20 張影像

#### 第四節 自動化標註方法效能評估

在本實驗中，將對 MRCOM 與以下兩種方法作一個評估及分析：

1. BCOM (Block Based Co-occurrence Model)：這個做法主要是利用 Block Based 影像切割取得影像中的物件，並且套用[Mori99]中所提出的 Co-occurrence Model 進行影像的標註。
2. RCOM (Region Based Co-occurrence Model)：這個做法主要是利用本論文第三章所提到的「Region Growing」的方式進行影像切割取得影像中的 Region，並視每一個 Region 為一個物件，再套用[Mori99]中所提出的 Co-occurrence Model 進行影像的標註。

實驗設計方面，我們從六大類中取出數個較能代表該類別的關鍵字，經由 BROM、RCOM、MRCOM 三個方法進行檢索，並分別將所得之效果進行評估。

每一大類代表性的關鍵字分別如下：

1. 風景類：山、太陽、河流、瀑布、原野
2. 交通類：車、飛機、船、火車、鐵路
3. 動物類：熊、老虎、馬、鹿
4. 植物類：玫瑰、向日葵、百合
5. 飾品類：金塊、項鍊、戒指
6. 建築類：建築物、羅馬建築

本實驗將透過上述各種類別代表性關鍵字分別以此三種方法所執行的系統效能進行評估，並且採用 11-point Interpolated Measure 的方法將執行結果繪製成 recall-precision 的對照圖，並且取得相對應的 MAP。每一個大類中的 11-point

Interpolated Measure 與 MAP 的取法主要是各個測試條件下所得效能之平均值。舉個例子來說，假設在查全率為 0.1 的時候，風景類中五個查詢條件之準確率分別為 0.8、0.75、0.75、0.8 與 1.0，則整體類別在查全率為 0.1 時所對應的準確率為  $(0.8+0.75+0.75+0.8+1.0)/5=0.82$ 。若此五個條件取得之 MAP 分別為 0.3、0.45、0.4、0.3 與 0.5，則整體類別之 MAP 為  $(0.3+0.45+0.4+0.3+0.5)/5=0.39$ 。

以下是三種方法分別在六大類別中所得的 11-point Interpolated Measure Graph、MAP 與分析：

1. 風景類：山、太陽、河流、瀑布、原野。

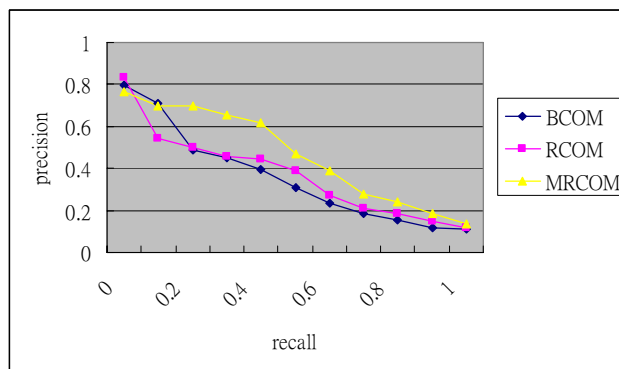


圖 32：風景類之 11-point Interpolated Measure Graph

MAP：

BCOM	RCOM	MRCOM
35.93%	37.31%	<b>46.65%</b>

表格 4：風景類之 MAP

由於在風景類的影像中，物件的變化程度較大，如不同影像中的瀑布有著不同的樣子，有的是放射狀的、有的是細長帶狀的；有的瀑布旁是一些大石塊、有些則是以樹林為其背景。所以在這個類別中利用取物件的方式比較難以去取得比較一致的物件型態，所以 BCOM 與 RCOM 在這個類別中，所得到的整體效能是

差不多的。在先前我們曾經提到，影像在擷取的過程，通常會把影像的主題放置在影像的中央以突顯其重要性。在 MRCOM 有特別針對這個問題去做處理，加重位於影像中央物件所代表的語意概念，所以在這邊 MRCOM 可以得到比其它兩者更佳的效果。

圖 33、34 與 35 分別是以「瀑布」為查詢關鍵字，由 BCOM、RCOM 與 MRCOM 三種方法所執行出來的效果。因為瀑布常常與山上的大石頭或是樹林一起出現，所以有關於瀑布的影像中主要包含的有細長帶狀的瀑布物件、以綠色為主的樹林物件、與一些黑色大石頭的物件。如果利用 BCOM 的方法，常常會把原本應該屬於同一個體的物件分割到好幾個不同的區塊中，如圖 36 中瀑布的影像利用 Block Based 影像切割的方式，將很容易把整個瀑布分散到不同的物件當中（如圖 37）。所以用 Block Based 影像切割的方法很容易把整個畫面切成好幾塊，裡面包含著以黑色跟白色（石頭和瀑布）混合為主或是以黑色跟綠色（石頭和樹林）為主的 Blocks。如圖 33，BCOM 所找出來的影像大部分是以黑色和白色混合為主，或是以黑色和綠色為主的影像。



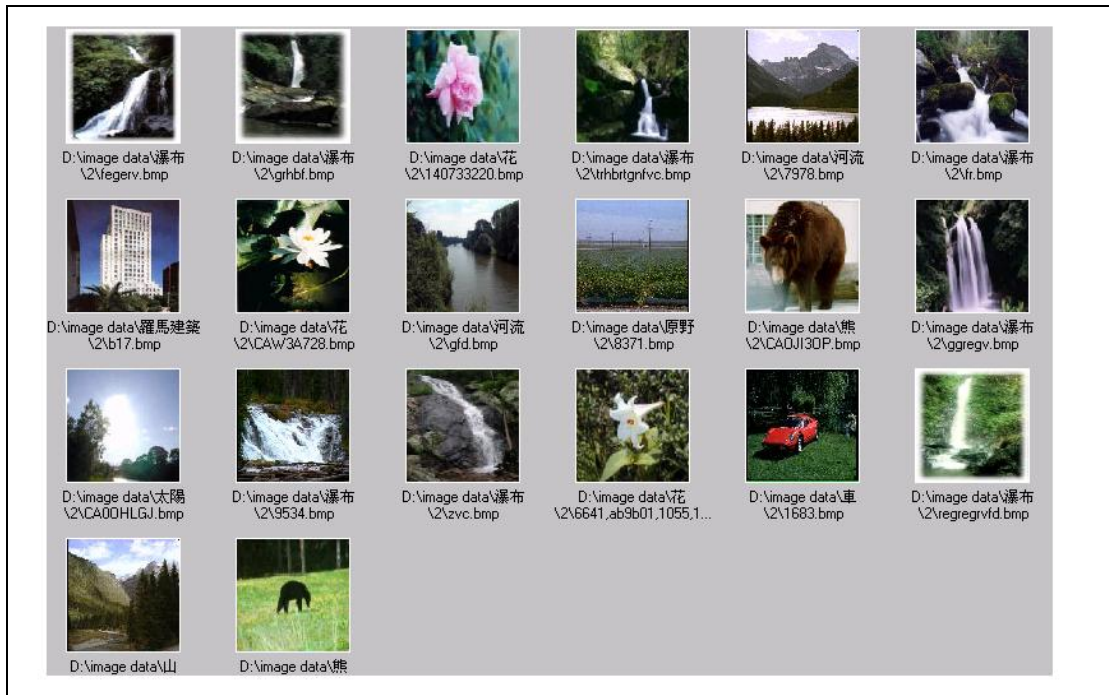


圖 33：以瀑布為查詢條件，BCOM 執行結果

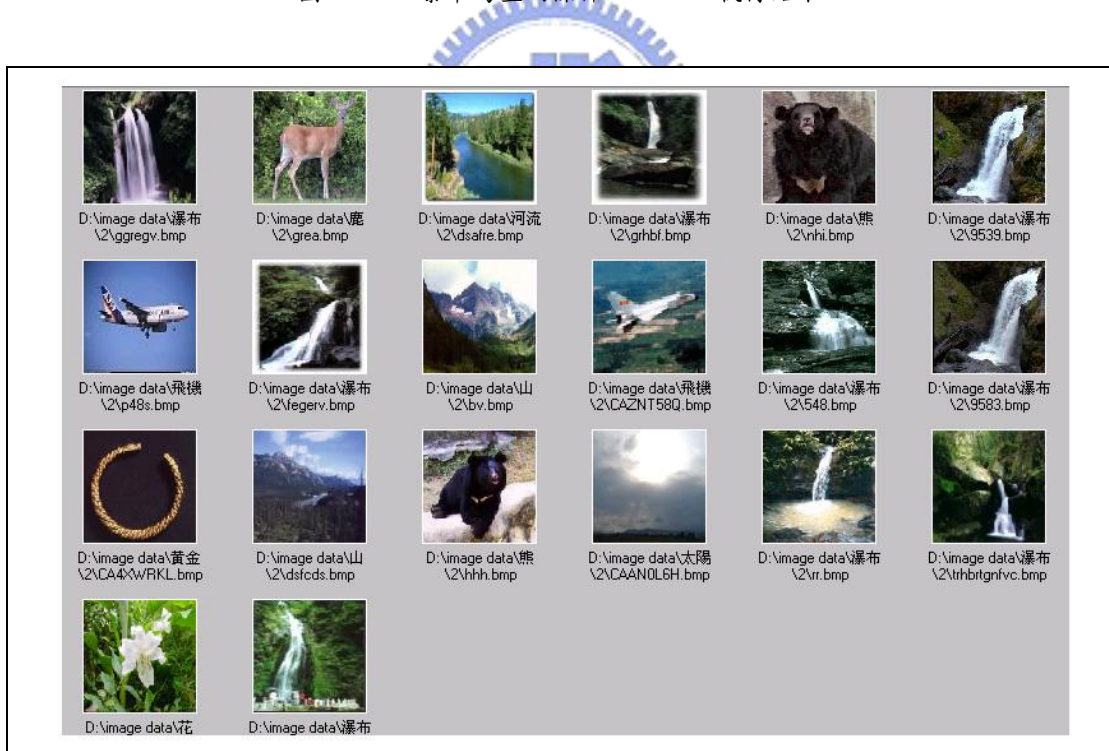


圖 34：以瀑布為查詢條件，RCOM 執行結果



圖 35：以瀑布為查詢條件，MRCOM 執行結果





圖 36：瀑布影像



圖 37：將圖 34 作 block 影像切割所得結果

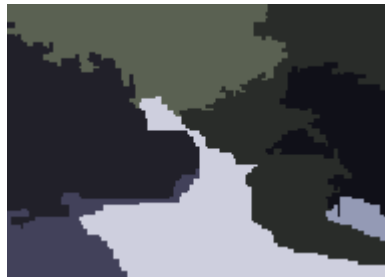


圖 38：將圖 34 作 region 影像切割所得結果

圖 38 是以 Region Based 影像切割將圖 36 作切割的結果。雖然利用這種方式可以將瀑布、樹林與石頭分別切割出來，但是由於在 Co-occurrence Model 中並沒有考慮到物件與影像位置對應關係的重要性，所以白色帶狀的物件，都很有可能被視為是瀑布。如圖 34 中，透過 RCOM 方法檢索出了很多內容含有白色帶狀物件的影像，但其內容並不是有關於瀑布的影像。因為在另一方面 MRCOM 考慮到位於影像中央的物件，比較能代表影像的主題而去加重其所代表語意權重，因此在圖 35 透過 MRCOM 的方法，找出了更多中央是以白色帶狀物件為主的影像，而其中大部分就是我們所要尋找的瀑布影像。

2. 交通類：車、飛機、船、火車、鐵路。

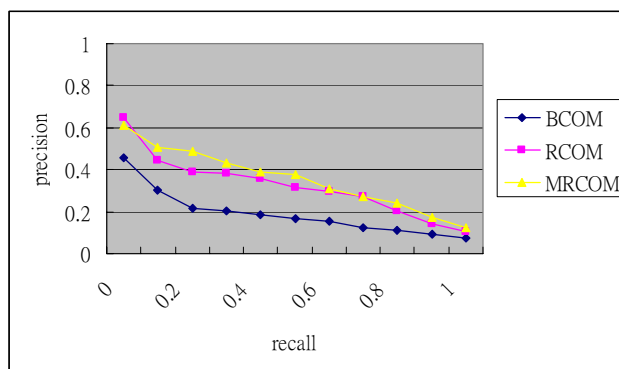


圖 39：交通類之 11-point Interpolated Measure Graph

MAP：

BCOM	RCOM	MRCOM
18.99%	32.19%	<b>35.56%</b>

表格 5：交通類之 MAP

在本實驗分類中，第二大類到第六大類是以特定物件為主題的影像，BCOM 常常會將特定的物件分割到不同的 Block 中；RCOM 與 MRCOM 的方法則可以將這些特定的物件比較完整地切割出來，所以在這幾個類別中 RCOM 與 MRCOM 表現的都比 BCOM 要來的好。

在 BCOM 與 RCOM 的做法中，是將所切割出來的物件對應到訓練資料集中最為接近的類別，但是這種做法有可能會把物件對應到外觀很相似但是在語意概念上錯誤的類別。MRCOM 是採取對應到多個相似類別的方式，以取得更為精確的語意概念，所以整體上 MRCOM 所得的效能要比 BCOM 與 RCOM 來的好。

3. 動物類：熊、老虎、馬、鹿。

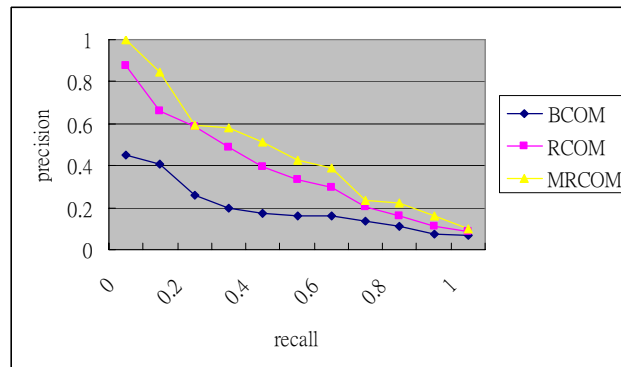


圖 40：動物類之 11-point Interpolated Measure Graph

MAP：

BCOM	RCOM	MRCOM
20.04%	38.09%	<b>46.01%</b>

表格 6：動物類之 MAP

4. 植物類：玫瑰、向日葵、百合。

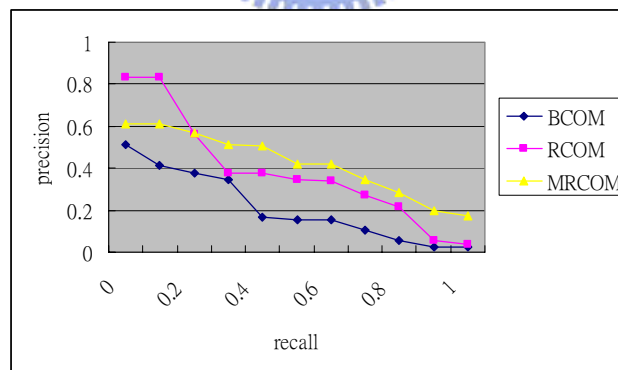


圖 41：植物類之 11-point Interpolated Measure Graph

MAP：

BCOM	RCOM	MRCOM
21.26%	38.60%	<b>42.24%</b>

表格 7：植物類之 MAP

5. 飾品類：金塊、項鍊、戒指。

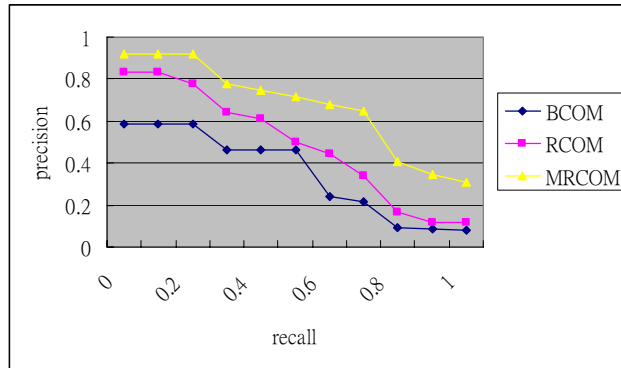


圖 42：飾品類之 11-point Interpolated Measure Graph

MAP：

BCOM	RCOM	MRCOM
51.75%	48.92%	<b>67.06%</b>

表格 8：飾品類之 MAP

6. 建築類：建築物、羅馬建築。

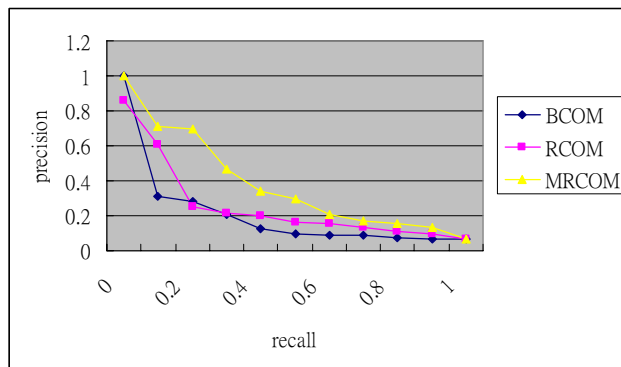


圖 43：建築類之 11-point Interpolated Measure Graph

MAP：

BCOM	RCOM	MRCOM
21.90%	25.96%	<b>38.50%</b>

表格 9：建築類之 MAP

圖 44 與表格 10 為計算在所有類別之整體 11-point Interpolated Measure 與 MAP。其中，在 11-point Interpolated Measure 中特定的查全率所對應的準確率的計算方式是計算所有相對的查全率之準確率平均，整體 MAP 是取各類別所得之 MAP 的平均。透過圖 44 與表格 10 可以很清楚的看到整體效能的表現最佳的為 MRCOM、RCOM 次之、BCOM 最差。

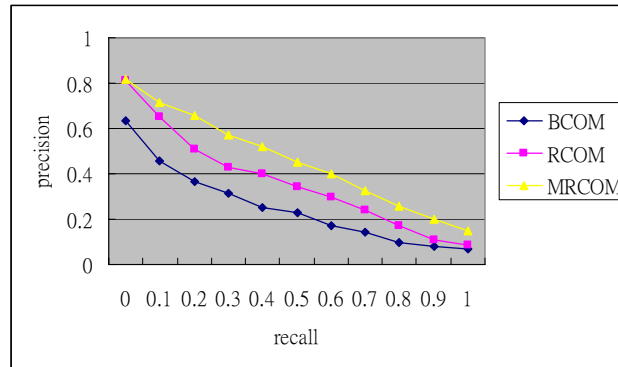


圖 44：六大類別整體之 11-point Interpolated Measure Graph

BCOM	RCOM	MRCOM
25.55%	36.84%	<b>46.00%</b>

表格 10：六大類別之整體 MAP

## 第五節 使用者回饋機制效能評估

在這個實驗中，將評估透過與使用者互動使系統進行學習，是否能提升系統效能。在每一次檢索的過程中，使用者可以指定系統回傳的結果與所要尋找的目標是否相關，並且將這些資訊回傳給系統。以下是 MRCOM 配合使用者相關回饋的機制檢索老虎的每一次查詢結果：

- 第一次檢索老虎：在回傳的前 20 張影像中 (如圖 45)，真正是老虎的影像有 11 張，MAP 為 55.35%。

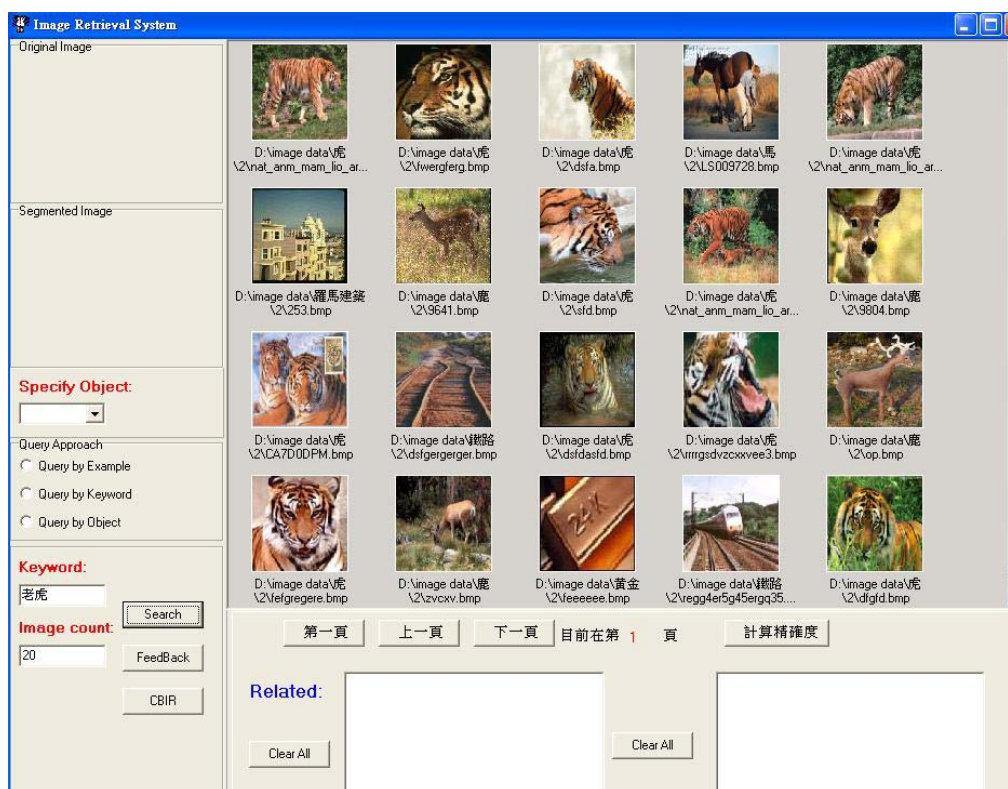


圖 45：老虎第一次檢索結果

- 第一次相關回饋後檢索老虎：在回傳的前 20 張影像中 (如圖 46)，真正是老虎的影像有 14 張，MAP 為 71.39%。比起前一次檢索進步了三張，MAP 進步了 16.04%。



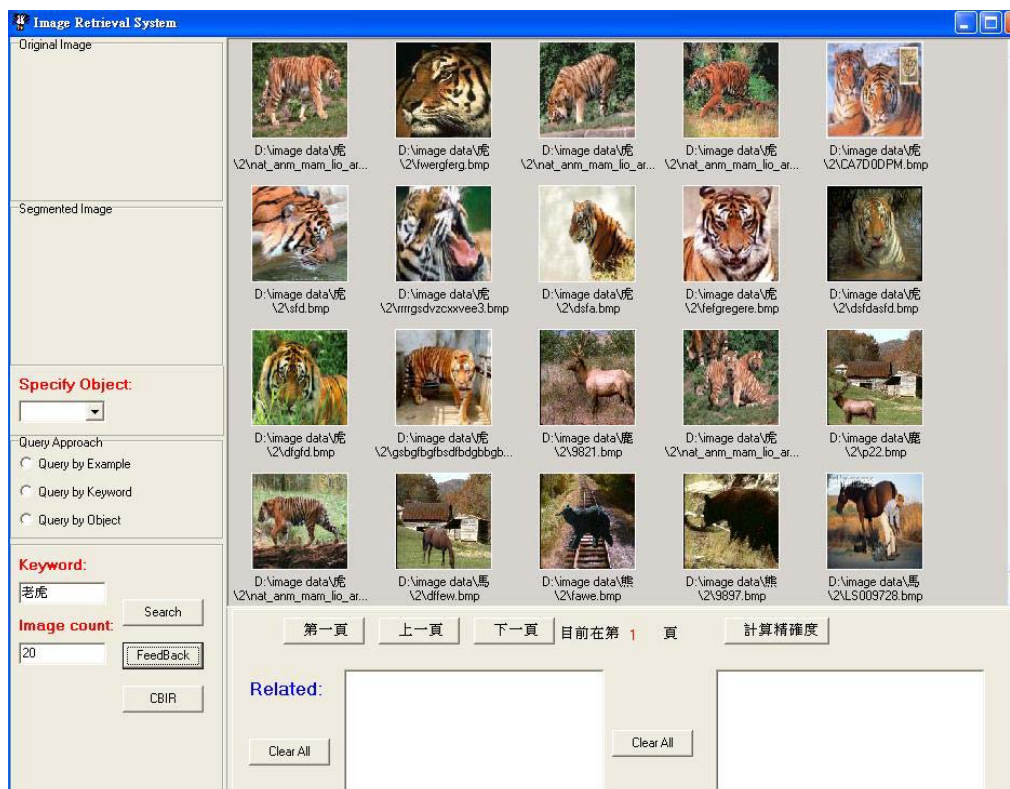


圖 46：老虎第一次相關回饋後檢索結果

- 第二次相關回饋後檢索老虎：在回傳的前 20 張影像中 (如圖 47)，真正是老虎的影像有 17 張，MAP 為 77.51%。比起前一次檢索進步了三張，MAP 較前一次進步了 6.12%。

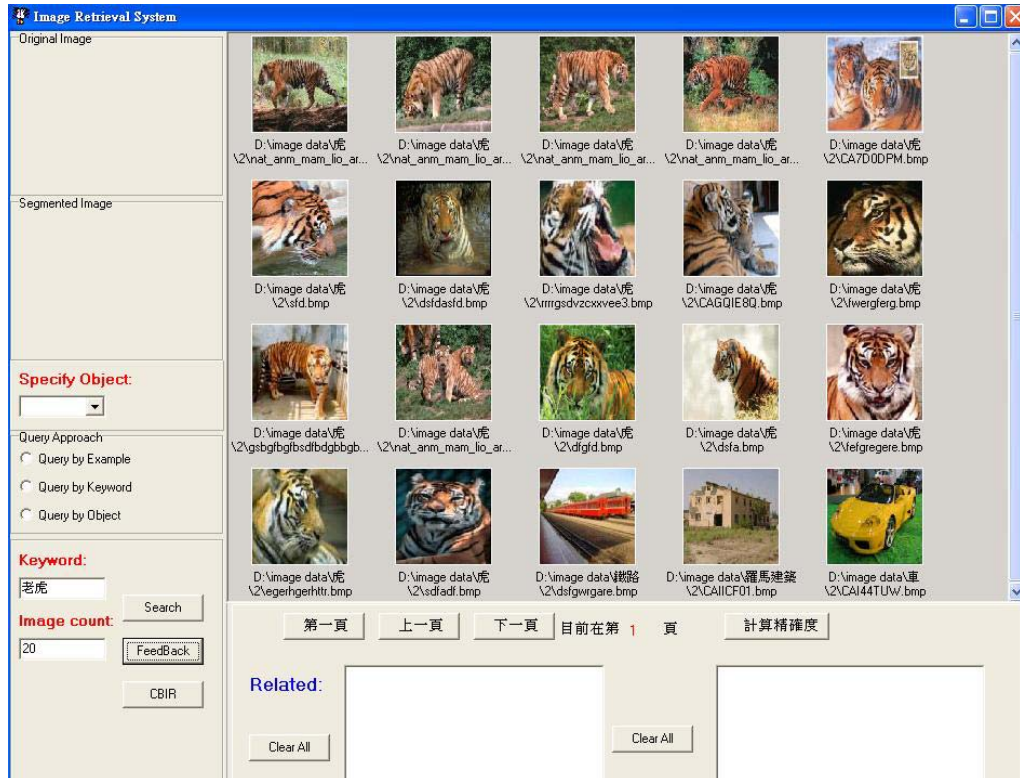


圖 47：老虎第二次相關回饋後檢索結果

第三次相關回饋後檢索老虎：在回傳的前 20 張影像中 (如圖 48)，真正是老虎的影像有 19 張，MAP 為 81.49%。比起前一次檢索進步了兩張，MAP 較前一次進步了 3.97%



圖 48：老虎第二次相關回饋後檢索結果

在以上幾次與使用者互動進行相關回饋的學習，可以很明顯看到系統效能的提升。在整個實驗中，根據在第一節中所提出的每一個類別具代表性的關鍵字進行檢索與學習，在使用者分別經過五次的相關回饋後，整體類別的 MAP 從原本的 46.00% 增加到 82.70%。圖 49 為在進行五次使用者相關回饋過程中的 11-point Interpolated Measure Graph，圖 50 為整體 MAP。由圖 49 與圖 50 中可以很清楚地看到，MRCOM 配合相關回饋的機制的確可以提升整體系統的效能，並且在經過 4 到 5 次使用者相關回饋的學習後，系統進步的幅度逐漸趨於緩和，且整體上已經達到一個蠻不錯的效能。

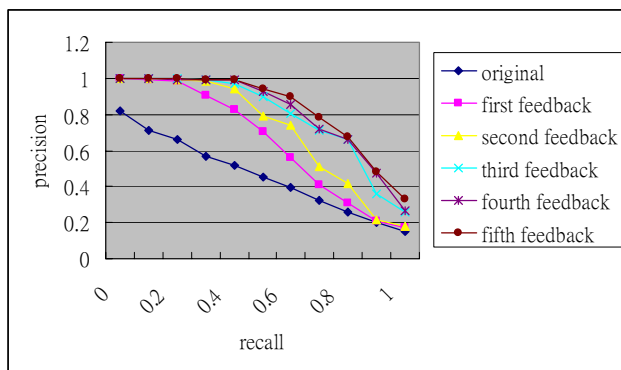


圖 49：六大類別相關回饋後之 11-point Interpolated Measure Graph

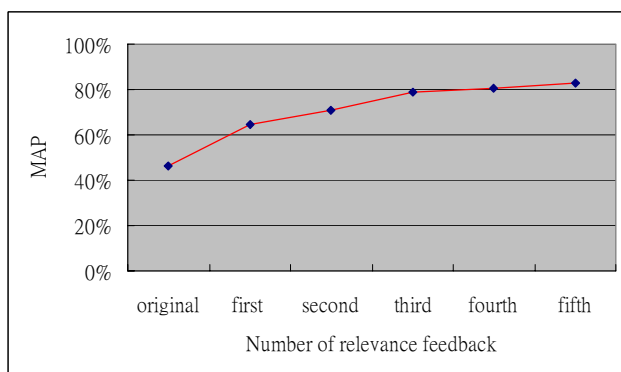


圖 50：六大類別相關回饋後之 MAP

## 第五章 結論與未來研究方向

本章總結本論文並說明未來的研究方向。第一節討論在本論文中所開發出的回饋式影像自動化標註與檢索系統的整體效能與發展過程的相關議題；第二節討論未來可能的相關研究方向。

### 第一節 結論

本論文提出了一種改良式共同出現模組的方法 (MRCOM)，主要針對 [Mori99] 作了以下改良：

1. 採用 Region Based 的模式，將影像切割成幾個跟人們視覺上比較吻合物件，再來進行 Region 的分群動作，把類似的 Regions 歸類成相同的 Blobs。
2. 在 [Mori99] 中採用將所取出之 Block 對映到最相似的群中，此種做法太過極端不適當，因為最接近的群可能不足以代表該 Block 的語意概念。所以本論文改採用尋找與 Region 最接近的前三群進行正規劃 (Normalize) 取得更適合 Region 的語意概念。
3. 一般存在影像中央的物件，在人們的感官中會將它認為是這一張影像的主題或是扮演比較重要的角色，所以當 Region 位於影像之中央時，將加強它占整張影像之語意概念權重。

從第四章的實驗中，可以看出利用 RCOM 方法比利用 BCOM 方法在六大類別中進行檢索，在 MAP 進步了 11.29%。本論文提出之 MRCOM 比 BCOM 與 RCOM 在效能上亦有顯著的提升，MAP 分別進步了 20.45% 與 9.16%。本論文所提出之使用者相關回饋機制在提升系統效能方面亦達到不錯的成果，整體上在經過五次的使用者回饋後，MAP 可以從原來的 46% 提升到 82.7%。

在第四章實驗中以 IBAM 與 MRCOM 進行效能的評估中，可以看到當影像的內容為風景類時，IBAM 執行效能優於 MRCOM。這主要是風景類影像中的物件本身變化程度比較大，所以透過 Region Based 影像切割比較難以取得其中的物件，且在同樣風格的風景類影像在整張影像的特徵都十分相近，所以 MRCOM 在此類影像中，效能比 IBAM 差。然而，在其它以特定物件為主題的影像，MRCOM 在效能上是優於 IBAM 的，這主要是由於同樣的物件可能會出現在不同的場景，如果以 IBAM 的方法進行影像標註將有可能因為背景的影響而忽略了影像中特定物件的重要性。但是 MRCOM 可以透過影像切割取得影像中的特定物件，因此可以避免上述發生在 IBAM 的情形。

## 第二節 未來研究方向



以下概述本論文未來研究方向：

1. 在進行分群的過程中，如何有效地判定先前分群完成的類別是否夠具代表性，進而將類別進行合併 (Merge) 或是分割 (Split) 的動作。若能將一些語意概念相近的類別進行合併則可以減少一些不必要存在的類別。若是在一個類別中，內部的語意概念資訊過於雜亂，則可將這一個類別分割成多個類別，使得類別中的語意資訊更具代表意義。
2. 在標註的過程中，所用來進行標註的關鍵字主要都是由訓練資料集中所得來。未來我們將研究如何幫助系統增加標註時所用的關鍵字，使標註內容更為豐富。
3. 許多資訊檢索系統導入查詢擴展 (Query Expansion) 的技術來提升系統使用效能。未來我們將研究如何透過有效的查詢擴展進而提升系統使

用效能。



## 參考文獻

- [Albuz95] E. Albuz, E. D. Kocalar, and A. A. Khokhar, "Quantized CIELab\* space and encoded spatial structure for scalable indexing of large color image archives," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 6, 2002, pp. 1995-1998.
- [Barnard03] K. Barnard, P. Duygulu, N. D. Freitas, D. Forsyth, D. Blei, and M. I. Jordan, "Matching Words and Pictures," *Journal of Machine Learning Research*, Vol. 3, 2003, pp. 1107-1135.
- [Barnard01] K. Barnard and D. Forsyth, "Learning the semantics of words and pictures," *In International Conference on Computer Vision*, Vol.2, 2001, pp. 408-415.
- [Bimbo97] A. D. Bimbo and P. Pala, "Visual Image Retrieval by elastic matching of user sketches," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 19, Issue 2, 1997, pp. 121-132.
- [Blei03] D. Blei, Michael, and M. I. Jordan "Modeling annotated data" *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, 2003, pp. 127-134.
- [Bilmes98] J. Bilmes, "A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models," *Technical Report, University of Berkeley, ICSI-TR-97-021*, 1998.
- [Bradshaw00] B. Bradshaw, "Semantic based image retrieval: A probabilistic approach," *ACM Multimedia*, 2001, pp. 167-176.



- [Carson99] C. Carson, M. Thomas, S. Belongie, J. M. Hellerstein, and J. Malik, "Blobworld: A system for region-based image indexing and retrieval," *In Third International Conference on Visual Information Systems, Lecture Notes in Computer Science*, 1999, pp. 509-516.
- [Cinque01] L. Cinque "Color-based image retrieval using spatial-chromatic histograms," *Image and Vision Computing*, Vol. 19, Issue 13, 2001, pp. 979-986.
- [Cho02] S. B. Cho, "A Human-Oriented Image Retrieval System Using Interactive Genetic Algorithm," *IEEE Transactions On System, And Cybernetics Part A: Systems and Humans*, Vol. 32, Issue 3, 2002, pp. 452-458.
- [Duygulu02] P. Duygulu, K. Barnard, N. D. Freitas, and D. Forsyth, "Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary," *In Seventh European Conference on Computer Vision*, 2002, pp. 97-112.
- [Ehrig04] M. Ehrig and Y. Sure, "Ontology Mapping – An Integrated Approach," *In Proceedings of the First European Semantic Web Symposium*, 2004.
- [Flickner95] M. Flickner, H. Sawhney, W. Niblack, and J. Ashley, "Query by Image and Video content: The QBIC System," *IEEE Computer*, Vol.28, Issue 9, 1995, pp. 23-32.
- [Forstner94] W. Forstner, "A framework for low level feature extraction," *Digital Photogrammetry and Computer Vision on Spatial Information*, 1994, pp. 383-394.
- [Freeman74] H. Freeman, "Computer processing of line drawing images," *ACM Computing Surveys (CSUR)*, vol. 6, Issue 1, 1974, pp. 57-97.

- [Funt95] B. V. Funt and G. D. Finlayson, "Color constant color indexing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, Issue. 5, 1995, pp. 522-529.
- [Garding96] J. Garding and T. Lindeberg, "Direct computation of shape cues using scale-adapted spatial derivative operators," *International Journal of Computer Vision*, vol. 17, Issue 2, 1996, pp. 163-191.
- [Gupta97] A. Gupta and R. Jain, "Visual Information Retrieval," *Communications of the ACM*, Vol. 40, Issue 5, 1997, pp.70-79.
- [Hopfield82] J. J. Hopfield, "Neural network and physical systems with collective computational abilities," *Proceedings of the National Academy of Science*, vol. 79, Issue 4, 1982, pp. 2554-2558.
- [Hu02] J. Hu and E. Hadjidemetriou, "Spatial color component matching of images," *16th International Conference on Pattern Recognition, 2002. Proceedings*, vol. 3, 2002, pp. 11-15.
- [Iqbal02] Q. Iqbal "Combining Structure Color and Texture For Image Retrieval: A Performance Evaluation," *16th International Conference on Pattern Recognition (ICPR)*, Vol. 2, 2002, pp. 438-443.
- [Iqbal00] Q. Iqbal and J. K. Aggarwal, "Low-level and Higher-level Approaches to Content-based Image Retrieval," *Proceeding of the IEEE Southwest Symposium on Image Analysis and Interpretation*, 2000, pp.197-201.

- [Jeon03] J. Jeon, V. Lavrenko, and R. Manmatha, "Automatic Image Annotation and Retrieval using Cross-Media Relevance Models," *ACM Conference on Research and Development in Information Retrieval*, 2003, pp. 119-126.
- [Jing02] F. Jing, M. J. Li, H. J. Zhang, and B. Zhang, "An efficient and effective region-based image retrieval framework," *IEEE Transactions on Image Processing*, Vol. 13, Issue 5, 2002, pp. 699-709.
- [Kelly95] P. M. Kelly, M. Cannon, and D. R. Hush, "Query by Image Example: The CANDID approach," *In SPIE Vol. 2420 Storage and Retrieval for Image and Video Databases III*, 1995, pp. 238-248
- [Ko02] B. C. Ko and H. Byun, "Multiple Regions and Their Spatial Relationship-Based Image Retrieval," *Image And Video Retrieval Lecture Notes In Computer Science 2383*, 2002, pp. 81-90.
- [Kohonen84] T. Kohonen, "Self-organization and associative memory," Springer Verlag, 1984.
- [Lee02] K. M. Lee, "Incremental feature weight learning and its application to a shape-based query system," *Pattern Recognition Letters*, Vol. 23, Issue 7, 2002, pp. 865-874.
- [Lim99] J. H. Lim, "Learnable visual keywords for image classification," *Proceedings of the fourth ACM conference on Digital libraries*, 1999, pp. 139-145.
- [Lu00] Y. Lu, C. Hu, X. Zhu, H. J. Zhang, and Q. Yang, "A Unified Framework for Semantics and Feature Based Relevance Feedback in Image Retrieval Systems," *ACM Multimedia*, 2000, pp. 31-37.

- [Malik90] J. Malik and P. Perona, "Preattentive texture discrimination with early vision mechanisms," *Journal of the Optical Society of America A*, vol. 7, Issue 5, 1990, pp. 923-932.
- [McQueen67] J. McQueen, "Some methods for classification and analysis of multivariate observations," *Proc. of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1967, pp. 281-297.
- [Mori99] Y. Mori, H. Takahashi, and R. Oka, "Image-to-word transformation based on dividing and vector quantizing images with words," *First International Workshop on Multimedia Intelligent Storage and Retrieval Management*, 1999.
- [Mostafa02] T. Mostafa, H. M. Abbas, and A. A. Wahdan, "On the use of hierarchical color moments for image indexing and retrieval," *IEEE International Conference on Systems, Man and Cybernetics*, vol. 7, 2002, pp.6.
- [Ogle95] V. Ogle and M. Stonebraker, "Chabot: Retrieval from a relational database of images," *Computer*, Vol. 28, Issue 9, 1995, pp. 40-48.
- [Paschos01] G. Paschos, "Perceptually Uniform Color Spaces for Color Texture Analysis: An Empirical Evaluation," *IEEE Transactions on Image Processing*, Vol. 10, Issue 6, 2001, pp. 932-937.
- [Pentland96] A. Pentland, R. Picard, and S. Sclaroff, "Photobook: Content-based Manipulation of Image Database," *International Journal of Computer Vision*, Vol. 18, Issue 3, 1996, pp. 233-254.
- [Pinheiro00] A. M. G. Pinheiro, E. Izquierdo, and M. Ghanhari, "Shape matching using a curvature based polygonal approximation in scale-space," *International Conference on Image Processing*, Vol. 2, 2000, pp. 538-541.

- [Prasad01] B. G. Prasad, S. K. Gupta, and K. K. Biswas, "Color and Shape Index for Region-Based Image Retrieval," *Lecture Notes in Computer Science (LNCS 2059)*, 2001, pp. 716-725.
- [Ravishankar98] K. C. Ravishankar, B. G. Prasad, S. K. Gupta, and K. K. Biswas, "Dominant color region based indexing for cbir," *International Conference on Image Analysis and Processing*, 1998, pp. 887-892.
- [Rui97] Y. Rui, T. S. Huang, and S. Mehrotra, "Content based image retrieval with relevance feedback in MARS," *International Conference on Image Processing*, vol. 2, 1997, pp. 815-818.
- [Rui99] Y. Rui and T. S. Huang "A Novel Relevance Feedback Technique in Image Retrieval," *ACM Multimedia*, 1999, pp. 67-70.
- [Smith95] J. R. Smith and S. F. Chang, "Single Color Extraction and Image Query," *Image Processing, 1995. Proceedings, International Conference, 1995*, pp. 528-531.
- [Smith96] J. R. Smith and S. F. Chang, "Visualseek: a fully automated content-based image query system," *In Proceedings of ACM Multimedia*, 1996, pp. 87-98.
- [Swain91] M. J. Swain and D. H. Ballard, "Color indexing," *International Journal of Computer Vision*, vol.7, Issue 1, 1991, pp. 11-32.
- [Tello95] R. Tello, "Fourier descriptors for computer graphics," *IEEE Transactions on Systems, Man and Cybernetics*, Vol. 25, Issue 5, 1995, pp. 861-865.
- [Wang01] J. Z. Wang and Y. Du, "Scalable Integrated Region-based Image Retrieval

- using IRM and Statistical Clustering,” *January 2001 Proceedings of the first ACM/IEEE-CS joint conference on Digital libraries*, 2001, vol. pp. 268-277.
- [Yoo02] H. W. Yoo, D. S. Jang, S. H. Jung, J. H. Park, and K. S. Song, “Visual information retrieval system via content-based approach,” *Pattern Recognition*, vol. 35, no Issue 3, 2002, pp. 749-769.
- [Zhang02] D. Zhang and G. Lu, “Enhanced generic Fourier descriptors for object-based image retrieval,” *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 4, 2002, pp. IV-3668 - IV-3671.
- [Zhang02] D. Zhang and G. Lu, “Improving retrieval performance of Zernike moment descriptor on affined shapes,” *IEEE International Conference on Multimedia and Expo*, vol.1, 2002, pp. 205-208.
- [Zhou01] X. S. Zhou and T. S. Huang, “Edge-Based Structural Feature for Content-Based Image Retrieval,” *Pattern Recognition Letters*, vol. 22, no. 5, 2001, pp. 457-468.
- [Zhou02] X. S. Zhou and T. S. Huang, “Unifying Keywords and Visual Contents in Image Retrieval,” *IEEE Multimedia*, Vol. 9, Issue 2, 2002, pp. 23-33.