物件式影像自動標註與檢索系統

鄭佳彬　鄭培成*　　柯皓仁　　楊維邦
國立交通大學資訊科學系　國立交通大學圖書館　國立東華大學資訊管理系 & 國立交通大學資訊科學系
300 新竹市大學路 1001 號
TEL：03-5712121 ext.56647
*E-mail: cpc@cis.nctu.edu.tw

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化影像標註方法，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，將影像切割成多個與人類視覺上較為吻合的物件；2.將所取出之物件對映到最接近的前三群，並進行正規化，以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。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%。

關鍵字：影像標註、關鍵字式檢索、共同出現模式
Keyword：image annotation, keyword-based image retrieval, co-occurrence model

1. 簡介

現行影像檢索的方法大致可分為範例式檢索(Query by Example)與關鍵字式檢索(Query by Keyword)兩大方向。當使用範例式影像檢索系統時，系統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[1]，讓使用者描繪所欲檢索的影像輪廓，接著系統根據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，就很難去檢索使用者想要找尋的影像。

在另一方面，利用關鍵字式影像檢索系統則有兩個主要的問題：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。

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，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的檢索方式(Semantic Query)。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)來協助建立影像註解。本論文利用影像處理(Image Processing)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。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(Semantic Concept)，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。

2. 相關研究工作

本論文的相關研究工作主要分為兩大類，一類為自動化影像標註的相關方法，這一類的研究主要是著重於如何透過影像低階特徵進行影像語意概念的推導；另一類是影像標註的表示法，探討如何建構影像中所包含的語意概念之表示方法。

2.1 自動化影像標註的相關方法

影像的內容是由存在於影像之中的物件所組成的，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，一個是區塊式(Block Based)，一個是區域式(Region Based)。其中，區塊式在實作上較區域式更為簡單，因為區塊式只是單純地將影像切割成數個等大小的矩形，進一步由一個或多個矩形組成物件。但是區塊式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，利用區塊式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。區域式影像切割將影像切割成比較吻合人們所認定的物件。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？所以區域式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[3][4]。

在取出影像中的物件後，將透過這些物件的低階特徵(如顏色、形狀、紋理等…)推論可能存在的語意概念並為影像進行標註。在[7]中提出了共同出現模式(Co-occurrence Model)計算出現在相同類別(或稱 “群(Cluster)”)中的物件與關鍵字的頻率，進行機率的統計。在[3]中提出了一種翻譯模組的做法，將翻譯模組視為一種辭典，利用機器翻譯的將影像中的物件翻譯到另一種語言。在[4]中指出，影像中的物件與文字之間並非是一對一的對映關係，因為有一些字可能是多個物件共同出現時，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。

2.2 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[4]、語意網路模式[6]、機率式標註向量模式[4]，以下是這三種模組的大致介紹。

2.2.1 固定式標註模式

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，對它的的標註方式為（*I* | 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。

2.2.2 語意網路模式

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。在影像 *I* 中，若包含與關鍵字 *w* 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* 之間的關聯程度，權重的值越高表示二者的關連性越大。

2.2.3 機率式註解向量模式

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。

3. 物件式影像標註自動標註與檢索系統

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)，主要是針對以下三點改進共同出現模式：(1)以區域式影像切割代替[7]中以區塊式影像切割取得物件；(2)一個物件將對映到多個群(Cluster)以取得更精確的語意概念；(3)加重位

於影像中央之物件的重要性。

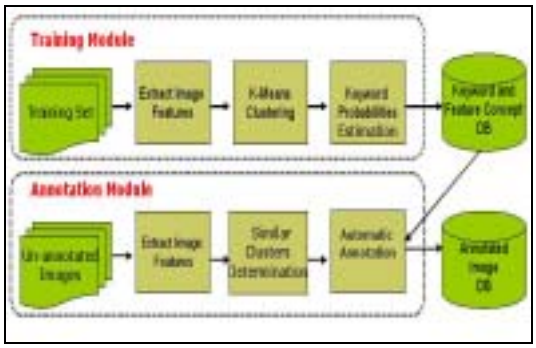

圖 1 系統架構

3.1 系統架構

本系統架構如圖 1 所示，共分為兩個模組，一為訓練模組(Training Module)；另一為自動標註模組(Annotation Module)。首先將訓練資料集的影像透過訓練模組找出同樣風格物件的語意概念，接著再將未標註影像透過自動標註模組計算出影像中的語意概念進行標註的動作。

3.2 影像特徵的擷取

本論文採用顏色與形狀當作影像的特徵。其中，在顏色的部分，會先將影像中的顏色透過[7]中的顏色對照表將顏色量化到 25 種顏色，接著透過[2]所提出之顏色特徵進行相似度計算。在形狀部分則取用[9][5]中所提之形狀特徵進行相似度計算。影像中兩物件之相似度計算公式如(Eq. 1)所示，其中，$sim\_c_{i,j}$ 與 $sim\_s_{i,j}$ 分別代表兩個物件在顏色與形狀的相似度，w_c 與 w_s 為顏色與形狀所佔權重。

$$sim_{i,j} = w_c \times sim\_c_{i,j} + w_s \times sim\_s_{i,j} \quad \text{(Eq.1)}$$

3.3 訓練模組

本論文以區域式切割取代[7]中以區塊式切割取得物件，訓練過程如下：

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進行區域式影像切割
3. 由訓練資料集內某影像 *I* 所切割出來之所有物件將繼承 *I* 的所有標註文字作為其標註
4. 取出影像中的物件特徵
5. 以 K-Means Clustering 將所有物件進行分群
6. 計算每一群中每個關鍵字出現的機率

步驟六中，計算關鍵字 w_N 出現在第 K 群的機率 $P_{K,N}$ 公式如((Eq. 2)：

$$P_{K,N} \approx \frac{|w_N|}{\sum_{i=1}^{M} |w_i|} \quad \text{(Eq. 2)}$$

其中，$|w_i|$ 表示關鍵字 w_i 出現在某一群中的次數；M 表示所有關鍵字個數。

3.4 自動標註模組

在傳統共同出現模式中，將所取出物件對應到訓練模組中最相似的群，這種做法太過極端且不適當，所以本論文改採找尋與物件最接近的前三群進行正規化。關鍵字 w_N 出現在區域 r 中的機率 $P_{r,N}$ 可利用(Eq. 3)得之：

$$P_{r,N} = \frac{\sum_{j=1}^{3} s_j^r \times P_{C_j,N}^r}{\sum_{i=1}^{3} s_i^r} \quad \text{(Eq.3)}$$

其中，s_j^r 表示與 r 第 j 相似群之對映相似度；$P_{C_j,N}^r$ 表示關鍵字 w_N 出現在與 r 第 j 相似群的機率。

此外，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區域 r 是位於影像 I 之中央，則加重其語意概念在整個影像中的權重。本論文中定義中央區域的方法如下：

1. 將影像切割為九個等大小之矩形(如圖 2)
2. 計算所有區域中的點位於第 5 個矩形範圍內占整個區域的比例
3. 在步驟 2 中比例最大的區域即位於中央之區域

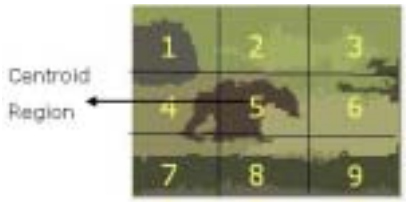

圖 2　定義中央區域

最後，計算關鍵字 w 出現在影像 I 中的機率 $P(I_w)$ 可由(Eq. 4)：

$$P(I_w)=\frac{1}{N}\sum_{j=1}^{N}\alpha_j \times P_{rj,w} \quad \text{(Eq. 4)}$$

其中，N 為在影像 I 的區域個數；α_j 為影像 I 第 j 個區域所佔之權重，$P_{rj,w}$ 為關鍵字 w 在第 j 個區域可能出現之機率值。

4. 實驗與討論

本節簡介本論文系統實作並針對系統進行效能評估。

4.1 系統實作

為了解決前述使用者查詢影像所可能遭遇到的問題，我們開發了一套複合式影像檢索系統，系統介面如圖 3，系統共提供了以下三種檢索方式：

圖 3　系統操作介面

(1) 範例式查詢：實作[2]中所提出之影像特徵進行檢索。
(2) 物件式查詢：透過系統所提供之影像切割工具讓使用者選擇所感興趣之物件進行檢索(如圖 4)。
(3) 關鍵字式檢索：透過 MRCOM 為影像進行語意概念分析，並允許使用者透過關鍵字進行檢索(如圖 5)。

圖 4　以物件式查詢「瀑布」

圖 5　以關鍵字查詢「車」

4.2 系統效能評估

本論文主要是透過區域式影像切割並改良傳統共同出現模式對影像進行標註的動作。為了比較區塊式影像切割與區域式影像切割在套用共同出現模組時的效能，除了 MRCOM 外，我們另外實作了兩套系統進行三種方法之效能評估，分別為：

BCOM (Block Based Co-occurrence Model)：這個做法主要是利用區塊式影像切割取得影像中的物件，並且套用傳統共同出現模式為影像進行標註。

RCOM (Region Based Co-occurrence Model)：這個做法主要是利用區域式影像切割取得影像中的物件，並且套用傳統共同出現模式為影像進行標註。

在實驗中，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。在訓練過程中，針對訓練資料集中的影像分別給予 1 到 7 個關鍵字進行影像標註。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。我們將實驗資料分為六大類 (如表格 1)，並且針對每大類以 2 到 5 個代表性關鍵字進行檢索以評估效能。

類別	代表性關鍵字
風景	山、太陽、河流、瀑布、原野
交通	汽車、飛機、船、鐵路、火車
動物	熊、老虎、馬、鹿
植物	玫瑰、向日葵、百合
飾品	金塊、項鍊、戒指
建築物	建築物、羅馬建築

表格 1　實驗資料分類與代表性關鍵字

在效能評估方式方面，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)。

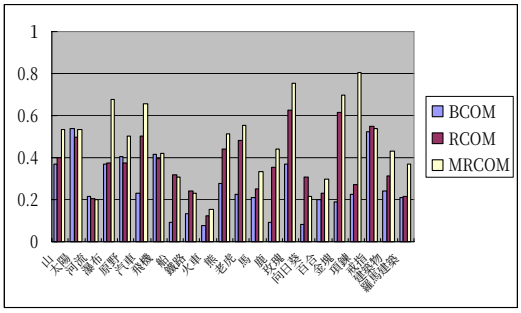

圖 6　各代表性關鍵字之 MAP

	BCOM	RCOM	MRCOM
風景	**35.93%**	**37.31%**	**46.65%**
交通	**18.99%**	**32.19%**	**35.56%**
動物	**20.04%**	**38.09%**	**46.01%**
植物	**21.26%**	**38.60%**	**42.24%**
飾品	**51.75%**	**48.92%**	**67.06%**
建築物	**21.90%**	**25.96%**	**38.50%**

表格 2　六大類別在三種方法下之 MAP

BCOM	RCOM	MRCOM
25.55%	**36.84%**	**46.00%**

表格 3　整體 MAP 之比較

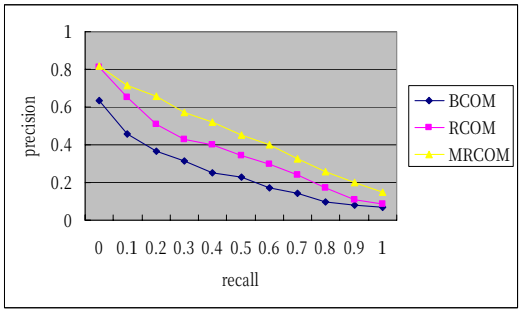

圖 7　三種方法之 11-point Interpolated Measure Graph

圖 6 為所有測試關鍵字分別以前述三種方法所得之 MAP，表格 2 為六大類別測試關鍵字之整體 MAP，表格 3 為整體效能之 MAP，圖 7 為整體效能之 11-point Interpolated Measure Graph。由圖 6 中看出 MRCOM 普遍在各個測試的關鍵字中得到較佳的效能。並且由以上的實驗數據很明顯看到 MRCOM 的整體效能最佳，RCOM 次之，BCOM 最差。透過區域式影像切割所得到的影像物件會比以區塊式影像切割得到之物件更接近人們視覺上所認同的，所以 RCOM 所執行的效能比 BCOM 要來的好。由於 MRCOM 是以區域式影像切割的方式取得影像中的物件，透過前述的改良，最後可以在實驗數據中看到本論文所提出的改良方式確實提升了系統的效能！

5. 結論與未來研究方向

本論文提出了一種改良式共同出現模組的方法-MRCOM，主要透過區域式影像切割並且對傳統共同出現模式進行前述的改良，從第 4 節的實驗中可以看到 MRCOM 在整體的效能上分別比 BCOM 與 RCOM 進步了 20.45%與 9.16%。

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

注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。

6. 參考文獻

[1] Bimbo, A. D., Pala, P. Visual Image Retrieval by Elastic Matching of User Sketches. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19(2), 1997, 121-132.

[2] Cinque, L., Levialdi, S., Olsen, K.A., Pellicanò. Color-based image retrieval using spatial-chromatic histograms. *Image and Vision Computing*, 19(13), 2001, 979-986.

[3] Duygulu, P., Barnard, K., Freitas, N. D., and Forsyth, D, Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. *In Seventh European Conference on Computer Vision*, 2002, 97-112

[4] Jeon, J., Lavrenko, V., and Manmatha, R. Automatic Image Annotation and Retrieval using Cross-Media Relevance Models. *ACM Conference on Research and Development in Information Retrieval*, 2003, 119-126.

[5] Ko, B. C., and Byun, H. “Multiple Regions and Their Spatial Relationship-Based Image Retrieval,” *Image And Video Retrieval Lecture Notes In Computer Science 2383*, 2002, 81-90.

[6] Lu, Y., Hu, C., Zhu, X., Zhang, H. J., and Yang, Q. “A Unified Framework for Semantics and Feature Based Relevance Feedback in Image Retrieval Systems,” *ACM Multimedia*, 2000, 31-37.

[7] Mori, Y., Takahashi, H., and Oka, R. “Image-to-word transformation based on dividing and vector quantizing images with words,” *First International Workshop on Multimedia Intelligent Storage and Retrieval Management*, 1999.

[8] Ravishankar, K. C. , Prasad, B. G., Gupta, S. K., and Biswas, K. K. “Dominant color region based indexing for cbir,” *International Conference on Image Analysis and Processing*, 1998, 887-892

[9] Zhang, D., and Lu, G. “Improving retrieval performance of Zernike moment descriptor on affined shapes,” *IEEE International Conference on Multimedia and Expo*, vol.1, 2002, 205-208.