行政院國家科學委員會專題研究計畫　成果報告

類神經網路影像檢索之研究(II)：影像資料索引之建構

計畫類別：個別型計畫
計畫編號：NSC94-2213-E-009-139-
執行期間：94 年 08 月 01 日至 95 年 07 月 31 日
執行單位：國立交通大學資訊工程學系(所)

計畫主持人：傅心家
共同主持人：包曉天

報告類型：精簡報告
報告附件：出席國際會議研究心得報告及發表論文
處理方式：本計畫可公開查詢

中　華　民　國 95 年 10 月 27 日

中文摘要：

多媒體如數位影像等的大量增加使得以人工對影像標記關鍵字來建立索引和進行搜尋的方式相當耗費人力與時間，因此，如何自動化處理多媒體的索引建立以及搜尋就顯得日益重要。然而，日常的語意所定義出的物件，卻難以在影像上精確地劃分，而所謂相關/相似影像，更常莫衷一是。故而有研究指出可以使用低階影像特徵作為檢索依據，只要低階影像特徵輸入正確，便可快速精確地找到所求影像。但是以低階特徵為檢索條件的方式，在介面使用上對一般使用者往往不夠直覺且難以使用。所以影像檢索最困難的部分就是在於如何將使用者心目中的概念影像(conceptual image)，以具體的簡圖或是以數學算式，或影像特徵等方式表現出來。

使用者心中的概念影像千變萬化，很難已對所有的物件都給定相同的特徵維度，過高有維度詛咒(Curse of Dimensionality)問題，過低則又無法充分表現使用者的概念。我們認為應對於概念影像複雜度而動態給予不同的特徵維度表示。然而在維度不同的兩物件如何比對其相似度，以及如何在使用者端以及伺服端上達成影像檢索的四個需求--(1)精確(2)快速(3)使用者親和(4)可擴張性，並能以自動化處理是本計劃的重心。在前期計劃中我們發展了一套方法，除可作為高階語意和低階影像特徵之間的橋樑。在本期計劃中更進一步發展出可精確地描述使用者的概念影像的方法，簡稱視覺字串(Visual string)，可以根據使用者點選圖片進而動態調整描述特徵向量的複雜度。由於每張影像所描述的複雜度不盡相同，所以本計畫發展了Generalized Probabilistic Decision-based Neural Network 來計算不同複雜度的影像特徵向量之間的相似程度，進而能比較影像資料庫中預先建立的索引與根據使用者點選影像所得到的視覺字串之間的相似度，而給出使用者想要的圖片。

關鍵詞：

Neural Network, Content-based Image retrieval (CBIR), Mixture Gaussian Model (GMM), Generalized ProbabilityDecision Based Neural Networks (GPDBNN)

Abstract：

The region-based image retrieval becomes a popular topic in the recent years for its capability to index an image with different levels of local contents in the image. In this project, we have developed a “visual string” to determine dynamically the number of feature dimensions according to the complexity of the "conceptual image".

The goal of this project intends to construct an indexing/query methodology that can bridge the image representation gap between the high-level semantic meaning and the low-level image features. However, the conventional methodologies about comparison similarity between two objects are based on that they are in the same feature dimension. If two objects are in different feature dimension, the conventional methodologies do not work. We have developed “Generalized Probabilistic Decision-based Neural Network(GPDNN)” to compare two objects with different feature dimension. By “visual string” and “GPDNN”, the system can build reliable indexing image database automatically and also can return good query answers.

前言：

近年來多媒體資料盛行，影像、聲音等資料大量出現在網路上，像 Youtube[1]、Video.google[2]、Flickr[3]這樣可以免費提供空間給使用者的影像平台已經相當常見，甚至也有使用者開始利用網路的空間以聲音做podcast[4]，資料的展現已經不再僅是文字，而是混合大量影像、聲音等多媒體資料。

而數位相機隨拍隨看以及輕便的特性，以及在手機上越來越普及的相機功能，數位相機甚至會取代部分傳統攝錄影機市場[5]。而近年來，手機製造商也常在手機上附加攝影功能的小型 CCD，甚至已經有可能取代部分低階數位相機市場[6]加上電子產業進步，電子儲存設備的價格低廉，數位影像的儲存代價相對變得低廉，僅個人使用就有可能儲存大量的數位影像，更遑論散佈在網際網路各處的數位影像的龐大資料量。過去，為了提昇搜尋引擎的準確性，常採用人工標記的方式，如 yahoo 等做法[7]。然而，在上述因素刺激下而大量成長的數位影像，若採行人工標記的方式，所耗費人力、財力之大將無可計數。因此，針對影像搜尋上，一個可以自動化建置影像資料，並提供搜尋的檢索方式的需求，不但是一般含有影像的網站需要，一般個人的需求也因為數位相機產品普及化而大量增加。

目前網路影像檢索多採用內容關鍵字搜尋方式[4,7]供使用者檢索所需影像，然而隨著影像資料的巨幅成長，若不能在影像內容上以自動化方式適當提供精確的描述，而仍要以人工方式來註解(Annotation)，那麼想要在網際網路的多媒體資料上，建構便於搜尋所需之索引或關鍵字，將會是一項耗費龐大人力、時間的工作。此外，建構一個可提供搜尋功能的影像資料索引所需考量的問題，事實上比建構文字索引更加複雜，主因在於：

雖然可以用範例影像本身的內容當作搜尋的指標，但是使用者在心中的目標影像(其後我們稱之為目標概念影像)並不出現在於單張影像中。

一張影像所表達的概念對於不同的人而言存在不同的意義，為了加速檢索作業在每一張影像先做具有概念意義的分類，難以達到絕對客觀的程度。

而通常又為了加速影像搜尋的速度，對於已儲存之影像常預先處理成較適合檢索的資料型態，例如將影像先切割成許多物件--如貓、狗、房屋等，然後再利用切割出影像中的物件來進行搜尋。然而，依語意所定義出的物件，在影像上難以相當精確地切割出來，加上所謂相關/相似影像，更是一種模糊概念，常常是見仁見智，缺乏放諸四海皆準的標的。這類困難使得影像檢索的需要難以被滿足[8][9][10]。因此有學者研究針對影像低階特徵進行檢索的方式[10-24]。儘管以低階影像特徵作搜尋，只要輸入的條件正確，就可以快速找到目標影像。但是，針對低階特徵所提供的檢索介面，對於一般使用者不夠直覺而難以使用[25]。

但是透過以人工方式註明每一張影像中所有包含的概念，藉此提昇對影像檢索的廣度及精確度。但是隨著影像資料的巨幅成長，人工標記方式已然不可實行。因此影像檢索的自動化建立是現今多媒體資料應用技術上的必然趨勢。

研究目的：

對於多媒體檢索而言，最大的困難始終存在於如何把使用者心中想要的概念，以數學方法，或是以電腦可以理解的方式表現出來。一個有效且直覺的多媒體的檢索應包含：直覺的檢索介面，檢索核心，資料自動化處理。直覺的檢索介面在於讓使用者可以很直覺地使用檢索系統，不需要額外的專門知識便可以使用。為了讓使用者不需要和系統互動多次才能找到所需要的資料，不但在直覺的檢索介面上可以確切獲得使用者心中的概念外，更需要一個有效的檢索核心。而為了加速多媒體上的搜尋，如果能對已經儲存在資料庫內的資料預先做處理，在搜尋時候就可以增加速度。由於多媒體的大量增加，使得我們幾乎不可能負擔人工作業方式所耗費的大量人力和時間。因此資料自動化處理除了可以在搜尋時候降低處理時間以外，更是勢在必行的趨勢。

在做影像檢索時，使用者常會於心中將影像切割成許多物件--如貓、狗、房屋等，並期望檢索系統能於影像資料庫中找尋出含有某特定物件的影像。為了能夠符合使用者的期望，在做影像檢索前，影像資料庫中的影像必須事先根據其包含的物件做索引。若用人工的方式對儲存於影像資料庫中的所有影像做索引，不僅費時費力，而且由於人的主觀意識，使得不同的人或不同的時間對相同的影像常常做出不同的索引。

為了能自動化的對影像做索引，本年度提出了 visual keyword 與 visual string 的概念來作影像的索引。所提之 visual keyword 是以複合式高斯機率分布函式(mixture Gaussian probability distribution function)來描述影像中某區塊的顏色、紋理與形狀等影像特徵。與目前文字搜尋相仿，使用者用 visual keyword 來搜尋影像，系統會找尋位於資料庫內包含所要搜尋的 visual keyword 的影像。visual keyword 整合了影像低階特徵，並以同於文字搜尋的 keyword 觀念來為影像作索引。由於使用多高斯分佈模型所描述影像的特徵，對於影像上的特徵變化有其容忍度。相對於整張影像的檢索方式，用多高斯分佈模型的方式，只要區域影像特徵上的分佈相近就會被篩選出來，對於在檢索區域以外的影像部分變化則不予考慮，減少非使用者感興趣的影像被檢索出來，也降低非檢索區域對檢索結果的影響。圖表 1 為用 visual keyword 來作影像索引的例子。經過自動化處理後，描述圖(a)中白馬的 visual keyword 索引如圖(b)所示。

(a) The original image

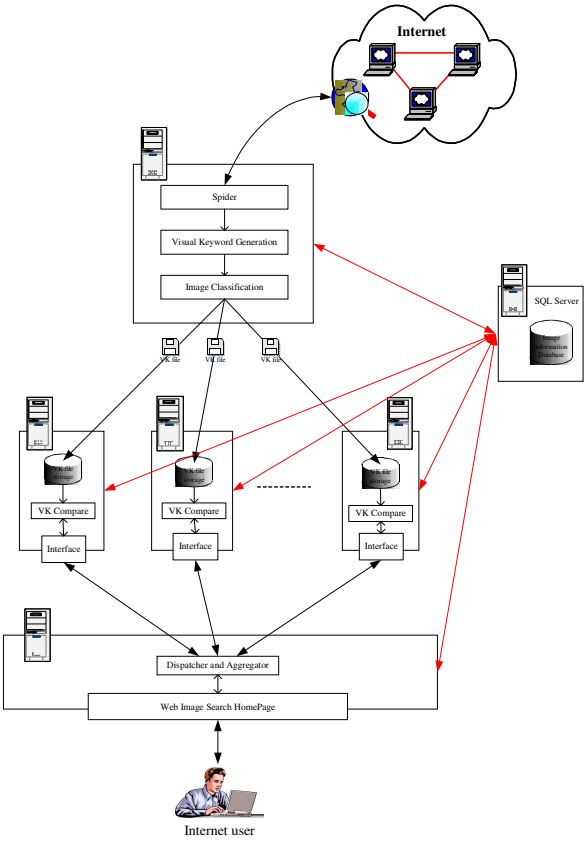
(b) the visual keyword to describe the white horse

圖表 1 An example of visual keyword indexing

而多個 visual keywords，根據之間相互的位置關係就可以形成 visual string。所以使用者不但可以找一匹白馬，藉由 visual string 的產生，更可以找一匹在草地上的白馬。

隨著影像資料庫的增大，如何有效且快速地在龐大數量的影像索引中做搜尋已是急

欲解決的課題。本計畫已經建構了一多伺服器之實驗平台。下圖為此實驗平台之架構圖。此實驗平台包含了 5 部 image query server、1 部 web server 及 1 部 SQL server。初步的實驗證實此架構確能加速搜尋及處理速度。此實驗平台的網址為 http://140.113.216.66/WebImageSearch。



文獻探討：

在影像檢索領域上，國內外均已有許多研究發表[11-24]。早期發展的系統，在國外如 QBIC[16]，Virage [17]，Photobook [18]，VisualSEEk 和 WebSEEk [19] 皆是以整張影像的特徵分布來當作搜尋的依據。然而以整張影像的特徵分布當作搜尋依據並不能有效掌握到局部特徵的分布。就一張影像而言，對其感興趣的區域，會因為每個人不同而有所不同，以這類方式做搜尋常需要多次尋找，或是與系統互動才能得到比較好的搜尋品質，也就是找到符合使用者心中想要的影像。在針對影像搜尋難以一次到位的困難，便有學者進行和使用者互動的系統，如 Netra[20] 和 Blobword [21]等，期望藉由使用者先規劃出感興趣物件的外型或其他特徵，進而掌握到局部區域特徵之分布。一般說來，這類檢索方式可以得到比整張影像進行搜尋的方法好，然而，以使用者預先規劃出感興趣物件之外型或特徵的方法，其檢索結果相當倚靠使用者所輸入的特徵，而對一般使用者來說，低階影像特徵通常不夠直覺，而使用者對於物件形狀的描述也往往不夠精確，使得這類方式不易得到良好之結果。

為了克服使用者在對物件描述上不夠精確，便有學者進行以電腦自動分割影像物件，藉由電腦精確地切割影像物件進而避免人為上的誤差。但是，影像上的物件精確自動化切割卻是不易達成的目標[8][9][10]，所以有 Y. Chen 和 J.Z. Wang [28]等學者，針對這類問題，提出以影像部分區域為基礎的乏晰特徵比對方法，根據顏色、材質等影像特徵，對整張影像分割成數個特徵相近的區域。由於影像區域的形成是根據乏晰的影像特徵，藉由乏晰的特質，可在比對上避免因為切割不精確所帶來的影響。近年來不少學者提出用高階語意(Semantics)的特徵[34]，來檢索靜態影像甚至動態影像內容，在檢索的精準度要比使用低階特徵好很多，但是尚未能看到自動化地將高階語意自影音資料中成功地抽取出來。這方面也是不少研究人員目前研究重點之一。

而國內學者，如交大資工系李素瑛教授則針對以區域影像特徵做檢索時，研究各區域間位置關係的有效表示法，即為二維字串比較法[23]。在二維字串比較法當中，每一塊區域都會標記該區域在影像的水平方向和垂直方向上，所佔的起始位置。根據這些標記，

就可以定義出區域之間位置的關係，以及它們在不同方向上是否有相疊的區塊。此表示法可以避免因為影像旋轉後，在語意上不同的困擾，而使各區域的相對關係仍然可以良好保持。

本實驗室也在六年前著手相關研究[26]，提出" EM based multiple infernecelearning " 概念來做影像檢索系統之開發與研究，也取得初步成果[29][30]。四年前更擴展到新聞影片上的研究[27]，累積許多多媒體及影像特徵擷取經驗，冀望透過上述經驗以及本實驗室多年在類神經網路的研究，可提供一直覺且有效的影像檢索系統。

研究方法：

研究的方法主要建立在三個方面: 影像特徵的抽取，特徵混合的展現，諸特徵關係的建立。

在影像特徵的抽取上，我們已經做了繁複的實驗得到可以良好表現圖像特徵，並適用於自動化影像檢索的特徵。我們利用了人類視覺會先察覺顏色，其次是形狀，最後才是動作[31]的性質，來當作影像的低階特徵。但若單純憑藉顏色會切出過多零碎的形狀，不足以構成一個整體，因此就必須考慮重複出現的顏色組合，也就是材質(Texture)。根據相同材質特性，雖可以決定形狀，但距離語意上，一個不可分離的物體來說，仍是不足。所以在這部分我們以多高斯分佈分別來描述顏色、形狀、材質等特徵。

但是對於影像上的一塊區域來說，顏色、材質的複雜程度是不可能預先知道的，而且除非我們能完全瞭解人類所可能表達的所有概念，否則對於使用者所想要表達的概念就不可能以固定數字的高斯數目來表示。針對這個問題，本實驗室交大資工系智慧型多媒體實驗室基於在語音上的研究與經驗，已經開發出可以在語音上，尋找高斯分布最適個數的方法[33]。以自我成長，並佐以BIC(Bayesian Information Criterion)判斷是否已達最適個數的方式。此方法在建立語者模型上已經得到了不錯的成效，因此在影像上，同樣是尋找高斯分布的最適個數的問題也採行這個方式而獲得解決。所以本計畫正是採行自我成長的方式來把使用者心中的概念影像以適當複雜度的 visual keyword 來表示。而為了能更進一步表達更複雜的概念，利用不同 visual keyword 之間相互的位置關係，定義出 visual string，使得使用者不但可以找尋白色的馬，更可以找尋站在綠色草地上的白馬。

然而由於不同複雜程度的特徵向量在相似程度的比對上以傳統使用 Euclidean 距離的方式或是任何以相同空間維度所表示的方法都不能適用。因為在任兩塊影像區域中我們不能保證他們經過處理後的高斯個數都會是相同的。因此我們發展了通用機率決策類神經網路(Generalized Probabilistic Decision-based Neural Network, GPDBNN)來做這些特徵的混合模型。通用機率決策類神經網路可針對因為複雜程度不同的輸入自動調整內部神經元個數，所以在不可預先知道影像複雜程度的情況下，通用機率決策類神經網路就可以解決影像特徵。而在通用機率決策類神經網路的訓練上，我們預先收集許多圖片，選擇具有相似物件的圖片群當作檢索影像。當檢索影像經過前處理，以及多高斯分布建立模型取得二維字串資料後，就會當成網路輸入。而 GPDNN 便可透過調整神經元個數以及參數尋得最適於檢索圖片群的高斯分布個數及參數。而其他將被比對圖片，經由(非精

確)的自動切割與多高斯分布建立模型，也輸入此網路後，便可自網路輸出得到相似值，透過相似值的比較，就可得到最相近於檢索圖片的影像。

結果與討論：

我們在本期計畫中除了發展適合影像檢索的核心 GPDNN、visual keyword、visual string、並發掘合用之影像特徵之外，我們還建立了直覺的檢索介面。我們也提供系統展示(http://140.113.216.66/WebImageSearch)可供外界使用。在這個系統中，我們認為對於一張影像中使用者真正感興趣的部分並不一定是整張影像，所以我們提供了點選機制。也就是說使用者只要在他感興趣的物件上以滑鼠點擊，系統就可以自動針對他所點擊的區域抽取相關的影像特徵，並且把系統自動找到的區域標示給使用者看，如果使用者覺得系統自動找到的區域並不完全包含他所感興趣的區域，使用者可以繼續點選直到他感興趣的區域完全被系統找到為止。而當使用者感興趣的區域被系統完全包含後，系統會只針對使用者所點擊的區域來搜尋，因此相較於針對整張影像作搜尋的方法可以有較好的準確率，而可達到七成以上。

參考文獻：

[1] Youtube 網站
http://www.youtube.com/
[2] Video.google 網站
http://video.google.com/
[3] Flickr 網站
http://www.flickr.com/
[4] Podcast
http://www.apple.com.tw/itunes/podcasts/
[5] 經濟部技術處 ITIS 產業資訊服務網及其市場報告
http://www.itis.org.tw
市場報告:
http://www.eettaiwan.com/ART_8800301924_617717,617727.HTM.f3197ae3
[6] 聯合報網路新聞
http://udn.com/NEWS/INFOTECH/INF1/1725705.shtml
[7] yahoo 搜尋引擎
http://tw.yahoo.com
[8] Jianbo Shi and Jitendra Malik, "Normalized cuts and image segmentation,"IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no.8, pp. 888–905, 2000.
[9] W.Y. Ma and B.S. Manjunath, "Edgeflow: a technique for boundary detection and image segmentation," IEEE Transactions on Image Processing, vol. 9, no.8, pp. 1375–1388, 2000.
[10] James Ze Wang, Jia Li, Robert M. Gray, and Gio Wiederhold, "Unsupervised multiresolution segmentation for images with low depth of field," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, no. 1, pp.85–90, 2001.
[12] James Ze Wang, Jia Li, and Gio Wiederhold, "SIMPLIcity: Semantics-sensitive integrated matching for picture LIbraries," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, no. 9, pp. 947–963, 2001.
[13] Y. Chen and J. Wang, "A region-based fuzzy feature matching approach to content-based image retrieval," IEEE Transactions on PAMI, vol. 24, no. 9, pp. 1252–1267, Sep. 2002.
[14] E. Jungert S. K. Chang and Y. Li, "Representation and retrieval of symbolic pictures using generalized 2d strings," University of Pittsburgh, vol. PA 15260, 1988.
[15] J. R. Smith and C.-S. Li., "Image classification and querying using composite region templates," Journal of Computer Vision and Image Understanding, 1999.
[16] Christos Faloutsos, Ron Barber, Myron Flickner, Jim Hafner, Wayne Niblack, Dragutin Petkovic, and William Equitz, "Efficient and effective querying by image content," Journal of Intelligent Information Systems, vol. 3, no. 3/4, pp. 231–262, 1994.
[17] Amarnath Gupta and Ramesh Jain, "Visual information retrieval," Communications of the ACM, vol. 40, no. 5, pp. 70–79, 1997.
[18] A. Pentland, R. Picard, and S. Sclaroff, "Photobook: Content-based manipulation of image databases," in Storage and Retrieval for Image, Video Databases II, San Jose, CA, Feb. 1994, SPIE, vol. 2185.
[19] John R. Smith and Shih-Fu Chang, "Visualseek: A fully automated content-based image query system," in ACM Multimedia, 1996, pp. 87–98.
[20] W. Y. Ma and B. S. Manjunath, "Netra: A toolbox for navigating large image databases," in Proc. IEEE Int'l Conf. Image Processing,

1997, pp. 568–571.
[21] Chad Carson, Megan Thomas, Serge Belongie, Joseph M. Hellerstein, and Jitendra Malik, “Blobworld: A system for region-based image indexing and retrieval,” in Third International Conference on Visual Information Systems, June 1999, pp. 509–516.
[22] Q. Y. Shi S. K. Chang and C. W. Yan, “Iconic indexing by 2-d strings,” IEEE Transactions on PAMI, vol. PAMI-9, no. 3, pp. 413–428, May 1987.
[23] M.C. Yang S.Y. Lee and J.W. Chen, “2d b-string: a spatial knowledge representation for image database systems,” in Proc. ICSC’92 Second Int. Computer Sci. Conf., 1992, pp. 609–915.
[24] John C. Russ, The Image Processing Handbook, Ron Powers, 3 edition, 1998. [25] Kayo Suzuki, Mitsuru Nagao, Hiroaki Ikeda and Yoshifumi Shimodaira, " Image Retrieval Using Sketched Image on Multimedia Networks: New Criteria for Design New Type of TV Sets," IEEE Trans. on Consumer Electronic, 2000.
[26]智慧型多媒體資訊處理系統的研究(二)(The study of intelligent multimedia information processing system(II)), 計劃編號：NSC 89-2213-E-009-094, 執行期限：88年8月1日至89年7月31日, 主持人：傅心家, 交通大學資訊工程學系.
[27] 智慧型網際網路新聞視訊查閱系統的研發(The study of the Intelligent Web-Based News Video Search System), 計畫編號：NSC 89-2213-E-009-015 (I), NSC 90-2213-E-009-047 (II), 執行期限：89年8月1日至91年7月31日, 主持人：傅心家, 交通大學資訊工程學系.
[28] Y. Chen and J. Wang, “A region-based fuzzy feature matching approach to content-based image retrieval,” IEEE Transactions on PAMI, vol. 24, no. 9, pp. 1252–1267, Sep. 2002.
[29] Y.Y. Xu T.M. Fang and H.C. Fu, ``Image classification and indexing by EM based Multiple-Instance Learning," in Proc. of pcm'2000, Sydney, Australia, 13-15 December 2000.
[30] Y.Y. Xu, H. T. Pao, and H.C. Fu, ``Image classification and indexing by EM based Multiple-Instance Learning,'' in Moroccan Journal of Control Computer
Science and Signal Processing, 2002.
[31] Hegel, G. W. F. Aesthetics, Vol. I trans. T. M. Knox, Clarendon Press, Oxford, 1975.
[32] H. C. Fu, H.Y. Chang, Y.Y. Xu, and H.T. Pao, ``User Adaptive Handwriting Recognition by Self-growing Probabilistic Decision-based Neural Networks," in IEEE Transaction on Neural Networks, Vol. 11, No.6, Nov. 2000.
[33] Y.H. Chen, C.L. Tseng, S.S. Cheng, Hsin-Chia Fu, and H.T. Pao, ``A self-growing probabilistic decision-based neural network with applications to anchor/speaker identification," in Proc. Of HIS2002, Santiago, Chile, Dec.1-4, 2002.
[34] Milind R. Naphade and Thomas S. Huang, “A probabilistic framework for semantic video indexing, filtering and retrieval,” IEEE Transactions on Multimedia, special issue on Multimedia over IP, vol. 3, no. 1, pp. 141–151, Mar. 2001