國立台灣師範大學
資訊工程研究所碩士論文

指導教授： 李忠謀 博士

以機率為基礎的語意分析之物件辨識研究

Generic Object Recognition Using Probabilistic-Based Semantic Component

研究生： 吳家維 撰

中華民國 九十八 年 六 月

中文摘要

以機率為基礎的語意分析之物件辨識研究

吳家維

使用影像中具有語意資訊的內容來作物件辨識，應該比使用低階特徵來辨識更為合理。為了克服語意隔閡，也就是高階與低階影像特徵之間的差距，我們提出一個非監督式的方法，藉由收集影像中的高階資訊，建構出一個新的影像表示法，我們將之命名為以機率為基礎的語意組成描述子(pSCD)。首先，我們將低階影像特徵量化，藉此得到一組視覺字組。接著我們利用修改過的pLSA模型來分析在視覺字組與影像間，包含哪些具有語意資訊的隱藏類別。利用這些隱藏類別，我們可以建構出pSCD，並將之應用在物件辨識上。另外，我們也會討論隱藏類別的數量多寡對pSCD的影響。最後，藉由物件辨識的實驗，我們證明了pSCD比起其它的影像表示法更加具有辨別性，例如袋字表示法或pLSA表示法。

關鍵字：物件辨識、語意隔閡、視覺字組、袋字模型、影像表示法

ABSTRACT

Generic Object Recognition Using Probabilistic-Based Semantic Component

by

Jia-Wei Wu

Object recognition based on semantic contents of images is more reasonable than that based on low-level image features. In order to bridge the semantic gap between low-level image features and high-level concepts in human cognition, we presents an unsupervised approach to build a new image representation, which is called probabilistic semantic component descriptor (pSCD), by collecting high-level concepts from images. We first quantize low-level features into a set of visual words, and then we apply a revised model of probabilistic Latent Semantic Analysis (pLSA) to analyze what kinds of hidden concepts between visual words and images are involved. After collecting these discovered concepts, we could build pSCD for object recognition. We also discuss how many hidden concepts are appropriate for pSCD to describe a set of images. Finally, through object recognition experiments, we demonstrate that pSCD is more discriminative than other image representations, including Bag-of-Words (BoW) and pLSA representations.

Keywords: object recognition, semantic gap, visual word, bag-of-words model,

image representation

誌謝

感謝李忠謀老師，在研究上給了我很大的揮灑空間，如果沒有老師的鼓勵和支持，也許就不會有到國外參加研討會的美好經驗了。感謝政杰學長，兩年來給予我的指導與幫忙，學長是個很無私且慷慨的人，我常常覺得，學長對我研究的貢獻，似乎比我自己還要多呢！感謝 Face 二人組—依佳和凱民，一直以來的互相討論和激勵，確實給了我很大的研究動力。感謝百璋、定翔、建斌和靖雅(按照筆劃多寡排序)，沒有你們，兩年的碩士生涯不可能如此多采多姿。最後，謝謝我的家人，總是在我最低潮的時候拉我一把，讓我有勇氣繼續向前邁進，謝謝你們。

目錄

圖目錄

表目錄

第一章　緒論

1.1 研究動機

在電腦視覺的領域中，物件辨識是個存在已久的問題。對於某些針對特定目標的辨識，現今已有了不錯的成效。但對於一般性物件的辨識，仍有非常多的挑戰要克服。物件辨識的第一步驟，通常會根據不同目的，從影像中擷取具有代表性的特徵，例如顏色、形狀、紋理(texture)、空間資訊...等等。這些被擷取出的特徵大都屬於低階的影像特徵，也就是說，它們僅具備影像處理階段的特性，卻不包含能夠描述影像內涵的語意資訊。但人類在辨識影像的內容時，所依賴的往往是更高階的，且帶有語意概念的特徵。所謂的語意隔閡(semantic gap) [28][15]，即是指低階特徵與高階特徵之間那道難以銜接的間隙。

如果能從高階的影像特徵來進行物件辨識，應該是更為合理的，且更接近人類理解影像的做法。因此，如何跨越語意隔閡，取得包含語意資訊的高階特徵，並藉此提升物件辨識的準確度，就成了我們的首要目標。

1.2 研究目的

為了處理語意隔閡的問題，我們設計了一個具備語意資訊的影像特徵，用以描述一張影像的內容。而此影像特徵，我們將之命名為「以機率為基礎的語意組

成描述子」(probabilistic semantic component descriptor)，以下簡稱為 pSCD。

基本上，要正確描述出一張影像內的組成，例如要明確的指出此張影像中包含行人、汽車、建築物等，是很困難的工作，因為這牽涉到精確的物體辨識問題。所以我們的想法是，先將多種不同類別的影像蒐集在一起，雖然不清楚這些影像的內容是由哪些成份組成，但我們可以分析這些組成份子中，約略可以分類成哪些群組，再利用這些群組來描述影像的內容。藉由以上的概念，我們建構出 pSCD 這個新的影像表示法(image representation)。

在此篇論文中，我們將探討 pSCD 的原理及特性，並將之應用在物件辨識的實驗中，藉此達到更好的辨識率。此外我們還會在實驗中，與另外兩種常見的影像表示法做比較，希望藉此看出 pSCD 的辨別性及強健性(robustness)。

1.3 研究範圍與限制

一般物件辨識是在已知有哪幾種類別的前提下，將輸入影像分類至某個類別中，因此有時也稱作物件分類(object categorization)。此篇論文中，我們將針對一張影像僅包含一種類別的物件做研究，而影像中同類物件的數量則不受限制。在此種狀況下，我們仍需面對以下問題：

1. 雜亂背景(cluttered background)：在物件辨識的問題中，雜亂背景一直是很大的干擾因素，它讓影像切割變得難以實行，因而無法過濾出物件所在的區域。如果無法預先判斷物件的位置，從影像中擷取的特徵將包含許多物件以外的

部分。而這些非物件的特徵將造成辨識時的混淆。

2. 姿勢變化：這邊的姿勢變化主要指的是大小及平移的改變，以及因視角不同所造成的姿勢變化。另外關於非剛性(nonrigid)物件，其自身各部位的相對移動也包含在姿勢變化中。

3. 光線變化：由於實驗的影像大都在不同的環境下拍攝，因此光線的改變是無法避免的。

除了影像的內容外，監督(supervision)的程度大小也是值得注意的問題。雖說數位影像的取得不虞匱乏，但包含文字註解的影像相對來講則少了許多。因此，如果監督的程度越小，訓練資料(training data)的取得就更加容易，在學習階段所受的限制自然也會越少。

在我們的方法中，pSCD 的建立是非監督式的，不需要任何影像註解或標籤(label)即可進行。而在分類器的學習階段，我們僅需知道每張影像中包含何種類別的物件，至於物件的位置和邊緣則完全不必知道，因此我們訓練分類器的方式屬於弱監督式的學習(weakly supervised learning)。

1.4 論文架構

本論文一共分為六個章節，其中第一章為序論，第二章為文獻探討，在第三章中，我們將詳細說明 pSCD 的原理和方法，第四章則會介紹我們使用的分類器，第五章為實驗結果與分析，第六章則是本論文的總結以及未來研究方向。

第二章　文獻探討

物件辨識可簡單分類為以模型為基礎(model-based) [1][29]或以外觀為基礎(appearance-based)的辨識方法[24]。以模型為基礎的物件辨識會利用三維模型來描述物件，例如立方體、球體、圓錐體、或其它更複雜的三維模型。而以外觀為基礎的方法，僅會使用影像上的二維資訊進行辨識。由於本研究的方法屬於後者，因此在接下來的篇幅中，僅針對以外觀為基礎的方法做討論。

早期以外觀為基礎的物件辨識研究中，從整張影像擷取出全域特徵(global features)是最常被使用的方法，例如顏色或紋理的直方圖[21][22][25]。它的優點在於擷取上，以及相似度的計算上都很方便，不過由於全域特徵包含了所有像素的資訊，因此非常容易受到視角、光線、遮擋、以及背景雜亂等因素的干擾。由於這些問題，使得全域特徵在近十年內，已逐漸被局部特徵所取代[34]。

近年來，利用特徵點(interest point)偵測與局部描述子(local descriptor)來擷取局部特徵的方法，在物件辨識的領域中相當盛行[8][32][26][23]。經由局部描述子的量化(quantization)可得到一個與文字相似的單位，在文獻中通常被稱為視覺字組(visual word)。所謂的袋字模型(bag-of-words model)，即是藉由各個視覺字組的出現次數來描述一張影像，而各個特徵點在影像上的空間關係則被忽略掉。因此，袋字模型不易受到遮擋與姿勢變化等問題的干擾。不過也因為喪失了空間

資訊，使得它無法區別前景和背景的特徵，所以仍舊無法避免背景雜亂的影響。

由於視覺字組與文字的相似性，使得原本應用在文件分析的機率模型，例如 pLSA (probabilistic latent semantic analysis) [10]以及 LDA (Latent Dirichlet Allocation) [3]，得以套用至影像的領域中。在本研究的方法中，pSCD 是以袋字模型為基礎，並修改 pLSA 的機率模型所得到的高階影像特徵。我們在底下將針對以上相關文獻做個介紹，並指出文獻中與我們的方法有哪些異同。

2.1 特徵點與局部描述子

特徵點的偵測有許多不同的方法[16][17][18][30] [24]。這些方法的目標，大都是希望特徵點在不同的視角或觀測環境中具有其不變性(invariance)。在[16]中提到的 DoG(difference-of-Gaussian)偵測法具有平移和比率(scale)不變性。另外有些方法，如[17][18][30]可以達到更高程度的不變性─仿射不變性(affine-invariant)。不過，不變性的程度越高，辨識上的效果不見得越好，因為某些對辨識有利的資訊可能會因此沒被選取到。除此之外，也有些特徵點的偵測法不考慮不變性的問題，例如沿著影像中的邊緣(edge)等距離選取特徵點，或是乾脆將影像等分切割成小方格，並將每個小方格當成一個特徵點。

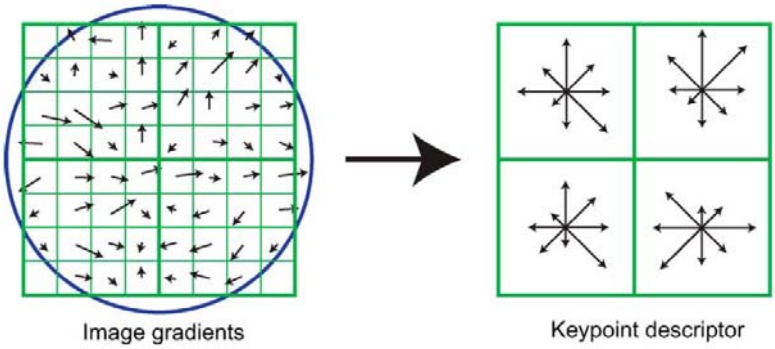


圖 2.1. SIFT 描述子的圖解[16]。值得注意的是，為了繪製上的方便，上圖僅使用了 2x2 個子區域。而實際上 SIFT 描述子總共使用了 4x4 個子區域。

除了利用特徵點達到不變性之外，局部描述子也可以具有不變性。例如 SIFT(scale invariant feature transform)描述子[16]對於旋轉、光線、及少量的幾何變形均具有不變性。簡單來說，SIFT 描述子為特徵點周圍灰階值的梯度統計結果，如圖 2.1 左邊所示，其中每個小箭頭均表示一個取樣點的梯度大小及方向，覆蓋在上方的圓圈則代表高斯分佈的權重。而圖 2.1 右邊的一個子區域，則代表 4x4 個取樣點經過統計後所得的梯度直方圖。若將所有直方圖的數值依序排列可得到一向量，此即為 SIFT 描述子。

目前 SIFT 描述子已廣泛被運用在電腦視覺各領域，並且已被證實，它的效能比現有的其它幾種描述子更好[19]，因此在我們的方法中亦採用 SIFT 描述子。此外，關於不同的特徵點偵測法和局部描述子，對影像分類結果的影響可在[23][12]中找到。

2.2 視覺字組

對影像的局部特徵進行分群(clustering)或量化，是產生視覺字組最常見的方

法。關於局部特徵可根據被分類的對象，套用不同的局部描述子，或是紋理、顏色等特徵[20]，而分群則大都使用 K-means。視覺字組目前被應用在許多領域，包含物件辨識[32]、景色分類[23]、影像註解[7]、及影像檢索[13]等等。

在[32]中，作者利用視覺字組構成的袋字表示法(bag-of-words representation)來描述一張影像，並將此應用在物件分類中。在最後的實驗，我們也會把 pSCD 和袋字表示法拿來做比較。

2.3 pLSA 和 LDA

pLSA 和 LDA 皆可藉由袋字模型來探索影像的隱藏類別。與 pLSA 相比，為了減少過適(overfitting)的問題，LDA 在它的多項分佈(multinomial distribution)中，加入了 Dirichlet 分佈作為共軛先驗(conjugate prior)。不過從[27]中可看出，兩者在影像分類的效果差異不大，且由於 pLSA 在使用及修改上較為容易，因此我們在本論文中採用了它。

近年來，pLSA 在影像分類的領域中得到不錯的成果[26]。但由於在袋字模型中，每張影像僅由視覺字組出現的次數來表示，並沒有考慮到特徵點在空間上的關係。因此，許多人修改了 pLSA 的模型，企圖在原本的演算法中加入空間資訊。例如在[14]的語意偏移(Semantic-Shift)演算法，以及[9]中所使用的 TSI-pLSA (Translation and Scale invariant pLSA)。

除了影像分類外，pLSA 與 LDA 亦被廣泛應用在影像註解的領域中[20][31][2]。

2.4 語意隔閡

如何減少語意隔閡，在影像檢索(image retrieval)的文獻中這是很常見的問題[15][Datt 08][Rui 99]。其中與機器學習有關的方法，可分成監督式與非監督式兩種。

在監督式的方法中，低階的影像特徵會與高階的語意資訊有明確的對應關係。因此在學習階段，必須要有大量的、包含標籤的訓練資料。而學習的方法可以使用 SVM(support vector machine)[33]、類神經網路(neural network)[Town 01]、或決策樹(decision tree)等。至於非監督式的學習方法，主要是利用輸入的特徵如何組織或群聚在一起的關係，藉此探索它們的高階語意資訊。常用的方法包括 K-means 與 NCut (Normalized cut)[Shi 00][Ng 02]等。而本研究的語意特徵擷取亦屬於非監督式的方法。

第三章　以機率為基礎的語意組成描述子

pSCD 是經由修改 pLSA 的機率模型所得到的影像表示法，而袋字模型則是 pLSA 得以應用在影像上的基礎。因此在第三章中，我們將從袋字模型及 pLSA 談起，最後再詳細說明 pSCD 的理論與方法。

3.1 袋字模型

袋字模型是一種利用視覺字組出現的次數來描述影像的方法。其步驟如圖 3.1 所示。

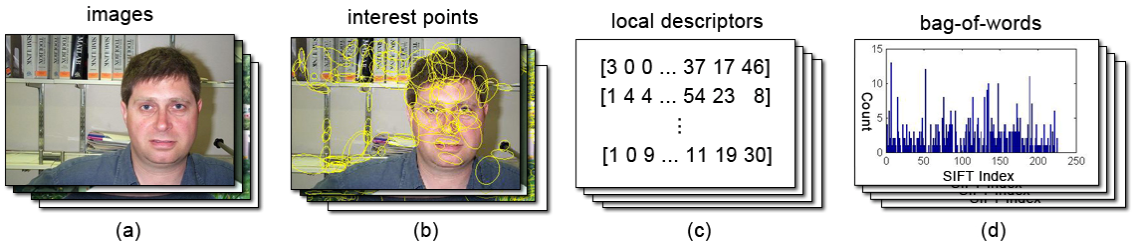


圖 3.1. 使用袋字模型描述影像的步驟。首先以不同大小及方向的橢圓區域，框出每張影像中的特徵點(b)，對每個橢圓區域計算出一個局部描述子(c)，接著將所有的局部描述子分成 k 類，即代表 k 個的視覺字組，最後藉由視覺字組出現的次數來表示一張影像(d)。

首先我們套用[17][18]這兩種方法，以不同大小及方向的橢圓區域，框出訓練資料中(training data)每張影像的特徵點(interest point)。並對每個橢圓區域計算出一個 SIFT 描述子(SIFT discriptor)，接著使用 K-means 分群法(K-means clustering)將全部影像的所有 SIFT 描述子分成 N_w 類。這 N_w 類的 SIFT 描述子，即代表了 N_w

個視覺字組，而這些視覺字組的集合則稱作一組字彙(vocabulary)。在圖 3.2 中，可看到兩個不同視覺字組的例子。當我們建立好一組字彙後，訓練資料中的任一張影像 d_i ，都可以由一個字組頻率向量(word frequency vector)來表示：

$$w(d_i) = \left(n(d_i, w_1), n(d_i, w_2), ..., n(d_i, w_{N_w}) \right) \quad (1)$$

其中 $n(d_i, w_j)$ 代表影像 d_i 中的 SIFT 描述子被分類到視覺字組 w_j 的數量。而此字組頻率向量亦被稱作袋字表示法(在本論文中，有時會簡稱為 BoW 表示法)。

另外，對於訓練資料以外的測試影像(test image)，一樣可以用袋字表示法來表示。我們利用已建立好的字彙，把測試影像中的每個 SIFT 描述子，歸類到距離最近的視覺字組，最後再統計每個視覺字組包含的數量即可。

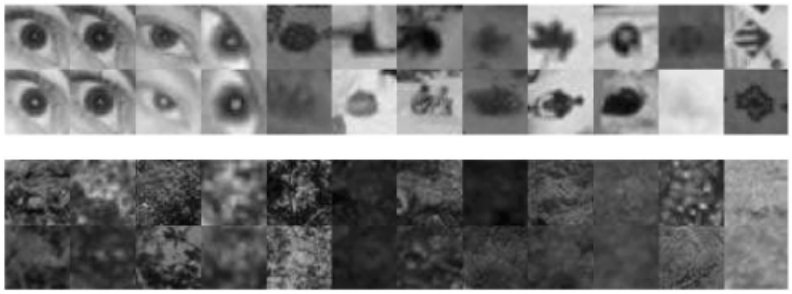


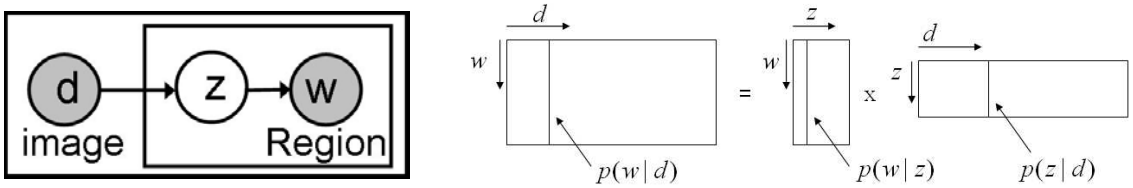
圖 3.2. 兩個不同視覺字組的例子[23]。

3.2 pLSA 模型

pLSA 是一個統計模型(statistical model)，最早由 T. Hofmann 在[10][11]中所提出，目的是為了用一個隱藏的類別變數(latent class variable)來描述一組文件和單字之間的關係。後來，藉由視覺字組的出現，使得 pLSA 也能被應用到物件辨識或影像分類的領域中。pLSA 的機率關係可由圖 3.3 表示。假如目前我們有 N_d 張影像，$\{d_1, d_2, ..., d_{N_d}\}$，且在這些影像中我們預先定義了 N_w 個視覺字組，

$\{w_1, w_2, ..., w_{N_w}\}$。在這些影像和視覺字組之間，我們假設有一個隱藏的語意資訊存在，而此語意資訊我們用預先定義好的 N_z 個隱藏類別來表示，$\{z_1, z_2, ..., z_{N_z}\}$。考慮上述定義，我們可以有以下的機率：

- $P(d_i)$：由這組影像內取出 d_i 這張影像的事前機率
- $P(z_k \mid d_i)$：已知影像 d_i 時，其中有隱藏類別 z_k 存在的機率
- $P(w_j \mid z_k)$：已知隱藏類別 z_k 時，會出現視覺字組 w_j 的機率



(a). pLSA 的圖模型　　(b). 機率上影像 d、視覺字組 w、與隱藏類別 z 之間的計算

圖 3.3. 在 pLSA 模型中，影像(d)與視覺字組(w)的機率關係。

從圖 3.3(a)的圖模型(graphical model)可看出，在已知隱藏類別 z_k 的情況下，d_i 和 w_j 是獨立的。由此關係可推導出下式

$$P(d_i, w_j) = P(d_i)P(w_j \mid d_i)\ ,\qquad P(w_j \mid d_i) = \sum_{k=1}^{N_z} P(w_j \mid z_k)P(z_k \mid d_i) \tag{2}$$

其中 $P(w_j \mid d_i)$ 的分解計算可從圖 3.3(b)的矩陣計算來觀察。我們從這組影像與視覺字組中取得由 $P(w \mid d)$ 構成的 $N_w \times N_d$ 矩陣，再設法拆解成兩個分別由 $P(w \mid z)$ 與 $P(z \mid d)$ 構成的矩陣相乘。

在 pLSA 模型的學習(learning)階段，為了求得 $P(z_k \mid d_i)$ 以及 $P(w_j \mid z_k)$ 的值，我們可以藉由最大概似估算法(maximum likelihood estimation)來計算。下式是我們欲最大化的概似函數(likelihood function)：

$$L = \prod_{i=1}^{N_d} \prod_{j=1}^{N_w} P(w_j \mid d_i)^{n(w_j, d_i)} \tag{3}$$

其中 $n(d_i, w_j)$ 代表在影像中 d_i 中出現視覺字組 w_j 的數量。

在有隱藏變數的模型中，EM(expectation maximization)演算法[4]是用來解決最大概似估算法的慣用手段，詳細的推導過程請參考[10]。經由概似函數的最大化，最後可得到 EM 演算法的 E 步驟及 M 步驟如下：

E-step

$$P(z_k \mid d_i, w_j) = \frac{P(w_j \mid z_k) P(z_k \mid d_i)}{\sum_l P(w_j \mid z_l) P(z_l \mid d_i)} \tag{4}$$

M-step

$$P(w_j \mid z_k) = \frac{\sum_i n(d_i, w_j) P(z_k \mid d_i, w_j)}{\sum_m \sum_i n(d_i, w_m) P(z_k \mid d_i, w_m)}, \tag{5}$$

$$P(z_k \mid d_i) = \frac{\sum_j n(d_i, w_j) P(z_k \mid d_i, w_j)}{n(d_i)} \tag{6}$$

其中 $n(d_i) = \sum_j n(d_i, w_j)$ 為影像 d_i 中所有視覺字組出現次數的總合，相當於影像 d_i 中特徵點的數量。另外，為了避免 pLSA 模型過適的問題，EM 演算法的循環計算會提早停止。通常循環的次數在 20-50 之間便已足夠。

我們可以利用求得的機率值 $P(z \mid d)$ 來做物件分類。例如在已知影像 d_i 的情況下，若 $z = z^*$ 時可以得到 $P(z \mid d_i)$ 的最大值，如底下公式(7)所示，則影像 d_i 將被分類到 z^* 這個類別中。

$$z^* = \arg\max_z P(z \mid d_i) \tag{7}$$

我們利用 pLSA 做了一個簡單的物件分類實驗，並從中挑出分類錯誤的例

子。如圖 3.4 所示，這幾張影像均被分類至同一個類別中。我們發現他們的相似處在於都擁有面積大且顏色均勻的背景，也就是說，他們並非因為前景，而是由於相似的背景而被錯分到同一個類別中。這樣的結果顯示，根據 $P(z|d)$ 所做的影像分類，會被某部分相似的視覺字組所支配，導致最後分類錯誤的結果。

圖 3.4. 使用 pLSA 作物件分類，被分至同一個類別的四張影像。

除了直接使用 $P(z|d_i)$ 的機率值做物件分類外，亦可把它當成另一種影像表示法[12][23]：

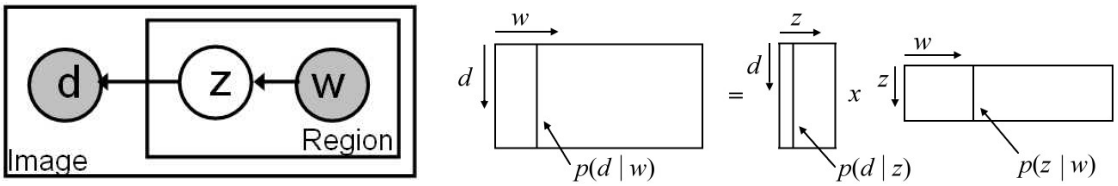
$$a(d_i) = \left(P(z_1|d_i), P(z_2|d_i), ..., P(z_{N_z}|d_i)\right) \tag{8}$$

在此篇論文中，我們將公式(8)稱作 pLSA 表示法。對於訓練資料以外的測試影像，則必須使用摺入法(fold-in method)[10]才能求得其 pLSA 表示法。首先利用訓練資料估算 $P(w|z)$，接著在固定此值的條件下，將測試影像代入原本 pLSA 的 EM 演算法中，換言之，在每次的 M 步驟，我們將不會更新 $P(w|z)$ 的值。使用此摺入法，最後便可求得測試影像的 $P(z|d)$，進而得到其 pLSA 表示法。

3.3 特徵擷取

考慮到在 pLSA 中，根據 $P(z|d_i)$ 的大小將影像 d_i 分類到某個 z^*，容易受到影像中不同成分的影響。我們把想法轉換成針對視覺字組來分析。因為每個視覺

字組，都代表著某些相似的橢圓區域。相較於整張圖片來說，針對這些面積較小的橢圓區域所做的分類可能更為可靠。於是我們把目標改為計算 $P(z|w)$，而非原本的 $P(z|d)$，藉此找出每個視覺字組最可能代表的隱藏語意。最後經由收集一張影像中出現過的視覺字組，以及它們各自的 $P(z|w)$，來組成這張影像的特徵向量。而這個帶有語意資訊的特徵向量，即是我們所謂的 pSCD。



(a). pSCD 的圖模型　　(b). 機率上影像 d、視覺字組 w、與隱藏語意 z 之間的計算

圖 3.5. 在 pSCD 中，影像(d)與視覺字組(w)的機率關係。

首先，我們將 pLSA 圖模型中的機率關係顛倒過來，如圖 3.5(a)所示。根據這個新的圖模型可將機率拆解如下：

$$P(d_i|w_j)=\sum_{k=1}^{N_z}P(d_i|z_k)P(z_k|w_j) \tag{9}$$

定理一：

假定我們的圖模型如圖 3.5(a)所示，則 $P(d_i|z_k)$ 與 $P(z_k|w_j)$ 可使用以下的 EM 演算法來估算：

E-step

$$P(z_k|d_i,w_j)=\frac{P(d_i|z_k)P(z_k|w_j)}{\sum_l P(d_i|z_l)P(z_l|w_j)} \tag{10}$$

M-step

$$P(d_i \mid z_k) = \frac{\sum_j n(d_i, w_j) P(z_k \mid d_i, w_j)}{\sum_m \sum_j n(d_j, w_m) P(z_k \mid d_j, w_m)}, \tag{11}$$

$$P(z_k \mid w_j) = \frac{\sum_i n(d_i, w_j) P(z_k \mid d_i, w_j)}{n(w_j)} \tag{12}$$

其中，$n(w_j) = \sum_i n(d_i, w_j)$ 代表此視覺字組出現在所有影像中的數量總合。

證明：

詳細證明請見附錄。

比較圖 3.5 (b)與圖 3.3 (b)可以發現，pSCD 乍看之下只是 pLSA 轉置後的結果，但隱藏語意在兩個模型上的意義卻有所不同。在 pLSA 的模型中，隱藏語意 z_k 存在的機率依賴於影像中，視覺字組的分佈情況；也就是說，若兩張影像所包含的視覺字組分佈類似，那麼 z_k 存在於此兩張影像的機率也會相似。然而在 pSCD 中，由於我們顛倒了原本在 pLSA 中的機率關係，隱藏語意 z_k 的機率將依賴於視覺字組在各個影像中的分佈情況；換句話說，兩個視覺字組在各影像中的分佈相似，那麼它們含有隱藏語意 z_k 的機率也會相差不多。

求得 $P(z \mid w)$ 之後，針對某個視覺字組，其最有可能代表的隱藏語意為：

$$z^* = \arg\max_z p(z \mid w_j) \tag{13}$$

由於這是一個非監督式的判斷方法，直接用最高分的 z^* 來描述這個視覺字組有些太過武斷，因此我們除了找出 $P(z \mid w_j)$ 中最高分的 z^* 之外，也取出第二高分

的 z^+：

$$z^+ = \arg\max_{z \neq z^*} p(z \mid w_j) \tag{14}$$

接著，我們計算出前兩名的差距 $\eta(w_j)$。藉此來判斷 w_j 最可能代表的隱藏語意，其可靠的程度有多少。

$$\eta(w_j) = P(z^* \mid w_j) - P(z^+ \mid w_j) \tag{15}$$

$\eta(w_j)$ 的值越大，代表使用 z^* 來描述這個視覺字組的可靠性越高。因此我們設定一個門檻值 h，所有 $\eta(w_j)$ 大於 h 的 $P(z^* \mid w_j)$ 才保留下來。也就是說，我們利用這個門檻值，過濾掉隱藏語意不是很明確的視覺字組。為此我們定義了一個函數 $\kappa(w_j, h)$。

$$\kappa(w_j, h) = \begin{cases} 1, \text{ if } \eta(w_j) \geq h \\ 0, \text{ if } \eta(w_j) < h \end{cases} \tag{16}$$

在後面的實驗中，為了比較上的方便，我們將門檻值 h 設定成過濾掉的視覺字組比例。舉例來說，$h = 0.1$ 代表有 10%的視覺字組由於語意不夠明確被過濾掉。假設目前隱藏語意的數目有 N_z 個，針對某一張影像 d_i，我們定義其 pSCD 為一個 N_z 維的向量 $f_{d_i} = (f_1^{d_i}, \ldots, f_{N_z}^{d_i})$：

$$f_k^{d_i} = \sum_{w_j \in d_i} \delta(z_k, z^*) \cdot \kappa(w_j) \cdot P(z^* \mid w_j) \tag{17}$$

其中 δ 是 Kronecker's delta function：

$$\delta(x, y) = \begin{cases} 1, \text{if } x = y \\ 0, \text{if } x \neq y \end{cases} \tag{18}$$

以上是擷取 pSCD 的方法。簡單來說，pSCD 收集了影像中語意較明確的資

訊，藉此來描述一張影像。由於 pSCD 是非監督式的方法，每個隱藏語意不見得能用精確的辭彙去描述。

另外，在擷取 pSCD 的過程中，有兩個必須預先決定好的參數。第一，我們必須先確認 N_z 的值，也就是隱藏語意的數量。N_z 的數值代表我們希望拿多少種語意資訊來描述這組影像。很明顯的，不同的 N_z 會造成 pSCD 維度的改變。對某些內容複雜的影像來說，太小的 N_z，容易造成語意資訊的喪失或混淆；但對於內容單純的影像來講，N_z 的值過大，也會使得語意資訊失去辨別性。因此如何選擇一個適當的 N_z 值，和我們所使用的資料集是有密切關聯的。第二，為了完成 pSCD 的擷取，我們還必須決定門檻值 h 的大小。因此在實際擷取 pSCD 之前，我們會先設計一個實驗，觀察不同的 N_z 以及 h 對實驗的影響，藉此來決定它們的數值。

第四章　物件辨識

當我們從某個多類別的資料集(data set)中，建立好一組字彙以及相對應的 $P(z|w)$ 後，就可以使用袋字表示法或 pSCD 來描述任何一張影像。在擷取出影像的特徵向量之後，我們選擇了 k-NN(k-nearest neighbor) [5]以及 GMM(Gaussian mixture model)作為分類器來進行辨識。

4.1 k-NN

由於 k-NN 不需要複雜的訓練過程，在使用上相當容易，對於辨識錯誤的影像進行分析也很方便。在我們實作的 k-NN 中，兩個物件之間的距離，即為它們的特徵向量在高維度空間中的歐基里德距離(Euclidean distance)。至於 k 值的大小，我們在四個類別的物件辨識實驗中，嘗試過 1 到 15 之間的數值，其中在 k=9 所得到的辨識率通常較高且較為穩定，因此我們在其它實驗中，亦使用 k=9 的 k-NN 來進行分類。

4.2 GMM

GMM 原本的功能在於，利用多個高斯分佈的組合，來近似於一個複雜的資料分佈。若要將 GMM 當作分類器，首先在訓練階段，我們必須對每一個類別的

資料建立一個 GMM。之後當測試資料輸入時，我們將該筆資料代入每個類別的 GMM 中，其中最大者，便將測試資料分類至該類別。

由於 GMM 是由多個高斯分佈組合而成，因此 GMM 分類器的效能，與其包含的高斯分佈個數有密切的關聯。通常高斯分佈的個數越高，對於訓練資料的辨識率也會跟著上升。但對於測試資料的辨識率，則會先上升一小段，接著由於過適的情況發生導致辨識率開始下滑。為了要決定 GMM 中高斯分佈的個數，我們會在訓練階段，利用訓練資料本身的交叉驗證(cross validation)，找出最合適的值。

第五章　實驗結果與分析

在此章中，首先會介紹我們使用的資料集，接著會談到實驗設定，包括視覺字組的建立與 pSCD 參數的決定。最後的實驗數據，我們將依照辨識的物件類別數來討論，其中包含了四個、七個與十個類別的辨識結果。

5.1 資料集

我們所使用的資料集為 Caltech 101-Object [6]。其中包含了 101 類的物件，且每類的影像數量不等，最少為 31 張，最多則包含 800 張。我們將由此資料集中，選出兩個子資料集，分別命名為 **D1** 與 **D2** 如下：

- **D1**：包含了四個類別的物件(飛機、機車、人臉、美洲豹)，其中每個類別均有 200 張影像。與 **D2** 相比，**D1** 中每個類別包含的影像數量較多，也因此會有較多的訓練資料。
- **D2**：包含了十個類別的物件(飛機、機車、人臉、美洲豹、手錶、盆景、汽車(側面)、雙桅縱帆船、枝型吊燈、玳瑁龜)，如圖 4.1 所示，其中前四個類別與 **D1** 相同。由於張數最少的玳瑁龜在 Caltech 101-Object 中僅包含 100 張影像，因此在 **D2** 中，每個類別的影像均只有 100 張，全部總共有 1000 張影像。

由於 SIFT 描述子的計算，僅需要影像上灰階值的梯度資訊，因此我們將所有的彩色影像先轉成灰階。另外，特徵點的數量與影像大小有很密切的關係，越大的影像，被偵測到的特徵點數量通常也越多。在我們使用的資料集中，同一個類別的影像大小相近，但不同類別的影像大小卻有很大的差異。為了不讓特徵點的數量多寡，影響最後的辨識結果，因此我們統一把所有影像的長邊，在維持原圖比例的情況下使用雙立方內插法(bicubic interpolation)縮放至 300 像素。

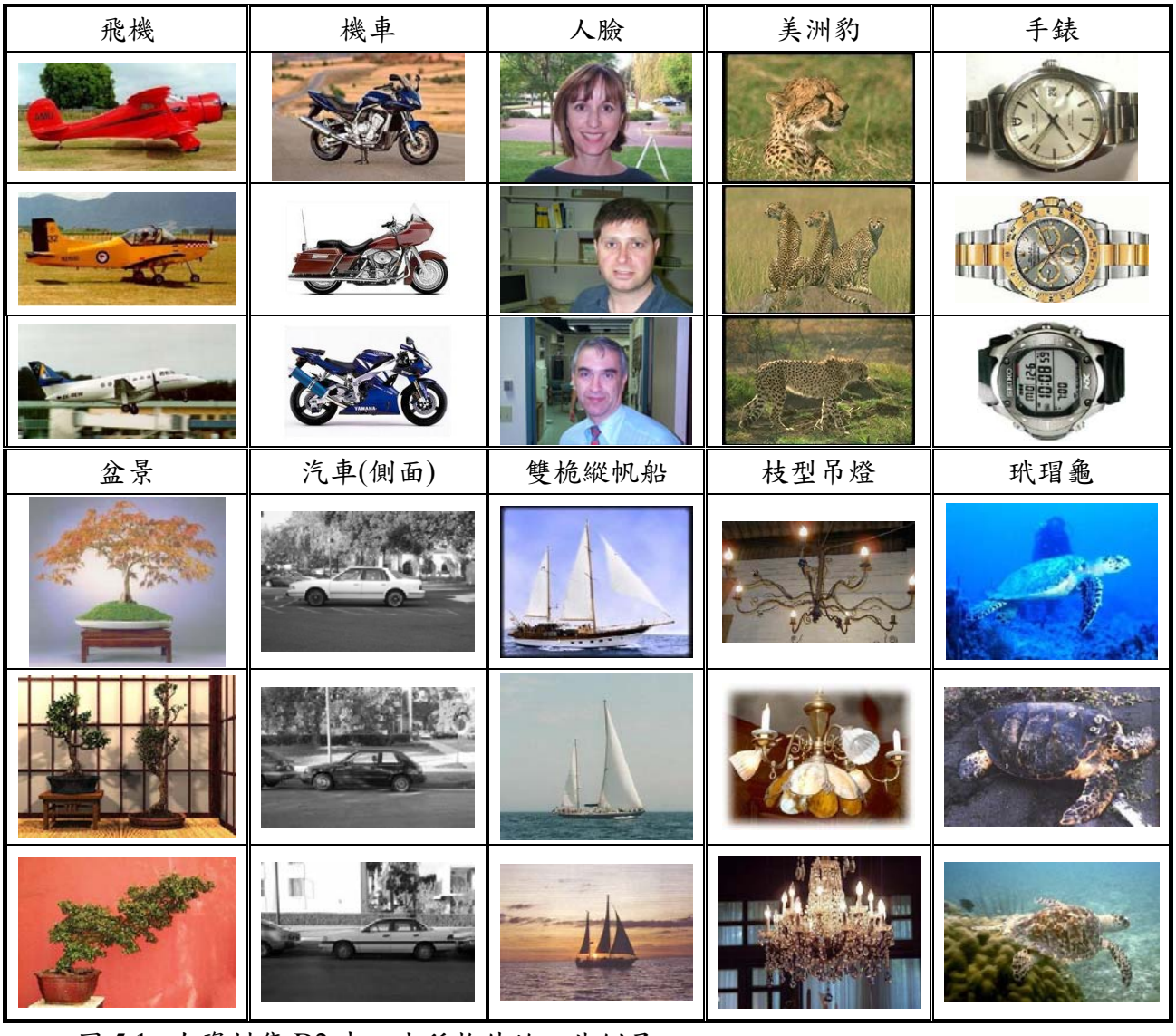


圖 5.1. 在資料集 D2 中，十種物件的一些例子。

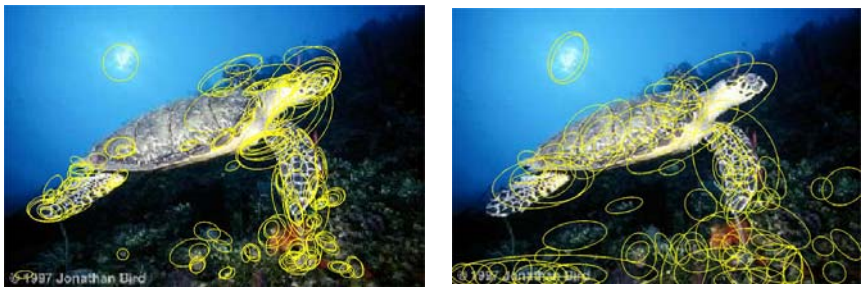
5.2 實驗設定

在建立袋字模型的階段，我們必須先確定視覺字組的數量，以及建立字彙所需要的資料集。此外，pSCD 的兩個參數：(1) 隱藏語意的數量 N_z，與(2) 欲過濾掉的視覺字組比例 h，也必須在正式進行物件辨識前決定好。我們會在此節中，詳細討論字彙的建立以及 pSCD 參數的決定。

5.2.1 視覺字組的建立

在[32]中有提到，人們的直覺通常會以為，若建立字彙所使用的資料集，與分類器學習時採用的訓練資料相同，最後可得到較好的辨識效果，然而實際上並非如此。由[23]與[32]的實驗皆可看出，只要建立字彙所用的資料集包含的類別與影像數量夠多，最後的辨識結果並不會有太大的差異。因此，我們可以使用某個特定的資料集，建立出一組通用字彙，並將之應用在其它資料集的特徵擷取中。

在接下來的實驗裡，我們選擇了類別數較多的資料集 **D2** 來建立一組通用字彙。首先，我們套用[17][18]的方法，擷取出兩組特徵點，如圖 5.2 所示。在資料集 **D2** 的 1000 張影像中，大約偵測到 330K 個特徵點。從相關文獻[12][17][18]中可發現，500 至 1000 左右的字彙數是很常見的選擇。因此，在本研究中我們設定的字彙數為 825 個，約略等同於一個視覺字組由 400 個特徵點所構成。



(a). 第一組特徵點[18]　　(b). 第二組特徵點[17]

圖 5.2. 我們採用的兩組特徵點的例子。

5.2.2 pSCD 參數的決定

為了決定 pSCD 的參數，我們設計了一個實驗，藉由不同的 N_z 與 h 來觀察 pSCD 的效果。首先利用四個類別的資料集 **D1** 訓練出一組 $P(z|w)$ 後，再擷取出 **D1** 中每張影像的 pSCD。接著，我們使用 k-NN (k = 9)，加上 leave-one-out 的方式來計算辨識率。不過，由於 $P(z|w)$ 是經由 EM 演算法求得，它的機率值不會是一個固定的數值。因此在這個實驗中，我們反覆跑了五次 EM 演算法，藉此得到五組不同的 pSCD。最後的辨識率，則是這五組特徵向量分別求得其辨識率的平均值。

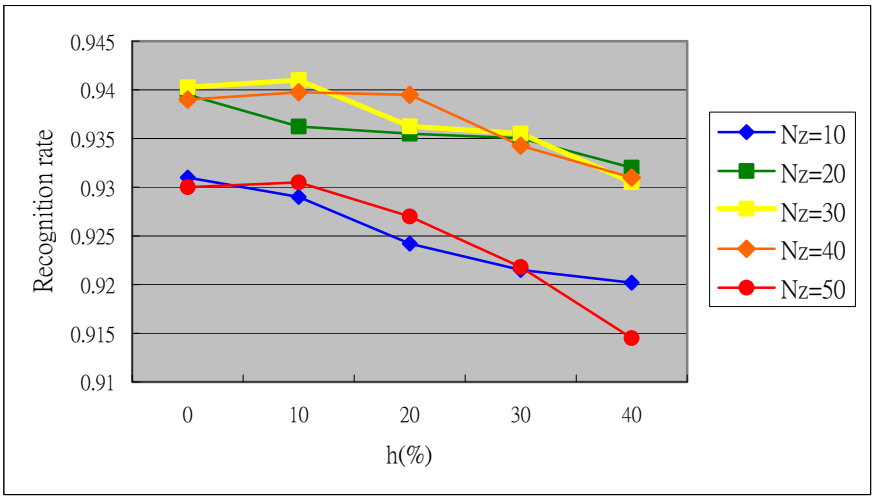


圖 5.3. 使用不同 N_z 與 h 的辨識率。

圖 5.3 我們是代入不同的 N_z 與 h 所得到的辨識率。由該圖可以看出，在四個類別的實驗中，隱藏語意的數量 N_z 在 20-40 之間有最好的辨識率。此外，當 N_z > 20 的時候，過濾掉 10%語意不明確的視覺字組對辨識率會有些微的提升。也許是因為當 N_z 較大的時候，擷取出的語意資訊會變得較為混亂，這時候適度的門檻值 h 便可過濾掉某些沒有辨別性的資訊。

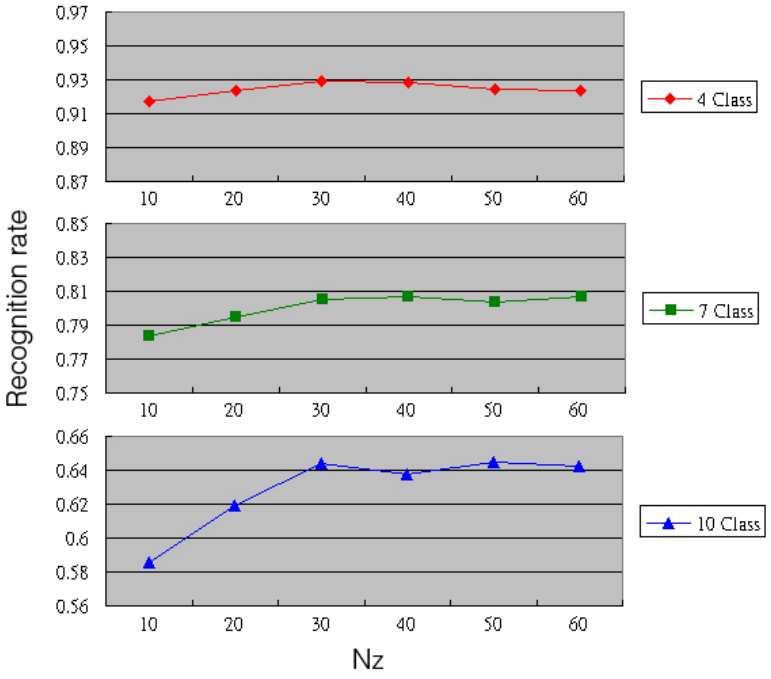


圖 5.4. 在類別數增加的情況下，N_z 對辨識率的影響。

接著，為了探討不同類別數對參數 N_z 的影響，首先我們令 h=0，並從資料集 **D2** 中挑選出四類、七類、與十類的物件 (關於類別的選擇請參考 5.3 節)，其餘設定則與之前的實驗相同。結果如圖 5.4 所示，在類別數較少的時候，容易受到 N_z 過大的影響，導致辨識率下滑。例如在四個類別的情況下，辨識率在 N_z 大於 30 後即開始下滑。反之，當類別數增加至七類或十類的時候，過小的 N_z 會有較大的影響，而辨識率在 $N_z \geq 30$ 的表現則較為穩定。

參考上述兩個實驗，我們決定將參數固定為 N_z=30 且 h=10%，在接下來的實驗中我們均使用這兩個參數值。

5.3 實驗結果

除了 pSCD 外，我們也採用了袋字表示法以及公式(8)的 pLSA 表示法作為比較。而各種影像表示法在實驗中的效能評估，我們均採用 4 折交叉驗證(4-fold cross-validation)來計算辨識率。也就是說，我們使用的資料集將被分成四等份，其中的每一份會輪流當作測試資料，而另外的三份則是訓練資料。因此我們在每個實驗中，皆可得到四個辨識結果的平均辨識率與標準差。

5.3.1 四個類別的物件辨識

在表 5.1 中，可看到三種影像表示法，在不同的分類器以及改變資料集大小的情況下，對於四類物件(飛機、人臉、美洲豹、機車)的辨識結果。其中，由於袋字表示法的維度遠大於我們的訓練資料量，在 GMM 的訓練過程中，會因為共變異矩陣(covariance matrix)的行列式為零導致訓練失敗。所以在表 5.1 中，即省略了袋字表示法在 GMM 分類器下的結果。

由表 5.1 可以很明顯的看出，在各種相同的條件下，pSCD 的辨識率遠勝過袋字表示法，若與 pLSA 表示法相比也有較優異的表現。此外，關於分類器的比較，由實驗數據可看出兩者的差異並不大，若只看 pSCD 的結果，k-NN 的效能

比 GMM 稍微好一些。至於資料集大小的影響，在此實驗中我們使用的資料集 **D2**，僅挑出與 **D1** 相同的四個類別，由於 **D2** 的每個類別只有 100 張，相當於訓練資料的數量僅有 **D1** 的一半。從表 5.1 可看出，在資料集大小減半的情況下，所有的辨識率都只有些微的下滑。

表 5.1. 四個類別實驗的辨識率。其中在辨識率右邊，小括號中的數據為標準差。

分類器	影像表示法	辨識率% (標準差)	
		D1 (800 張影像)	**D2** (400 張影像)
k-NN (k=9)	BoW	65.25 % (5.48)	65.00 % (5.23)
k-NN (k=9)	pLSA	89.38 % (1.44)	87.75 % (3.30)
k-NN (k=9)	pSCD	**93.13 % (1.11)**	**92.00 % (1.63)**
GMM	pLSA	89.75 % (2.33)	88.75 % (5.32)
GMM	pSCD	92.25 % (1.55)	90.50 % (2.08)

表 5.2. 四個類別實驗的混淆矩陣。此表格的數據為 pSCD 在使用資料集 D2 與 k-NN 下的辨識率。

整體辨識率% (標準差) = 92.00 % (1.63)				
真實類別↓	分類機率			
	飛機	人臉	美洲豹	機車
飛機	**0.93**	0.02	0.01	0.04
人臉	0.07	**0.89**	0.03	0.01
美洲豹	0.05	0.01	**0.94**	0
機車	0.05	0.01	0.02	**0.92**

為了更仔細的分析實驗數據，我們在表 5.2 中顯示出對於資料集 **D2**， pSCD 在 k-NN 辨識下的混淆矩陣(confusion matrix)。在此例中，混淆矩陣的第二列(row)、第一行(column)代表的意義為：實際上為人臉的影像，最後被歸類為飛機

的機率為 0.07。由於在四個類別的辨識實驗中，整體辨識率相當高，因此我們可以發現混淆矩陣的對角項除了人臉外，機率值幾乎都高於 0.9。

5.3.2 七個與十個類別的物件辨識

在接下來的兩個實驗中，我們將使用資料集 **D2**，逐步提高欲辨識的物件類別數，藉此觀察 pSCD 其辨別性的下滑趨勢。首先，我們由 **D2** 挑出原本在 Caltech 101-Object 中，數量較多的七個類別來作實驗。

表 5.3 為不同的影像表示法，分別使用 k-NN 與 GMM 的辨識結果。其中袋字表示法的辨識率已小於 0.5，而 pSCD 依舊有 0.8 左右的辨識率，與 pLSA 表示法相比也略微領先。表 5.4 則為 pSCD 在 k-NN 辨識下的混淆矩陣。在七種類別的情況中，我們可看出某些物件開始有較嚴重的混淆。例如實際上為手錶的影像，會有將近五分之一的機率被歸類成飛機。造成此種情況的可能因素為，由於 SIFT 描述子擷取出的資訊，較接近於紋理資訊或是局部形狀。而同樣由金屬材質構成的飛機與手錶，兩者皆包含許多較為細長的局部形狀，因此使得它們視覺字組的分佈較相似，進而造成 pSCD 的混淆。

表 5.3. 七個類別實驗的辨識率。

分類器	影像表示法	辨識率% (標準差)
k-NN (k=9)	BoW	46.14 % (4.83)
k-NN (k=9)	pLSA	78.14 % (1.64)
k-NN (k=9)	pSCD	**80.29 % (3.25)**
GMM	pLSA	76.71 % (4.02)
GMM	pSCD	78.86 % (1.48)

表 5.4. 七個類別實驗的混淆矩陣。

整體辨識率% (標準差) = 80.29 % (3.25)							
真實類別↓	分類機率						
	飛機	盆景	汽車(側)	人臉	美洲豹	機車	手錶
飛機	**0.91**	0.03	0.03	0.02	0	0.01	0
盆景	0.04	**0.71**	0.06	0.03	0.15	0.01	0
汽車(側面)	0.01	0	**0.92**	0.06	0.01	0	0
人臉	0.07	0.02	0.13	**0.76**	0.01	0	0.01
美洲豹	0.04	0.02	0.02	0.02	**0.9**	0	0
機車	0.03	0.03	0.04	0.01	0	**0.86**	0.03
手錶	0.19	0	0.04	0.08	0.06	0.07	**0.56**

在最後一個實驗中，我們使用資料集 **D2** 的所有類別，並重複之前的實驗步驟，得到表 5.5 與表 5.6 的數據。在十個類別全上的情況下，整體辨識率下降得很快。不過即使如此，pSCD 的辨識率依舊勝過其他兩種影像表示法。

表 5.5. 十個類別實驗的辨識率。

分類器	影像表示法	辨識率% (標準差)
k-NN (k=9)	BoW	30.0 % (3.64)
k-NN (k=9)	pLSA	59.3 % (2.20)
k-NN (k=9)	pSCD	**62.0 % (4.44)**
GMM	pLSA	59.7 % (3.05)
GMM	pSCD	60.2 % (2.97)

從表 5.6 (a)的混淆矩陣可看出，辨識效果最差的物件是枝型吊燈，其分類正確的機率僅有 0.34，且在表 5.6(b)中，當我們使用袋字表示法進行辨識，枝型吊燈亦得到最差的結果。參照圖 5.1 可發現，不同的枝型吊燈不論在材質或形狀上，均有很大的差異，所以很有可能，此類影像在袋字表示法的階段，已經不具備太

多有辨別性的資訊。而建立在袋字模型上的 pSCD，自然也很難取得有辨別性的高階特徵。

表 5.6. 十個類別實驗的混淆矩陣。

(a). 使用 k-NN，影像表示法為 pSCD。

整體辨識率% (標準差) = 62.0 % (4.44)										
真實類別↓	分類機率									
	飛機	盆景	汽車(側面)	枝型吊燈	人臉	玳瑁龜	雙桅縱帆船	美洲豹	機車	手錶
飛機	**0.85**	0.01	0.01	0.01	0.07	0	0.04	0	0	0.01
盆景	0.03	**0.56**	0.07	0.06	0.03	0.19	0	0.04	0.02	0
汽車(側面)	0.03	0	**0.82**	0	0.08	0.01	0.05	0.01	0	0
枝型吊燈	0.07	0.09	0.01	**0.34**	0.02	0.22	0.04	0.11	0.08	0.02
人臉	0.05	0	0.17	0.01	**0.59**	0.02	0.16	0	0	0
玳瑁龜	0.03	0.02	0.06	0.05	0.02	**0.47**	0.03	0.31	0	0.01
雙桅縱帆船	0.13	0.01	0.12	0.01	0.2	0.06	**0.43**	0.03	0.01	0
美洲豹	0.02	0.01	0.01	0	0	0.21	0.03	**0.72**	0	0
機車	0.01	0.02	0.02	0.03	0.01	0.01	0	0	**0.88**	0.02
手錶	0.15	0	0.07	0.01	0.05	0.02	0.09	0.03	0.04	**0.54**

(b). 使用 k-NN，影像表示法為袋字表示法。

整體辨識率% (標準差) = 30.0 % (3.64)										
真實類別↓	分類機率									
	飛機	盆景	汽車(側面)	枝型吊燈	人臉	玳瑁龜	雙桅縱帆船	美洲豹	機車	手錶
飛機	**0.47**	0	0	0	0.01	0.08	0.38	0.06	0	0
盆景	0.29	**0.15**	0.01	0	0.03	0.13	0.11	0.28	0	0
汽車(側面)	0.02	0	**0.12**	0	0.05	0.11	0.58	0.12	0	0
枝型吊燈	0.17	0.03	0.02	**0.04**	0.03	0.14	0.16	0.39	0.02	0
人臉	0.06	0	0	0	**0.45**	0.05	0.41	0.02	0.01	0
玳瑁龜	0.03	0	0	0	0.02	**0.16**	0.14	0.65	0	0
雙桅縱帆船	0.11	0	0.01	0	0	0.18	**0.62**	0.08	0	0
美洲豹	0	0	0	0	0.01	0.16	0.25	**0.58**	0	0
機車	0.47	0	0	0	0.03	0.05	0.13	0.16	**0.16**	0
手錶	0.17	0	0.03	0.01	0.06	0.05	0.31	0.11	0.01	**0.25**

第六章　結論

6.1 結論

pSCD 的建立是一個非監督式的過程。在沒有任何影像註解或標籤的情況下，我們藉由分析大量的影像資料與視覺字組的關係，從中擷取隱藏的語意資訊。而 pSCD 即為這些隱藏語意中，較為可靠的部分所組成的影像表示法。與傳統的袋字表示法相比，pSCD 顯得更為簡潔，保留下的資訊也更有辨別性，因此我們不需要使用具有核心函數(kernel function)或是推進演算法(boosting algorithm)等較為複雜的分類器，即可得到良好的辨識效果。

6.2 未來展望

在 pSCD 的節取過程中，如何決定隱藏語意的數量 N_z，以及門檻值 h 的大小是一個困難的問題。在本論文中，我們是利用實驗的方式，挑選出一組固定的參數值。不過在辨識的物件類別數增加的情況下，N_z 與 h 也應該有所調整。因此在未來的研究中，發展一個理論性的評估方式，藉此在不同的狀況下動態決定參數值是有必要的。

此外，除了物件辨識的領域，我們亦可將 pSCD 拓展至別的應用中。例如，

藉由隱藏語意與影像標籤的對應，也許可以把 pSCD 與影像檢索或影像註解結合在一起。這些都是在未來研究中，值得深思探討的問題。

參考文獻

[1] P. Besl and R. Jain, "Three-Dimensional Object Recognition," *ACM Computing Surveys*, *vol. 17*, no. 1, pp. 75-145, 1985.

[2] D. Blei and M. Jordan, "Modeling Annotated Data," Technical Report CSD-02-1202, U.C. Berkeley Computer Science Division, 2002.

[3] D. Blei, Y. Andrew, and M. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, *vol. 3*, pp. 993-1020, 2003.

[4] A. P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *J. Royal Statiscal Soc., Ser. R*, *vol. 39*, no. 1, pp. 1-38, 1977.

[5] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd Ed. Wiley, 2001.

[6] L. Fei-Fei, R. Fergus, and P. Perona, "Learning Generative Visual Models from Few Training Examples: An Incremental Bayesian Approach Tested on 101 Object Categories," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, *vol. 12*, pp. 178-187, 2004.

[7] S. L. Feng, R. Manmatha, and V. Lavrenko, "Multiple Bernoulli Relevance Models for Image and Video Annotation," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, *vol. 2*, pp. 1002−1009, 2004.

[8] R. Fergus, P. Perona, and A. Zisserman, "Object Class Recognition by Unsupervised Scale-Invariant Learning," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, *vol. 2*, pp. 264-271, 2003.

[9] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman, "Learning Object Categories from Google's Image Search," *Proc. IEEE Int'l Conf. Computer Vision*, *vol. 2*, pp. 1816-1823, 2005.

[10] T. Hofmann, "Unsupervised Learning by Probabilistic Latent Semantic Analysis," *Machine Learning*, *vol. 42*, pp. 177-196, 2001.

[11] T. Hofmann, "Probabilistic Latent Semantic Indexing," *Proc. ACM SIGIR*, pp. 50-57, 1999.

[12] E. Hörster, T. Greif, R. Lienhart, M. Slaney, "Comparing Local Feature Descriptors in pLSA-Based Image Models," *Lecture Notes in Computer Science*, *vol. 5096*, pp. 446-455, 2008.

[13] F. Jing, M. Li, H.- J. Zhang, and B. Zhang, "An Efficient and Effective Region-Based Image Retrieval Framework," *IEEE Trans. Image Processing*, *vol. 13*, no. 5, pp. 699-709, 2004.

[14] D. Liu, and T. Chen, "Semantic-Shift for Unsupervised Object Detection," *Proc. IEEE Computer Vision and Pattern Recognition Workshop on Beyond Patches*, pp. 16-23, 2006.

[15] Y. Liu, D. Zhang, G. Lu, and W. Ying Ma, "A Survey of Content-Based Image Retrieval with High-Level Semantics," *Pattern Recognition*, *vol. 40*, no. 1, pp. 262-282, 2007.

[16] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *Int'l Journal of Computer Vision*, *vol. 60*, no. 2, pp. 91-110, 2004.

[17] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust Wide Baseline Stereo from Maximally Stable Extremal Regions," *Proc. British Machine Vision Computing*, pp. 384-393, 2002.

[18] K. Mikolajczyk and C. Schmid, "Scale and Affine Invariant Interest Point Detectors," *Int'l Journal of Computer Vision*, *vol. 60*, no. 1, pp. 63-86, 2004.

[19] K. Mikolajczyk and C. Schmid, "A Performance Evaluation of Local Descriptors," *IEEE Trans. Pattern Analysis and Machine Intelligence*, *vol. 27*, no. 10, pp. 1615-1630, 2005.

[20] F. Monay and D. Gatica-Perez, "Modeling Semantic Aspects for Cross-Media Image Indexing," *IEEE Trans. Pattern Analysis and Machine Intelligence*, *vol. 29*, no. 10, pp. 1802-1817, 2007.

[21] W. Niblack, R. Barber, W. Equitz, M. Fickner, E. Glasman, D. Petkovic, and P.

Yanker, “The QBIC project: Querying Images by Content Using Color, Texture and Shape,” *Proc. Storage and Retrieval for Image and Video Databases*, *vol. 1908*, pp. 173-187, 1993.

[22] M. Pontil and A. Verri, “Support Vector Machines for 3D Object Recognition,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, *vol. 20*, no. 6, pp. 637-646, 1998.

[23] P. Quelhas, F. Monay, J. Odobez, D. Gatica-Perez, and T. Tuytelaars, “A Thousand Words in a Scene,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, *vol. 29*, no. 9, pp. 1575-1589, 2007.

[24] P. M. Roth and M. Winter, “Survey of Appearance-based Methods for Object Recognition,” Technical Report ICG-TR-01/08, Graz University of Technology, Institute for Computer Graphics and Vision, 2008.

[25] B. Schiele and J. Crowley, “Recognition without Correspondence Using Multidimensional Receptive Field Histograms,” *Int’l Journal of Computer Vision*, *vol. 36*, no. 1, pp. 31-50, 2000.

[26] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman, “Discovering Objects and Their Location in Images,” *Proc. IEEE Int’l Conf. Computer Vision*, *vol. 1*, pp. 370-377, 2005.

[27] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman, “Discovering Object Categories in Image Collections,” Technical report, CSAIL, Massachusetts Institute of Technology, 2005.

[28] A. W. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, “Content-Based Image Retrieval at the End of the Early Years,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, *vol. 22*, no. 12, pp. 1349-1380, 2000.

[29] M. J. Tarr, P. Williams, W. G. Hayward, and I. Gauthier, “Three-Dimensional Object Recognition is Viewpoint Dependent,” *Nature Neuroscience*, *vol. 1*, no. 4, pp. 275-277, 1998.

[30] T. Tuytelaars and L. Van Gool, “Matching Widely Separated Views Based on Affine Invariant Regions,” *Int’l Journal of Computer Vision*, *vol. 59*, no. 1, pp.

61-85, 2004.

[31] Y. Wang, T. Mei, S. Gong, X.-S. Hua, “Combining Global, Regional and Contextual Features for Automatic Image Annotation,” *Pattern Recognition*, *vol. 42*, pp. 259-266, 2009.

[32] J. Willamowski, D. Arregui, G. Csurka, C. R. Dance, and L. Fan, “Categorizing Nine Visual Classes Using Local Appearance Descriptors,” *Workshop on Learning for Adaptable Visual Systems (LAVS),* Cambridge, U.K., 2004.

[33] L. Zhang, F. Liu, B. Zhang, “Support Vector Machine Learning for Image Retrieval,” *Int’l Conf. Image Processing*, *vol. 2*, pp. 721-724, 2001.

[34] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid, “Local Features and Kernels for Classification of Texture and Object Categories: A Comprehensive Study,” *Int’l Journal of Computer Vision*, *vol. 73*, no. 2, pp. 213-238, 2007.

附錄　定理一的證明

定理一：

假定我們的圖模型如圖 3.5 所示，則 $P(d_i \mid z_k)$ 與 $P(z_k \mid w_j)$ 可使用以下的 EM 演算法來估算：

E-step

$$P(z_k \mid d_i, w_j) = \frac{P(d_i \mid z_k)P(z_k \mid w_j)}{\sum_l P(d_i \mid z_l)P(z_l \mid w_j)}$$

M-step

$$P(d_i \mid z_k) = \frac{\sum_j n(d_i, w_j)P(z_k \mid d_i, w_j)}{\sum_m \sum_j n(d_j, w_m)P(z_k \mid d_j, w_m)},$$

$$P(z_k \mid w_j) = \frac{\sum_i n(d_i, w_j)P(z_k \mid d_i, w_j)}{n(w_j)}$$

其中，$n(w_j) = \sum_i n(d_i, w_j)$ 代表此視覺字組出現在所有影像中的數量總合。

證明：

(1) 參考[10]中 pLSA 的推導，我們可藉由 d 與 w 兩者的互換來證明定理一。

(2) 此外，我們亦可由原本 pLSA 的 EM 演算法，即公式(4-6)先求得 $P(z_k \mid d_i)$ 與 $P(w_j \mid z_k)$，再藉由貝式定理的轉換得到與定理一相同的結果，證明如下：

假設在 pLSA 的 EM 演算法中，第 t 次疊代求得的機率值分別為 $P_t(z_k \mid d_i, w_j)$、$P_t(w_j \mid z_k)$、$P_t(z_k \mid d_i)$。而在 pSCD 的 EM 演算法中，第 t 次疊代

的機率值分別為 $P_t^*(z_k \mid d_i, w_j)$ 、 $P_t^*(z_k \mid w_j)$ 、 $P_t^*(d_i \mid z_k)$ 。此外，

$$N = \sum_i \sum_j n(d_i, w_j),\ P(d_i) = \frac{1}{N} \sum_j n(d_i, w_j),\ P(w_j) = \frac{1}{N} \sum_i n(d_i, w_j) \tag{A.1}$$

其中 $n(d_i, w_j)$ 代表影像 d_i 中包含視覺字組 w_j 的數量。

在第一次疊代中， $P_t(z_k \mid d_i, w_j)$ 及 $P_t^*(z_k \mid d_i, w_j)$ 皆由亂數產生的 $P_t(w_j \mid z_k)$ 、 $P_t(z_k \mid d_i)$ 與 $P_t^*(z_k \mid w_j)$ 、 $P_t^*(d_i \mid z_k)$ 計算求得。為了讓 pLSA 與 pSCD 的 EM 演算法有相同的初始值，因此我們令 $P_t(z_k \mid d_i, w_j)$ 與 $P_t^*(z_k \mid d_i, w_j)$ 在 t =1 時有相同的機率值，即 $P_1(z_k \mid d_i, w_j) = P_1^*(z_k \mid d_i, w_j)$ 。

假設在 $t = s$ 時， $P_s(z_k \mid d_i, w_j) = P_s^*(z_k \mid d_i, w_j)$ 。接著在 pSCD 的 M-step 中：

$$\begin{aligned}
P_s^*(z_k \mid w_j) &= \frac{\sum_i n(d_i, w_j) P_s^*(z_k \mid d_i, w_j)}{n(w_j)} = \frac{\sum_i n(d_i, w_j) P_s(z_k \mid d_i, w_j)}{n(w_j)} \\
&= \frac{\sum_i n(d_i, w_j) P_s(z_k \mid d_i, w_j)}{\sum_m \sum_i n(d_i, w_m) P_s(z_k \mid d_i, w_m)} \times \frac{\sum_i \sum_m n(d_i, w_m) P_s(z_k \mid d_i, w_m)}{n(w_j)} \\
&= P_s(w_j \mid z_k) \times \sum_i \left(\frac{\sum_m n(d_i, w_m) P_s(z_k \mid d_i, w_m)}{n(d_i)} \times \frac{n(d_i)}{N} \right) \times \frac{N}{n(w_j)} \\
&= P_s(w_j \mid z_k) \times \sum_i \left(P_s(z_k \mid d_i) P(d_i) \right) \times \frac{1}{P(w_j)} \\
&= \frac{P_s(w_j \mid z_k) \times P_s(z_k)}{P(w_j)}
\end{aligned} \tag{A.2}$$

$$P_s^*(d_i \mid z_k) = \frac{\sum_j n(d_i, w_j) P_s^*(z_k \mid d_i, w_j)}{\sum_m \sum_j n(d_m, w_j) P_s^*(z_k \mid d_m, w_j)} = \frac{\sum_j n(d_i, w_j) P_s(z_k \mid d_i, w_j)}{\sum_m \sum_j n(d_m, w_j) P_s(z_k \mid d_m, w_j)}$$

$$= \frac{\sum_j \left(\frac{n(d_i, w_j) P_s(z_k \mid d_i, w_j)}{n(d_i)} \right) \times n(d_i)}{\sum_m \left(\sum_j \left(\frac{n(d_m, w_j) P_s(z_k \mid d_m, w_j)}{n(d_m)} \right) \times \frac{n(d_m)}{N} \right) \times N}$$

$$= \frac{P_s(z_k \mid d_i)}{\sum_m \left(P_s(z_k \mid d_m) \times P(d_m) \right)} \times P(d_i)$$

$$= \frac{P_s(z_k \mid d_i) \times P(d_i)}{P_s(z_k)} \tag{A.3}$$

由(A.2)與(A.3)可知，在 $P_s(z_k \mid d_i, w_j) = P_s^*(z_k \mid d_i, w_j)$ 的情況下，公式(10-12)的 $P_s^*(z_k \mid w_j)$ 、 $P_s^*(d_i \mid z_k)$ ，可經由公式(4-6)的 $P_s(w_j \mid z_k)$ 、 $P_s(z_k \mid d_i)$ 透過貝氏定理的轉換得到。

若 $t = s+1$ 時，考慮 pSCD 在 EM 演算法的 E-step：

$$P_{s+1}^*(z_k \mid d_i, w_j) = \frac{P_s^*(d_i \mid z_k) P_s^*(z_k \mid w_j)}{\sum_l P_s^*(d_i \mid z_l) P_s^*(z_l \mid w_j)}$$

$$\propto P_s^*(d_i \mid z_k) P_s^*(z_k \mid w_j) = \frac{P_s(z_k \mid d_i) P(d_i)}{P_s(z_k)} \times \frac{P_s(w_j \mid z_k) P_s(z_k)}{P(w_j)}$$

$$= P_s(z_k \mid d_i) P_s(w_j \mid z_k) \times \frac{P(d_i)}{P(w_j)}$$

$$\propto P_s(z_k \mid d_i) P_s(w_j \mid z_k)$$

$$\therefore \quad P_{s+1}^*(z_k \mid d_i, w_j) = \frac{P_s(z_k \mid d_i) P_s(w_j \mid z_k)}{\sum_l P_s(z_l \mid d_i) P_s(w_j \mid z_l)} = P_{s+1}(z_k \mid d_i, w_j) \tag{A.4}$$

由公式(A.2)與(A.3)得知，在 $P_{s+1}^*(z_k \mid d_i, w_j) = P_{s+1}(z_k \mid d_i, w_j)$ 的情況下，$P_{s+1}^*(z_k \mid w_j)$ 、 $P_{s+1}^*(d_i \mid z_k)$ 可經由 $P_{s+1}(w_j \mid z_k)$ 、 $P_{s+1}(z_k \mid d_i)$ 透過貝氏定理的轉換得到。

故我們可證明 pSCD 的 EM 演算法，可先經由 pLSA 的 EM 演算法，再透過貝式定理的轉換得到相同的結果。