Introduction to Low-Level Vision Learning

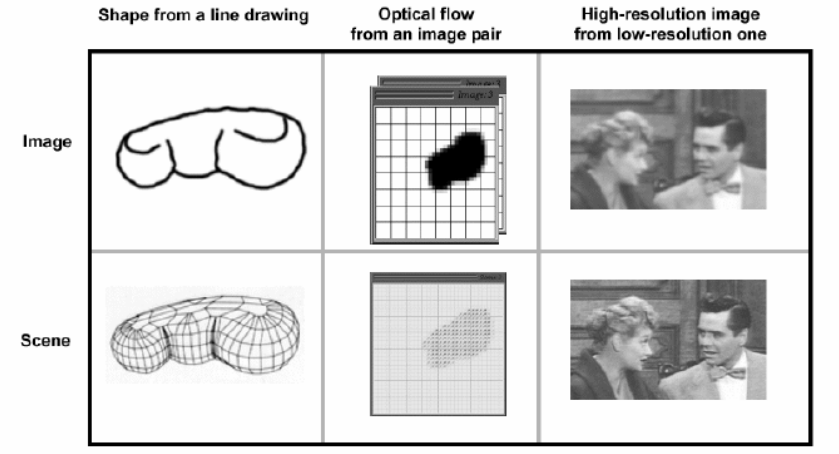
低階視覺學習簡介

鄭文皇 (wisley@cmlab.csie.ntu.edu.tw)，謝致仁 (jerry@cmlab.csie.ntu.edu.tw)

《內容大綱》

1. 低階視覺學習初步

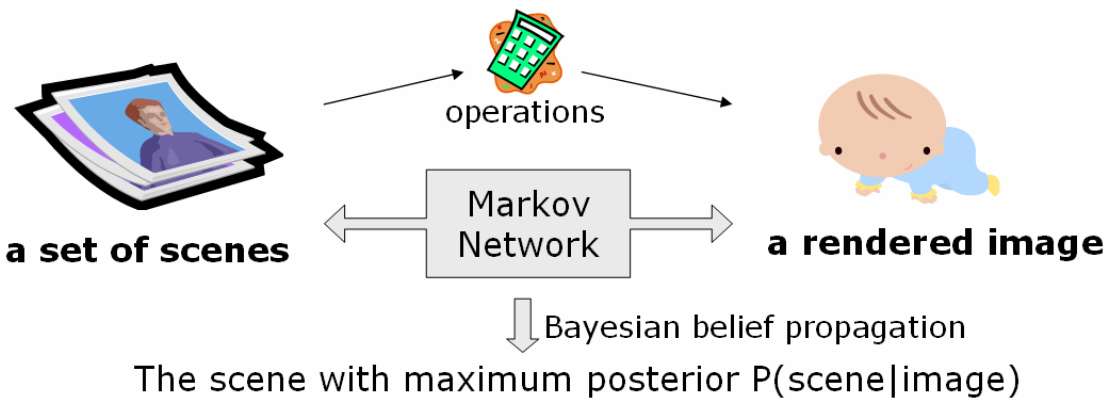
一般而言，低階視覺學習指的是在給定拍攝影像(image)資料的情況下，評估該影像所對應之真實場景(scene)參數的問題。舉例來說，所欲評估的場景參數可以是經投影後的物體速度(projected object velocity)、表面形狀與反射形態(surface shape & reflectance pattern)、缺損之高頻細節(missing high frequency detail)等等。



圖一、低階視覺學習問題實例。給定“影像”資訊，評估其所對應之真實“場景”。

在文獻中，傳統低階視覺學習問題的解法可分為兩大類。第一類是採用貝氏(Bayesian)與規則化(regularization)技巧。基本上，它們對於事件發生的事前機率(prior probability)或條件限制是經由假設而非學習而得，因此難以處理內容複雜的真實影像。第二類則是計算影像的統計資料，並將其與人類視覺系統(human visual system)的特性相關聯，例如用來分析並合成幾可亂真的影像材質(texture)。然而，它們並沒有處理視覺系統如何理解影像的問題，例如評估所對應的真實場景。

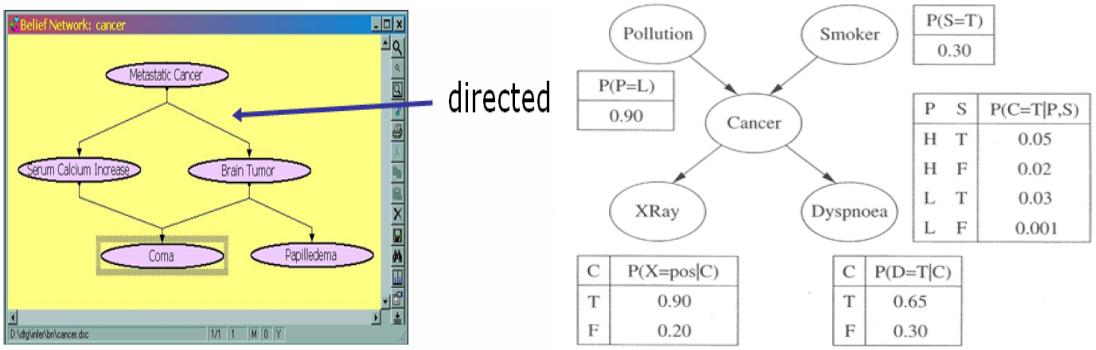
因此，目前文獻提出以機器學習的方式來整合兩類傳統解法的優點，稱為 VISTA 架構。具體來說，此種方式是針對某種場景參數，給定部份對應的影像與場景做為訓練資料(training data)，藉由研究影像的統計特性來學習如何由給定的任一影像推論(infer)出其相對應的場景。我們以圖二進一步說明，我們一開始先對原始真實場景進行特定運算(operations)以得到相對應的影像。此運算的選擇與我們的目的有關，若所欲評估的場景參數為移動估計(motion estimation)，則運算則是人為的影像平移；同樣地，若評估參數為超級解析度(super-resolution)，則運算為高頻濾波器(high-frequency filter)。如此一來，即可得到多組不同的場景/影像組，並透過馬可夫網路(Markov network)自動學習出影像與場景間的對應關係。得到這樣一個經訓練的馬可夫網路後，將來只要輸入任何一張影像，則可經由自動的貝氏可信度傳遞(Bayesian belief propagation)測試，找到該影像最可能的對應場景(使得事後機率 P(scene|image)最大)，而得知該影像所隱含的場景參數。



圖二、VISTA(Vision by Image/Scene TrAining)架構[1]。

2. 馬可夫網路

馬可夫網路是低階視覺學習架構的核心，其可視為貝氏網路(Bayesian network)的一種變形。與貝氏網路所不同之處在於其邊(edge)為無向性的(undirected)，因此可表示循環式的依賴關係(cyclic dependence)，圖三所示為貝氏網路的基本形式與應用於肺癌檢測關係推論的實例。



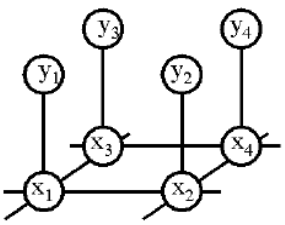
(a) (b)

圖三、貝氏網路(a)基本形式[6]與(b)推論應用實例。

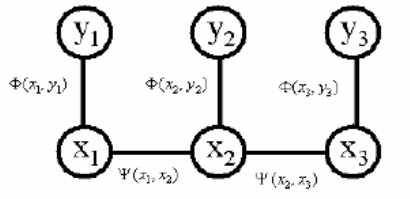
基本上，馬可夫網路可視為一個隨機變數(random variable)集合 X 中的所有聯合機率分布(joint probability distribution)的表示模型。就數學定義而言，一個馬可夫網路包含兩部份，分別為：

- 一個無向性圖(undirected graph) G = (V, E)，其中每個節點(vertex) v∈V 代表一個 X 中的隨機變數，而每個邊{*u*,*v*}∈E 代表隨機變數 u,v 間的依賴關係。
- 一群潛在函式(potential functions) φ_k 的集合，其中每一個函式對應於一個 G 中的派系(clique) k。每一個 φ_k 是一個由 k 中元素(element)所產生的可能聯合指定(joint assignment)到非負實數的對應。

在這裡我們先回憶一下之前所提的 VISTA 架構。在圖二中，我們知道當給定影像 y，其所得的對應場景即是使事後機率 P(x|y) = cP(x,y)為最大的場景 x，其中 c=1/P(y)為事前機率常數。此外，我們必需事先定義好事後機率 P 的損失函式(loss function)以得到一個最佳的場景評估 x。一般而言，損失函式需根據不同應用的特性決定，常用的有 MMSE(minimum mean squared error)及 MAP(maximum a posterior)兩種。為使影像/場景組能有效套入馬可夫網路，我們將影像與場景都分割成較小塊的補丁(patch)，然後將馬可夫網路中的一個節點對應到一個補丁，而連接的邊表示統計上的依賴關係，如圖四。當我們知道了位置 j 上的場景即可同時獲得該處影像與相鄰場景的資訊。因此，解決一個完整的馬可夫網路包含兩個階段：首先為學習階段(learning phase)，網路的連結參數將由訓練資料中學習而得；接著是推論階段(inference phase)，對應於特定影像的場景將由已訓練之網路中推論與評估而得。



圖四、用於低階視覺問題之馬可夫網路，每一個節點代表一個影像或場景的補丁。



圖五、適合性函式 Φ 與 Ψ。

在 VISTA 架構中，我們定義 Φ 與 Ψ 兩個適合性函式(compatibility function)，它們就是馬可夫網路中的潛在函式，由訓練資料中自動學習而得，如圖五。因此事後機率 P(x,y)可表示為

$$P(x_1,x_2,...,x_N,y_1,y_2,...,y_N)=\prod_{(i,j)}\Psi(x_i,x_j)\prod_k\Phi(x_k,y_k)\ ,$$

其中 N 為影像與場景的節點數。對於離散變數而言，以 MMSE 損失函式為基礎之 j 節點的事後機率 P 可寫為

$$\hat{x}_{jMMSE}=\sum_{x_j}x_j\sum_{all\ x_i,i\neq j}P(x_1,x_2,...,x_N,y_1,y_2,...,y_N)\ 。$$

同樣地，以 MAP 損失函式為基礎之 j 節點的事後機率 P 則可寫為

$$\hat{x}_{jMAP}=\arg\max_{[x_j]}\max_{[all\ x_i,i\neq j]}P(x_1,x_2,...,x_N,y_1,y_2,...,y_N)\ 。$$

(1) 無迴圈網路推論

在無迴圈網路中，可採用“信息傳遞(message-passing)”規則來計算 MAP 及 MMSE 的估計值。以圖五為例，節點 x_1 的 MAP 計算式可表示成

$$\begin{aligned}\hat{x}_{1MAP}&=\arg\max_{x_1}\max_{x_2}\max_{x_3}P(x_1,x_2,x_3,y_1,y_2,y_3,)\\&=\arg\max_{x_1}\max_{x_2}\max_{x_3}\Phi(x_1,y_1)\Phi(x_2,y_2)\Phi(x_3,y_3)\Psi(x_1,x_2)\Psi(x_2,x_3)\\&=\arg\max_{x_1}\Phi(x_1,y_1)\max_{x_2}\Psi(x_1,x_2)\Phi(x_2,y_2)\max_{x_3}\Psi(x_2,x_3)\Phi(x_3,y_3).\end{aligned}$$

我們可注意到其中每個子部份的計算僅需要知道該節點與其相鄰的節點資訊，在相鄰節點間進行本地端信息傳遞即可有效地計算出所需的估計值。也就是說，任一節點的估計值可寫成

$$\hat{x}_{jMAP}=\arg\max_{[x_j]}\Phi(x_j,y_j)\prod_k M_j^k\ ,$$

其中 k 代表節點 j 所有的相鄰場景節點，而 M_j^k 是由節點 k 傳遞至節點 j 的信息，也就是

$$M_j^k=\max_{[x_k]}\Psi(x_j,x_k)\Phi(x_k,y_k)\prod_{l\neq j}\tilde{M}_k^l\ 。$$

在文獻上已證明對於網路中的每個節點只需進行最多一次 M_j^k 的全域遞迴計

算，$\hat{x}_{jMAP}$ 即可給定每個節點 j 的最佳(optimal)估計值。

- 例一：在圖五的馬可夫網路中進行全域遞迴計算以求得節點 x_1 的 MAP 估計值。在經過第一次遞迴計算後，各 M_j^k 的值分別如下，

$$M_1^2 = \max_{[x_2]} \Psi(x_1, x_2)\Phi(x_2, y_2) \text{ ，}$$

$$M_2^3 = \max_{[x_3]} \Psi(x_2, x_3)\Phi(x_3, y_3) \text{ ，}$$

$$M_2^1 = \max_{[x_1]} \Psi(x_2, x_1)\Phi(x_1, y_1) \text{ ，}$$

$$M_3^2 = \max_{[x_2]} \Psi(x_3, x_2)\Phi(x_2, y_2) \text{ ，}$$

接著，經過第二次遞迴計算後，各 M_j^k 的值更新如下，

$$M_1^2 = \max_{[x_2]} \Psi(x_1, x_2)\Phi(x_2, y_2)\tilde{M}_2^3 \text{ ，}$$

$$M_2^3 = \max_{[x_3]} \Psi(x_2, x_3)\Phi(x_3, y_3) \text{ ，}$$

$$M_3^2 = \max_{[x_2]} \Psi(x_3, x_2)\Phi(x_2, y_2)\tilde{M}_2^1 \text{ ，}$$

$$M_2^1 = \max_{[x_1]} \Psi(x_2, x_1)\Phi(x_1, y_1) \text{ ，}$$

最後，經過第三次遞迴計算後，即可求得 x_1 的 MAP 估計值為

$$M_1^2 = \max_{[x_2]} \Psi(x_1, x_2)\Phi(x_2, y_2) \max_{[x_3]} \Psi(x_2, x_3)\Phi(x_3, y_3) \text{ ，}$$

$$\hat{x}_{1MAP} = \arg\max_{[x_1]} \Phi(x_1, y_1)M_1^2 \text{ 。}$$

(2) 迴圈網路推論

由於本地端信息傳遞的複雜性，在迴圈網路中無法以上節所提之計算方式求得精確的 MAP 或 MMSE 估計值，需以近似(approximation)技巧加以輔助。根據文獻上的討論，無迴圈與迴圈網路推論的精確性整理如下表。

表一、關於可信度傳遞之收斂(convergence)情況在無迴圈與任意(包含迴圈)網路推論的精確性比較。

Belief propagation algorithm	**Network topology**	
	no loops	*arbitrary topology*
MMSE rules	MMSE, correct posterior marginal probs.	For Gaussians, correct means, wrong covs.
MAP rules	MAP	Local max. of posterior, even for non-Gaussians.

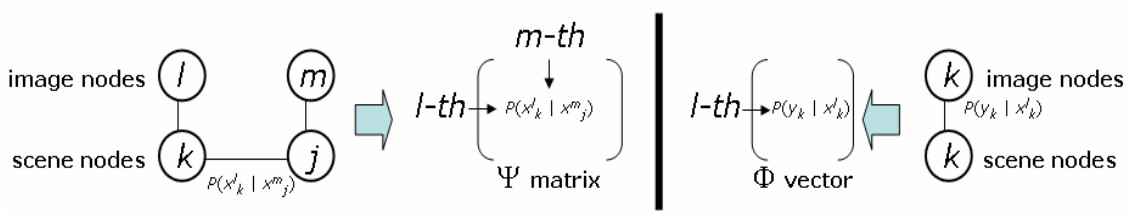
(3) 協調函式學習

在本節中，我們針對兩種較廣為接受的協調函式學習方法加以說明。第一種方法採用兩個信息傳遞規則，分別為

$$M_j^k = \max_{x_k} P(x_k \mid x_j) P(y_k \mid x_k) \prod_{l \neq j} \tilde{M}_k^l \ ,$$

$$x_{jMAP} = \arg\max_{x_j} P(x_j) P(y_j \mid x_j) \prod_k M_j^k \ 。$$

對於場景節點 k 的可能值 x^l_k 與相鄰節點 j 的可能值 x^m_j 所形成的所有機率組 $P(x^l_k|x^m_j)$進行求值運算(參考圖六)，得到相對的協調函式 Ψ 矩陣與 Φ 向量，即可於推論階段中以查表方式得到對應機率。



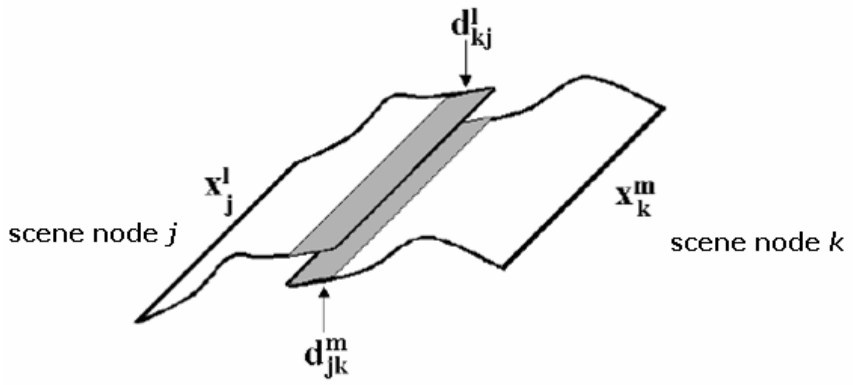
圖六、任意相鄰節點的協調函式 Ψ 矩陣與 Φ 向量在馬可夫網路中的關係示意圖。

第二種方法在分割影像與場景補丁時採用部份重疊(overlapping)的方式，因此協調函式的計算是直接由補丁重疊部份的差益性而得(參考圖七)：

$$\Psi(x_k^l, x_j^m) = \exp^{-|d^l_{jk} - d^m_{kj}|^2 / 2\sigma_s^2} \ ,$$

$$\Phi(x_k^l, y_k) = \exp^{-|y^l_k - y_o|^2 / 2\sigma_i^2} \ ,$$

其中 y_o 為正確的場景，而 σ^2_s 與 σ^2_i 為高斯雜訊的共變異數。

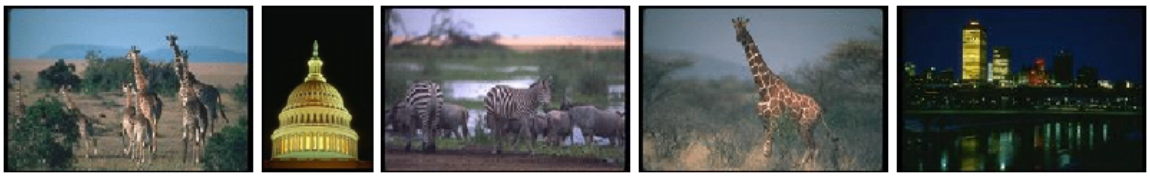


圖七、部份重疊補丁示意圖。

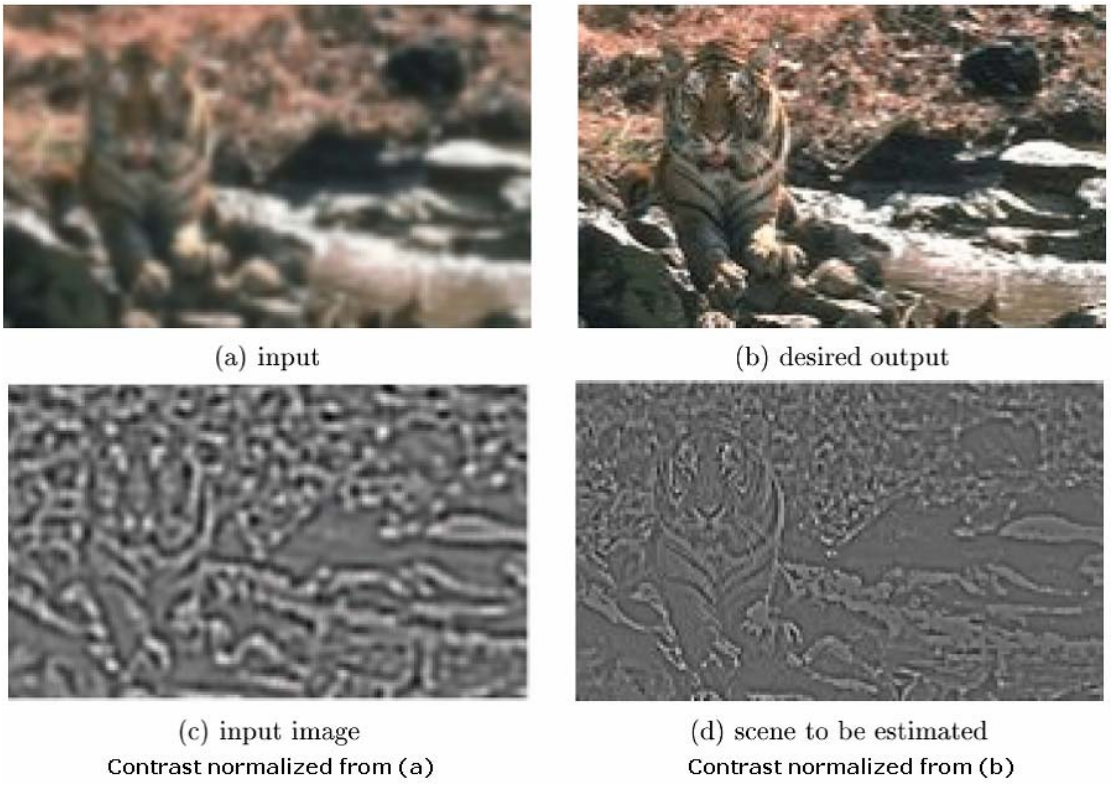
3. 相關應用

(1) 超級解析度

超級解析度指的是當給定一張低解析度(low-resolution, LR)影像時，求得其相對應的高解析度(high-resolution, HR)場景。因此其目標是希望從訓練資料中建立模糊化(blurred)LR 影像與銳利化 HR 場景之間的相對關係(如圖八)。值得注意的是，此處的定義與一般超級解析度問題並不相同。



(A) 部份訓練資料擷圖



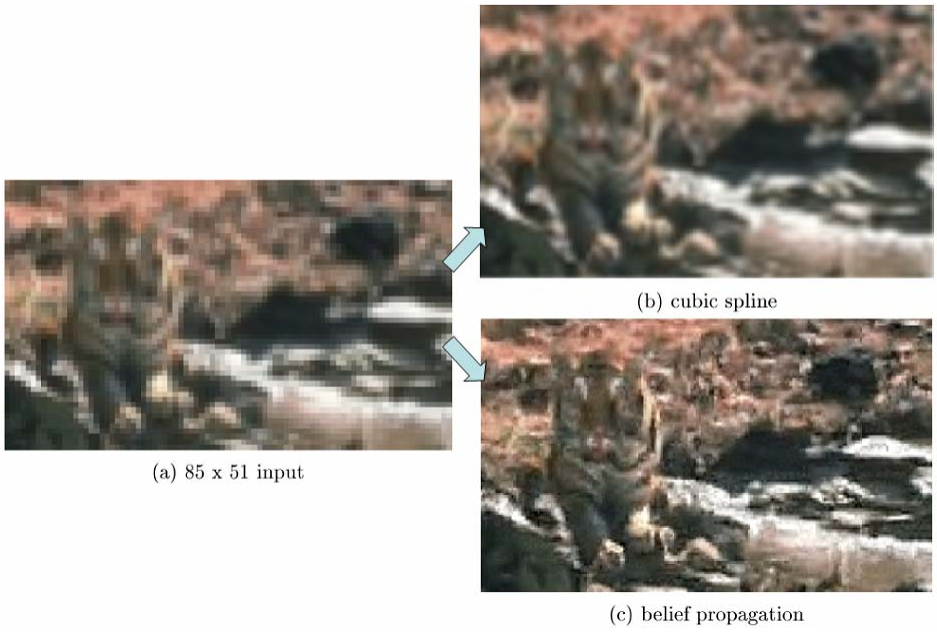
(B) LR 影像/HR 場景與相對應之高頻影像

(C) 實際訓練資料(高頻影像)

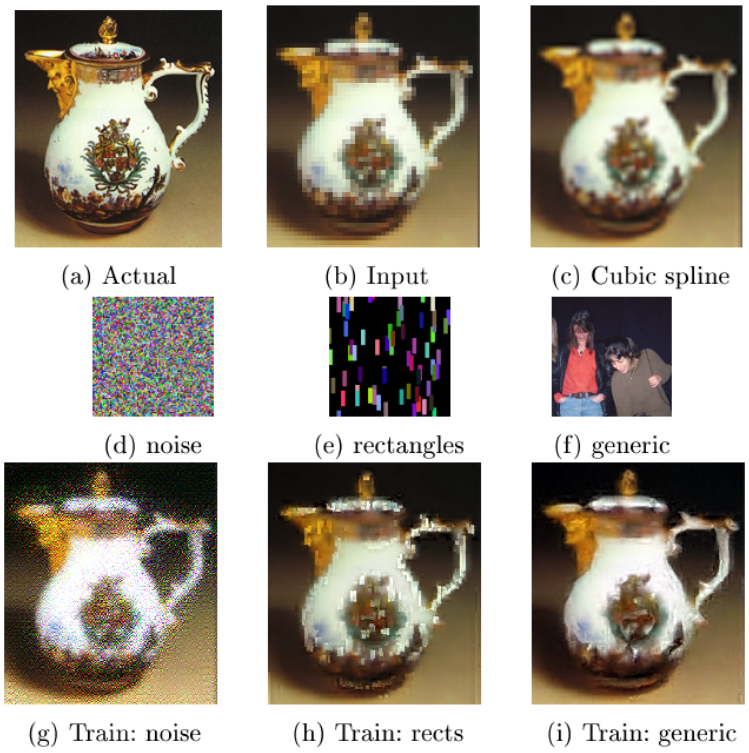
圖八、超級解析度實例。

為簡化問題的討論，我們給定兩個假設，分別為：

- 在給定第二高解析度影像(M)的情況下，最高解析度影像(H)是條件獨立(conditionally independent)於最低解析度影像(L)，也就是說 $P(H|M,L) = P(H|M)$。此假設是為了減少所需儲存的資料量。
- 不同解析度影像間的統計關係是獨立於影像對比性，與縮放倍數無關。此假設是為了減少影像對比可能值的資料量。

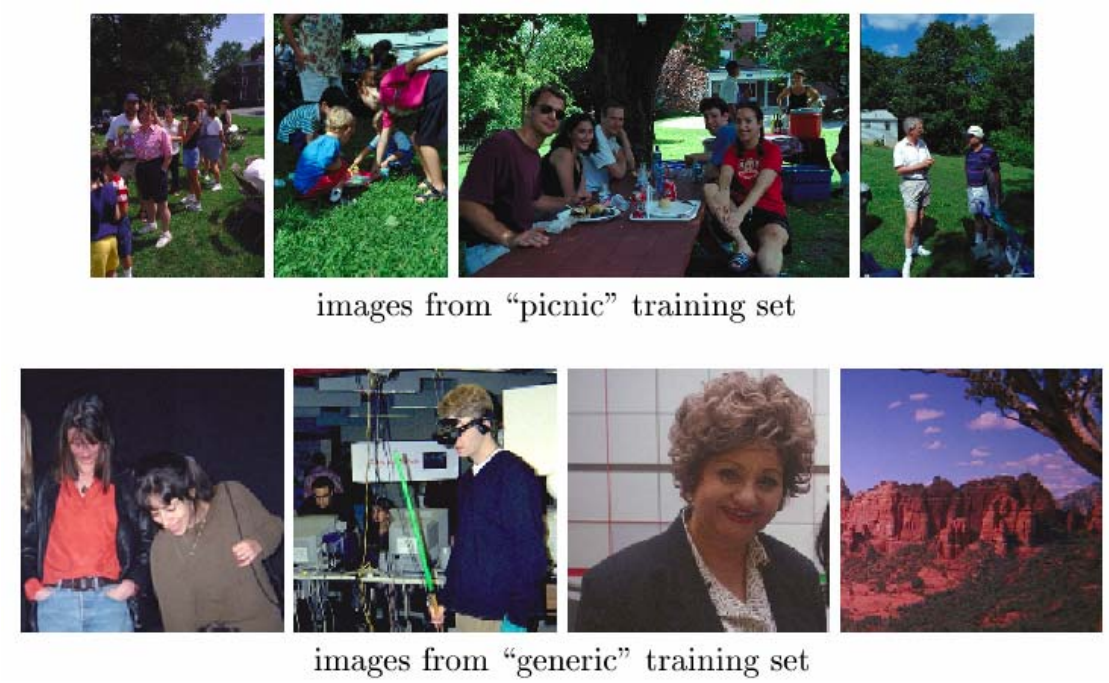


圖九、對(a)相同影像(b)直接內插放大與(c)可信度傳遞實驗結果比較。

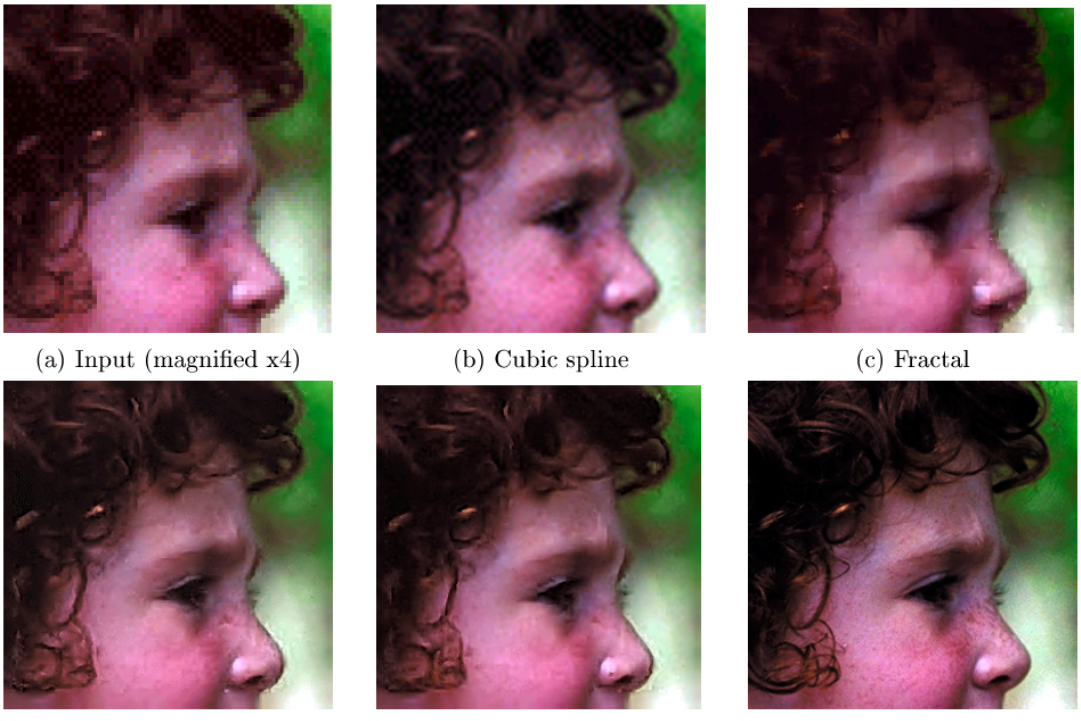


圖十、以不同雜訊對訓練資料進行干擾的實驗結果。

圖九顯示將 LR 影像經用直接內插放大(cubic-spline)與經由可信度傳遞所得結果的比較。圖十的實驗則顯示當以不同雜訊對訓練資料進行干擾的影響。清楚地，經由不同干擾的訓練資料在對應的實驗結果中會產生與雜訊型態相同的視覺缺損。圖十一與圖十二為另外一個實例。



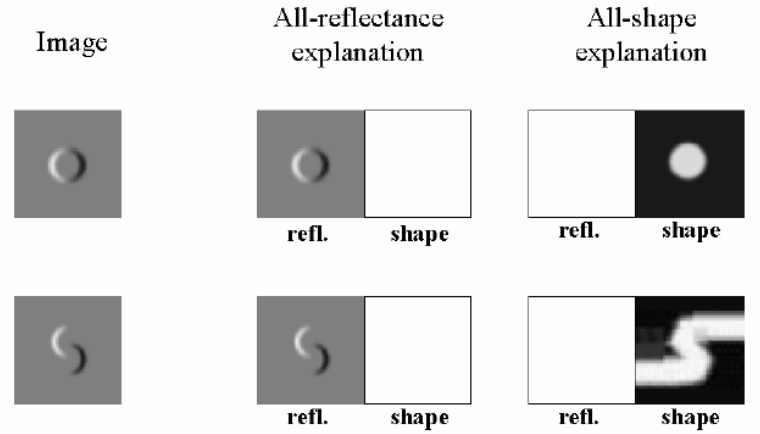
圖十一、兩個以自然影像所建立的訓練資料組。



圖十二、以不同方法所產生的超級解析度實驗結果。

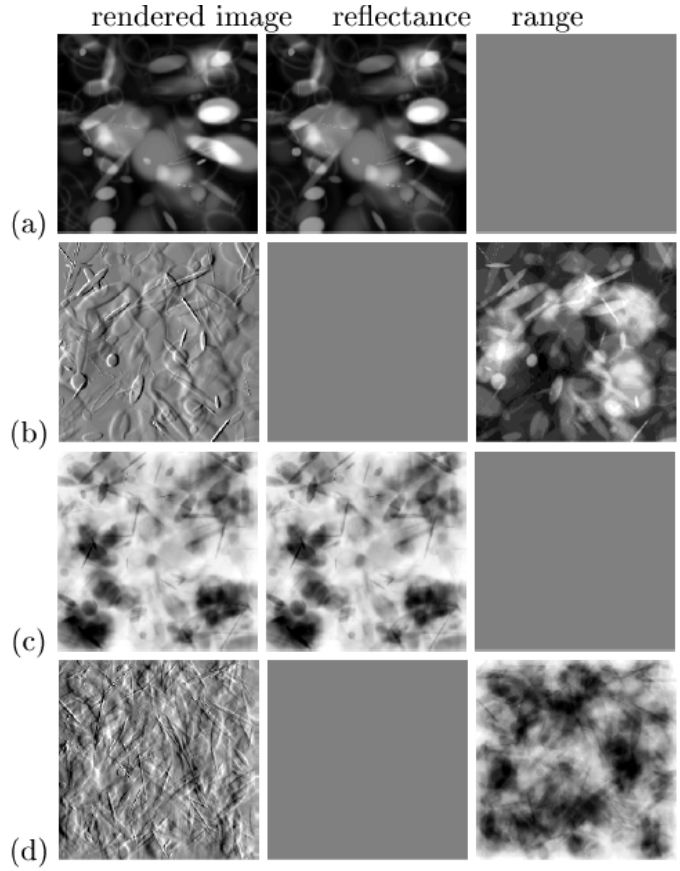
(2) 遮光與反射估計

遮光與反射估計是由給定的一張影像中求出對應的遮光與反射特性，如圖十三。

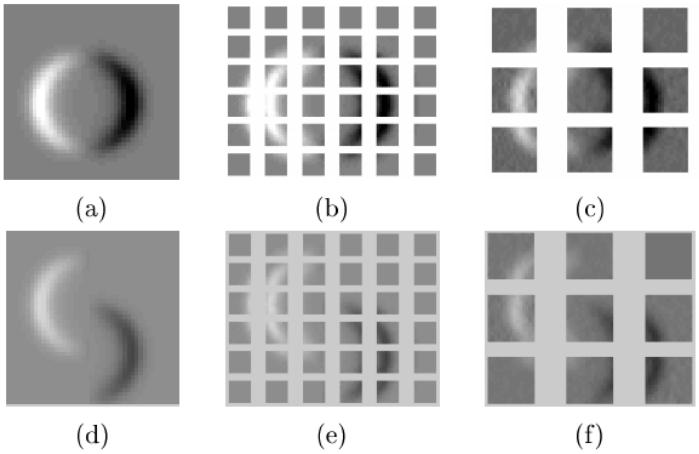


圖十三、遮光與反射估計實例。

為簡化問題的討論，我們假設只有單一方向的光源存在，並且反射函式為恆定。首先，產生訓練資料的影像及場景，其中每個場景包含兩個像素陣列(pixel array)，一個用以表示反射函式，而另一個則是用以表示形狀，稱為範圍圖(range map)，範圍圖中的像素強度(值)表示該點距離攝影機的遠近。參考圖十四。



圖十四、訓練資料實例。

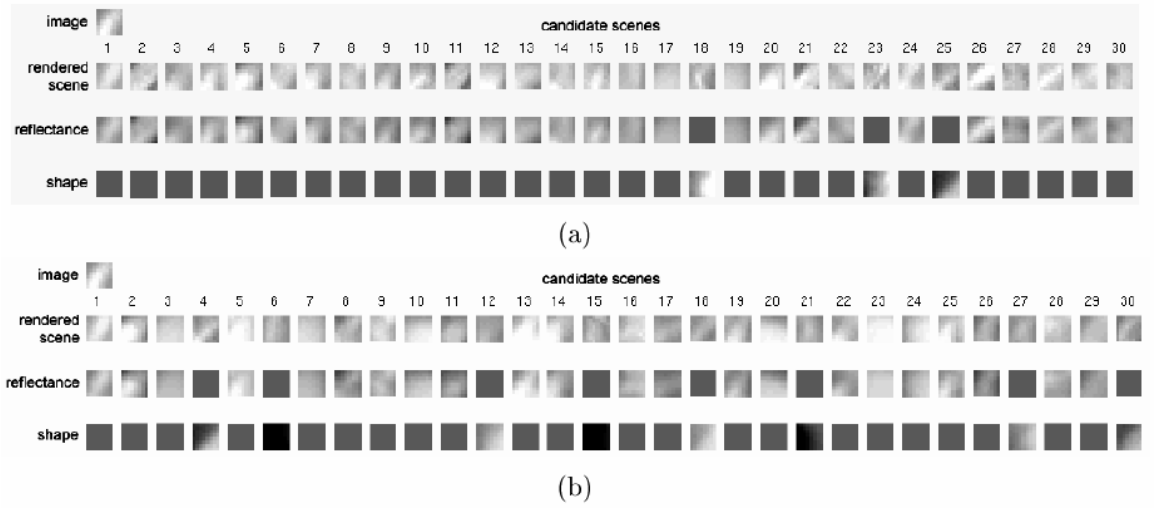


圖十五、訓練資料補丁。

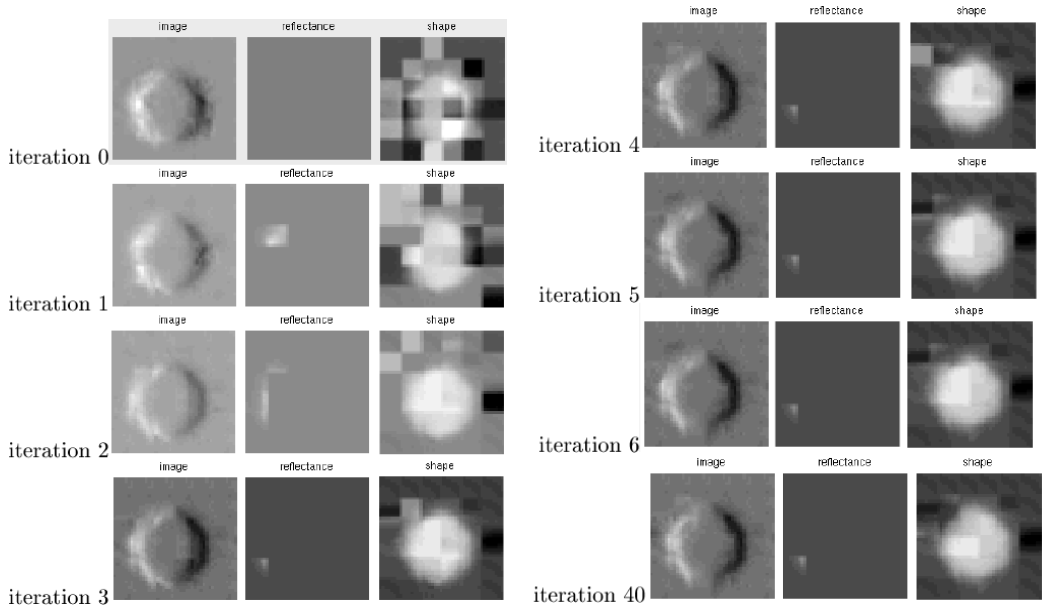
接著，將訓練資料分割成補丁，如圖十五。對每一個影像補丁，由訓練資料中選擇一組可能的對應場景。所選擇的場景組必須要有足夠的廣泛性，也就是說該組中的每個場景之間要具高度差異性。因此，我們定義並最大化一個選擇函式如下：

$$\hat{P}(S) = \max_{x_i^j \in S} P(|\hat{x}_i - x_i^j| < \varepsilon \mid y_i) \text{ ，}$$

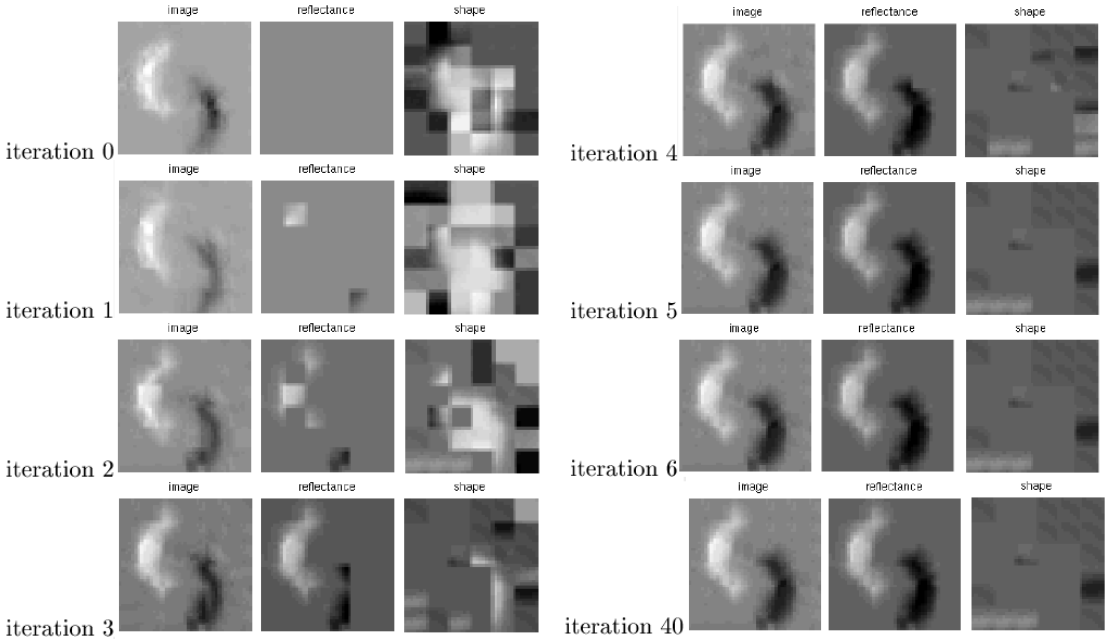
其中 S 為場景組，$\hat{x}_i$ 及 x_i^j 分別為正確的與可能的場景。而 S 的選擇可以使用貪婪演算法(greedy approach)來獲得，並對場景組中的每一個可能場景計算出一個利用值(utility value)。最後，圖十七與圖十八分別為三組利用可信度傳遞所得到不同的實驗結果。



圖十六、反射與形狀場景組實例。



圖十七、經由 40 次遞迴計算所得到的實驗結果。

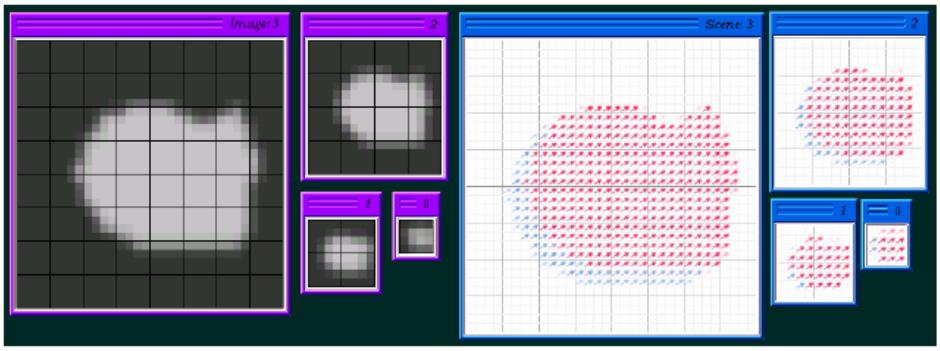


圖十八、經由 40 次遞迴計算所得到的實驗結果。

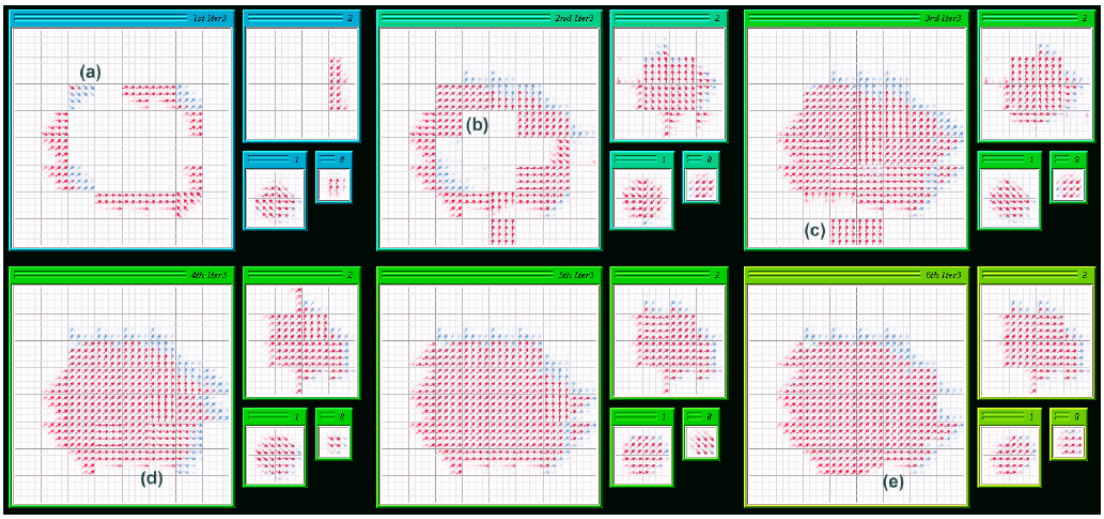
(3) 移動估計

在移動估計問題中，所欲求得的場景參數是移動物體的投影速度(projected velocity)，而影像資料是兩張連續的影格(frame)，如圖十九。同樣地，利用馬可夫網路與可信度傳遞估計，可算出影像中的物體移動速度，參考圖二十。值得注意的是，在圖二十中剛開始時所求得移動量的一致性較低，甚至呈現矛盾的現象，這是在移動估計問題中著名的“孔徑問題(aperture problem)”。也就是說，同質輪廓(homogeneous contour)的移動會有局部難以辨識的情況，這是由於物體的移動是由有限的視覺接受區域(receptive field)所觀察到的，就像是我們“以管窺

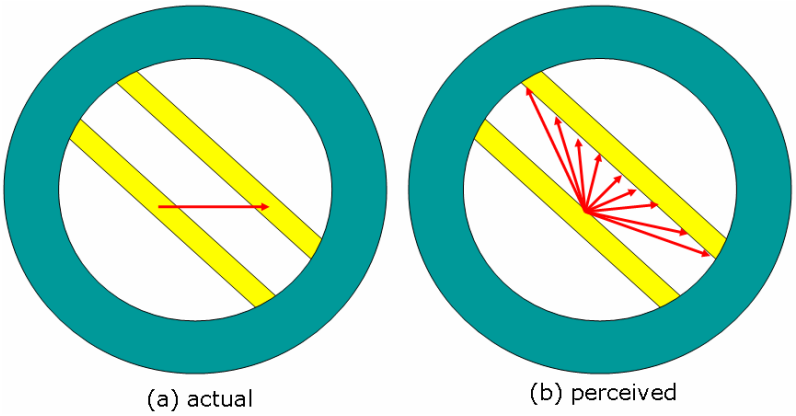
天”無法正確瞭解其全貌。以圖二十一為例，我們由孔徑中觀察由左向右水平移動的方形長棍(圖二十一(a))，由於只看到物體的局部，因此每個人或自己每次所感覺到的移動方向均不同(圖二十一(b))。



圖十九、左為原始影格，右為相對應之兩張連續影格中所求得之物體投影速度(移動前後之物體位置分別以紅色與藍色像素表示)。



圖二十、經由可信度遞迴計算所得到的實驗結果。



圖二十一、孔徑問題示意圖。

4. 結論

本篇文章初步介紹了低階視覺學習問題與相關文獻上所提出的解決方法，其中以 VISTA 架構為例子，它顯示了應用機器學習理論於多媒體資料上以解決視覺理解(visual interpretation)問題的優點。而 VISTA 架構的一個缺點在於給定任意的影像資料後，必須能找到或決定對應的場景資訊，如此才能順利導入馬可夫網路中進行自動學習。因此，如何將給定的問題定性化與定量化，並有效取得所須的訓練資料，將是低階視覺學習問題的一大挑戰。

5. 參考資料

本簡介之文字敘述與圖表摘自或參考以下參考資料：

[1] W. T. Freeman, E. C. Pasztor, and O. T. Carmichael, “Learning low-level vision,”*IJCV*. 2000.

[2] W. T. Freeman, T. R. Jones, and E. Pasztor, “Example-based super-resolution,”*IEEE CGA*. 2002.

[3] J. S. Yedidia, W. T. Freeman, and Y. Weiss, “Understanding belief propagation and its generalizations,”*IJCAI*. 2001.

[4] M. F. Tappen and W. T. Freeman, “Comparison of graph cuts with believe propagation for stereo, using identical MRF parameters,”*ICCV*. 2003.

[5] B. J. Frey and N. Jojic, “A comparison of algorithms for inference and learning in probabilistic graphical models,”*IEEE PAMI*. 2005.

[6] http://research.microsoft.com/adapt/MSBNx/