

感知



本章中，我們把電腦與原始的、未經修飾的世界聯繫起來。

感知透過解釋感測器的回應，而提供了代理人其所處的世界之相關資訊。感測器量測環境的某面向時，其形式可供代理人程式作為輸入。這個感測器可簡單如一開關，其利用一位元判別是否處於開啟或關閉；或者複雜如人眼。目前的人工代理人已可以使用許多不同的感測模態。其中與人類共有的包括視覺、聽覺和觸覺。無輔助之人類不能獲得的感測模態包括如，無線電波、紅外線、GPS 以及無線訊號。有些機器人會進行**主動感測**(active sensing)，也就是說它們發射出一個信號，例如雷達信號或超音波，然後感覺從環境中反射回來的信號。我們並不想要討論所有的感測方式，本章中將深入探討其中一個模態：視覺。

我們看到在 POMDPs 的描述(第 17.4 節)，在部份可觀察環境中的一個基於模型的理論決策代理人，其具有一個**感測器模型**——給定世界一狀態下，在感測器所提供的證據上的機率分布 $\mathbf{P}(E | S)$ 。貝式公式可以用來更新對狀態之估計。

對於視覺，感測器模型可以被拆成兩個部份：一個**物體模型**描述了存在於可視世界中的物體——人們、建築物、樹木、車輛等。物體模型可以包含一個取自電腦輔助設計(computer-aided design, CAD)系統的精確 3D 幾何模型，或可為含糊的限制，例如人眼間距通常為 5 到 7 公分的這個事實。一個**渲染模型**(rendering model)描述的則是，來自世界、會產生刺激的物理、幾何、及統計過程。渲染模型相當準確，但是有含糊部份。例如，在弱光下的白色物體看起來顏色可能像在強光下的黑色物體。一個較近的小物體，可能看起來會和較遠的大物體一樣大。沒有額外的證據下，我們無法分辨螢幕上的影像，是一個酷斯拉怪獸玩具，還是一隻真正的怪獸。

這種含糊部份可利用先驗知識來加以掌握——我們知道酷斯拉不是真的，因此這個影像一定是玩具——或選擇性地決定忽略這個含糊部份。例如，自動駕駛車輛的視覺系統可能不能解讀出距離很遠之物體，但是代理人可以選擇忽略這個問題，因為不太可能會撞上距離數英哩遠的物體。

理論決策代理人並不是使用視覺感測器的唯一架構。舉例來說，果蠅(*Drosophila*)某種程度上為反射代理人：牠們的頸椎巨神經束，從牠們的視覺系統到翅膀肌肉形成一個直接的路徑，以便產生一個逃脫的反應——這是一個立即的反應，不會有考慮行為。蒼蠅和許多會飛行的動物都使用封閉迴圈控制架構，來停留在物體上。視覺系統擷取與該物體的距離估計資訊，而控制系統則據以調整翅膀肌肉，從而允許飛行方向能快速變化，且不需要一個物體的詳細模型。

相較於其他感測器(例如一個吸塵器機器人是否撞到牆壁的單一位元資料)所得到的資料，視覺觀察資料則是非常之豐富，不論是能顯示的細節或是其產生的資料絕對數量。對於機器人所使用的攝影機，以 60 Hz 產生一百萬個 24 位元的像素資料；每分鐘 10 GB 的速度。因此對於一個有視覺能力的代理人，其問題變成是：在這麼多的視覺刺激中，哪一部份應該考慮來協助代理人做出好的行動選擇，而那個部份應該忽略？視覺——以及所有的感知——是用以加強代理人的目標，而非作為本身的目的。

我們可將這個問題的解決方式分成粗略三類。特徵擷取(feature extraction)的方法，如同果蠅身上所展現的，其強調的是直接應用於感測器觀察結果的簡單計算。在辨別(recognition)的方法中，代理人利用視覺及其它資訊，來在遭遇之物體間找出區別。辨別可指對每個影像標上是或否，即是否包含我們應尋找的食物，或是包含祖母的臉龐。最後，在重建(reconstruction)的方法中，代理人將會從一張影像或是一組影像來建立實際世界的地域性模型。

最近三十年的研究已經產出了許多強力工具與方法來處理這些方法。要了解這些方法需要了解影像藉以形成的處理過程。因此，我們現在將說明，影像生成時發生的物理與統計現象。

24.1 成像

影像會使物體外貌失真。比方說，一張從一端看鐵軌的照片，可能讓人會以為鐵軌最後會合在一起。另一個例子是，如果你將自己的手擋在眼睛前面，你可以擋住月亮，但實際上月亮遠大於你的手。當你前後移動手掌或是將手掌傾斜，你的手呈現的影像看起來會放大或是縮小，但是實際上並不是這樣(圖 24.1)。這些效應的模型對於辨別與重建都非常重要。

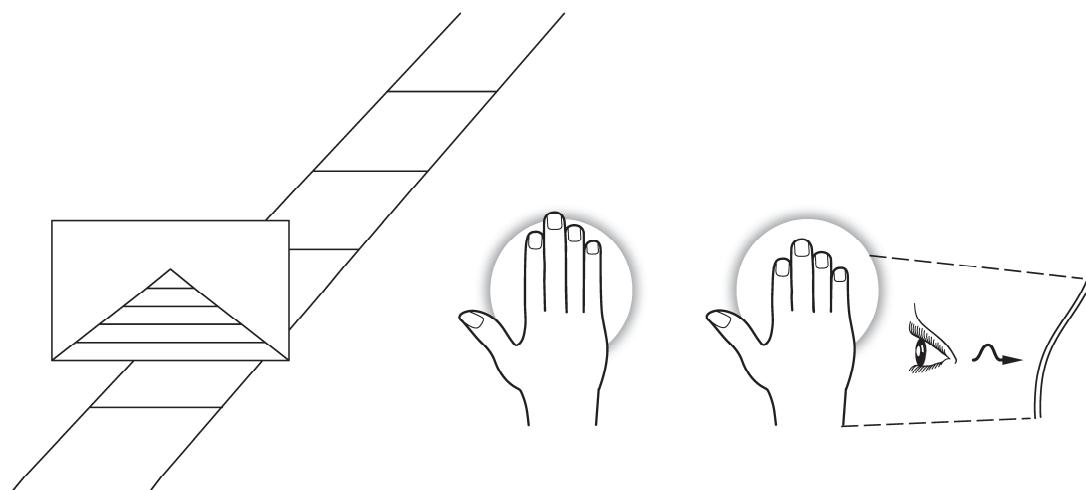


圖 24.1 影像使幾何關係失真。平行線看起來可能會在遠處相交，就如同在圖中左側的鐵路影像。圖中央的小手掌可以擋住大部份的月亮。圖右則是一個透視縮短效應：當手擺離眼睛時，看起來會比中間的圖要短

24.1.1 無透鏡影像：針孔照相機

影像感測器收集在一場景或環境中從物體散射出來的光線，並且產生一個二維影像。在眼睛當中，這個影像會成像於由兩種細胞所組成的視網膜處這兩種細胞包括對於各個寬廣波長光線很敏感的一億個桿狀細胞，以及五百萬個錐狀細胞。錐狀細胞主要用於辨別色彩，其有三種主要的形式，每一種形式會對應到特定波長。在照相機中，影像會成像於一平面上，此平面可以是帶有鹵化銀的薄底片，或者是一個小方塊上帶有數以百萬計的感光像素，其可以是互補金氧半導體(complementary metal-oxide semiconductor, CMOS)或者是感光耦合元件(charge-coupled device, CCD)。當每個光子抵達感測器時會產生一個光電效應，其強度由光子的波長決定。感測器的輸出，是來自某時段中所有光子的總效應，這表示影像感測器所感測到的是抵達感測器的光平均強度。

要看到一個聚焦影像，我們首先要確定所有光子大約是來自於環境中同一點，並且大約抵達影像平面上的同一點。最簡單的成像方法，莫過於使用**針孔照相機**(pinhole camera)，它的組成包括一個盒子，其前部的一個能透光的針孔 O ，以及盒子後部的影像平面(圖 24.2)。來自場景中的光子必須通過針孔，因此若針孔夠小，那麼場景中在附近的光子將會聚集在影像平面附近，而影像就會聚焦。

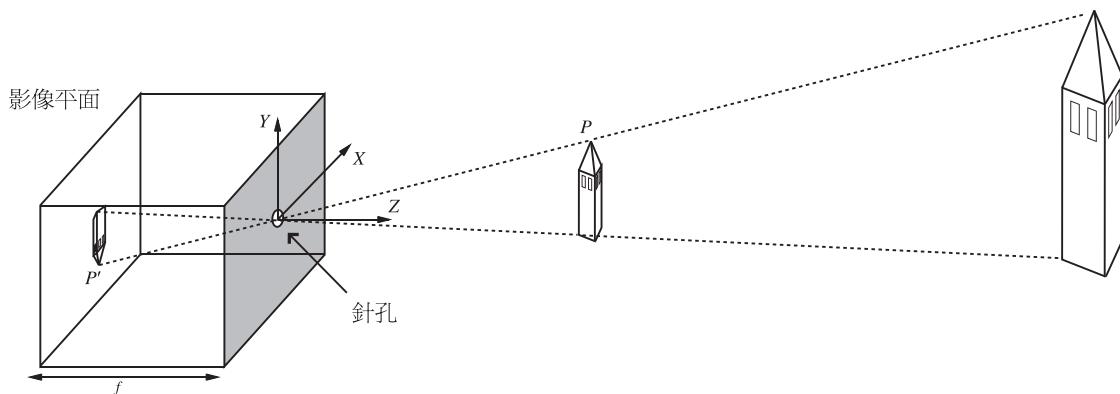


圖 24.2 在針孔攝影機後方的影像平面上，每個影像光敏元件接受到穿過針孔的一小部分方向的光線。若針孔夠小，其結果是在針孔後方的一個聚焦影像。投影的過程表示一個很遠很大的物體，看起來會和較近較小的物體一樣。注意到，影像被投影時是上下顛倒的。

針孔攝影機內的環境與影像的相對位置是最容易了解的。我們將採用一個以針孔為原點的三維坐標系，並考慮場景中的一點 P ，其座標為 (X, Y, Z) 。 P 被投影到影像平面上的點 P' ，座標為 (x, y, z) 。設 f 是從針孔到影像平面的距離，那麼根據相似三角形的性質，我們得到以下公式：

$$\frac{-x}{f} = \frac{X}{Z}, \frac{-y}{f} = \frac{Y}{Z} \Rightarrow x = \frac{-fX}{Z}, y = \frac{-fY}{Z}$$

這些公式定義了一個成像過程，稱為**透視投影**(perspective projection)。值得注意的是，分母上的 Z 意味著物體離得越遠，它的影像越小。還要注意到負號表示影像相對於實際場景是上下、左右顛倒的。

在透視投影的情況下，距離很遠的物體看起來會小。這也就是為什麼你可以用手掌把月亮遮住(圖 24.1)。這個效應的另一個重要結果是兩平行線最後會相交於地平線上一點。(考慮鐵軌的樣子，圖 24.1)。在場景中通過點 (X_0, Y_0, Z_0) ，且方向為 (U, V, W) 的一條直線可被描述為點集合 $(X_0 + \lambda U, Y_0 + \lambda V, Z_0 + \lambda W)$ ，其中 λ 在 $-\infty$ 和 $+\infty$ 之間變化。選擇不同的 (X_0, Y_0, Z_0) 點，將可以使不同的線平行於另外一條。這條直線上的一點 P_λ 到影像平面上的投影由下式給出：

$$\left(f \frac{X_0 + \lambda U}{Z_0 + \lambda W}, f \frac{Y_0 + \lambda V}{Z_0 + \lambda W} \right)$$

當 $\lambda \rightarrow \infty$ 或 $\lambda \rightarrow -\infty$ ，上式將會變成 $p_\infty = (fU/W, fV/W)$ ，當 $W \neq 0$ 。這表示從不同點離開的兩條平行線，在影像中會合在同一點——對於較大的 λ 值，影像點非常靠近於 (X_0, Y_0, Z_0) 點(再次考慮鐵軌的樣子，圖 24.1)。我們稱 p_∞ 為與方向為 (U, V, W) 的直線族相關聯的消失點(vanishing point)。方向相同的直線具有同一個消失點。

24.1.2 透鏡系統

針孔攝影機的缺點是我們需要有一個針孔讓影像聚焦。但是針孔越小，穿過針孔的光子就越少，表示影像會很暗。我們可以透過將延長針孔開啓時間以收集到更多光子，但是我們會讓影像動態模糊——當環境中的物體在移動時，因其會在不同位置傳送光子到影像平面上，因此會看起來很模糊。若我們不能讓針孔開啓的時間更長，我們可以使它看起來大一些。更多光線將會進入，但是從場景中小物體來的光線將會在影像平面上佈成一塊，造成一個模糊的影像。

脊椎動物的眼睛以及現代攝影機會使用透鏡系統來收集足夠光線，以便使影像聚焦。一個帶有透鏡的較大開口，可以將物體附近所發出的光線，聚焦在影像平面的某個區域。然而，透鏡系統可能會有景深的限制：他們僅能聚焦某些從固定深度來的光線(主要是來自聚焦平面)。在此範圍之外的物體將無法在平面上聚焦。為了要移動聚焦平面，可以透過改變眼睛中的透鏡；而在相機中，透鏡可以前後移動(圖 24.3)。

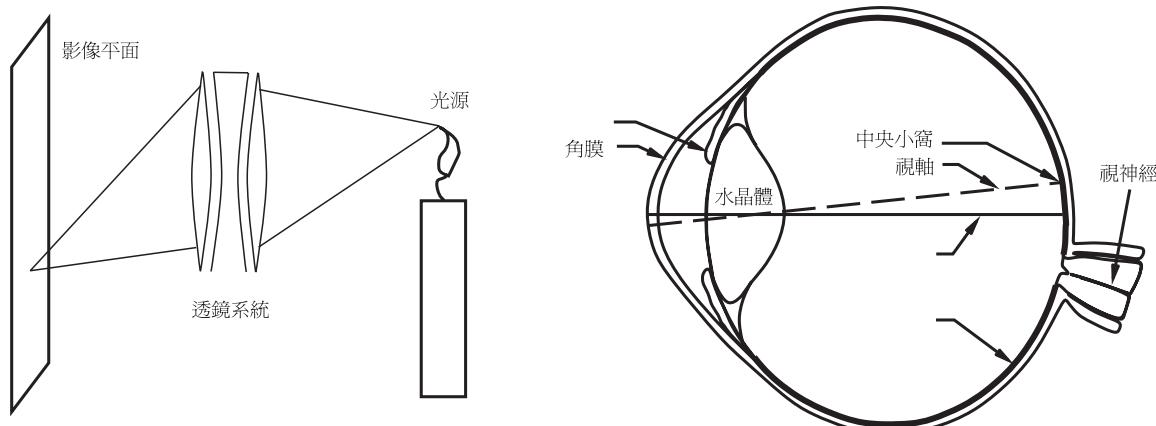


圖 24.3 透鏡收集離開場景某點所發射出的一方向範圍內的光線，並且將這些光線全部引導到影像平面上的單一點。聚焦對於空間中在聚焦平面附近的點有效；其他點則不會被正確聚焦。在照相機中，透鏡系統的元件可作移動而改變聚焦平面，而在眼睛中，透鏡形狀可以透過專用肌肉群而改變

24.1.3 縮放正交投影

透視效應並不總是很明顯。比方說，在很遠地方的美洲豹身上的斑點可能看起來會很小，因為美洲豹是在很遠的地方，但是兩個相鄰的斑點應該會有相同的大小。這是因為斑點之距離差異，相較於我們觀察它的距離是很小的，因此我們可以簡化投影模型。合適的模型是**縮放正交投影**。其想法如下：設物體上點的深度 Z 在某個範圍 $Z_0 \pm \Delta Z$ 內變化，其中 $\Delta Z \ll Z_0$ ，則透視比例因數 f/Z 可以近似為一個常數 $s = f/Z_0$ 。從場景座標 (X, Y, Z) 到影像平面的投影公式變成 $x = sX$ 以及 $y = sY$ 。成縮放正交投影是一個近似，且僅當場景中的物體沒有太大的內部景深差異時可以使用。比方說，縮放正交投影對於遙遠建築前端的特徵，可以作為很好的模型。

24.1.4 光與影

影像當中某個像素的亮度，是由場景中某個表面部份的亮度函數，投影到像素上而決定的。我們會假定其是一個線性模型(目前照相機在極亮與極暗處會使用非線性模型，但在中間值部分大多數是線性)。影像亮度是一個(若模糊的來說)很強的物體形狀暗示，並以此得到物體識別。人們通常區分不同亮度的三個主因，並且將它們重建而得到物體的性質。第一個原因是**整體光強度**。即使一個全白的物體放在陰影之下，其有可能比在直射的太陽光下的黑色物體看起來要暗淡，眼睛可以清楚分辨相對亮度，並且感知白色物體是白色的。第二，場景中的不同點可能會**反射**不同程度光線。通常，其結果是人類會感知到這些點是較亮或是較暗，並且了解物體的紋理或標記。第三，面光的表面區域通常會比側光或背光面較亮，而後者便是我們所知的**陰影**。一般來說，人類可以分辨來自於物體形狀的陰影，但是有時候陰影和標記會混淆。比方說，在顴骨下方畫上深色的線條，通常會看起來像是陰影，而使得讓臉看起來比較瘦。

大部份的表面是透過**漫射反射**的過程而反射光線。漫射反射是將光線散射，平均地橫跨離開表面的所有方向，所以散射表面的亮度並取決於所視方向。大多數衣物、圖畫、粗糙木質表面、植被、以及粗糙石塊的表面均是漫射。鏡子則不是漫射，因為你看鏡子的角度將會決定你在鏡中會看到什麼。一個理想鏡面的行為稱之為**鏡面反射**。某些表面——例如刷過的金屬、塑膠、或者潮濕地板——某些小部分會產生鏡面反射，稱為**鏡面反射性**。這很容易區分出來，因為這些區域通常很小同時很明亮(圖 24.4)。對於幾乎所有用途，這足以將所有表面模擬為具有鏡面反射的漫射。

在室外，照明的光源大多數是太陽，其所發射出來的光線都是平行光。我們將此情況，用一**遠處點光源**模型來代表。這是照明最重要的模型，而且不管是室內或室外場景都相當適合。在此模型中由某塊表面所能收集到的光線，是由光線照射方向以及此表面的法向量之夾角 θ 所決定。

某個漫射表面被遠處點光源所照射，其會反射它所收集到的部份光線；反射的比例稱之為**反照率**(albedo)。白紙和雪擁有高反照率，大約在 0.90 左右，而黑色天鵝絨與木炭的反照率就低到大約 0.05(這表示 95% 的入射光均被天鵝絨的纖維或是木炭內的孔所吸收)。**蘭式餘弦定律**描述了某塊漫射面的亮度可由下式表示

$$I = \rho I_0 \cos\theta$$

其中 ρ 為漫射反照率， I_0 是光源強度， θ 是入射光方向與表面法向量之夾角(參考圖 24.5)。蘭式餘弦定律預測了影像中的亮點來自某塊直接面對光源的表面，而暗點則是來自正切於光線的表面，因此陰影可以提供我們一些有關形狀的資訊。我們將在第 24.4.5 節中更詳加討論此議題。若某塊表面沒有被光源照到，那麼它就會是在陰影下。陰影通常不會是均勻的黑，因為陰影區通常還是會接收到來自其他光源的光線。在室外，這種光源最重要的來源是明亮的天空。在室內，來自其他表面的反射光可以照亮陰影區塊。這些交互反射會在其他表面的亮度上產生很顯著的效應。這些效應有時候可以用以下的模型來說明，藉著在預測亮度上加上一個固定環境照明度。

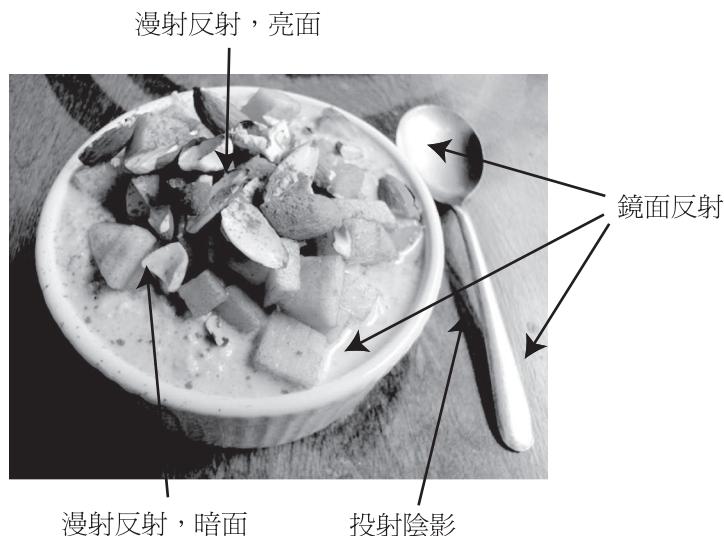


圖 24.4 各種照明效果。在金屬湯匙與牛奶表面有鏡面反射產生。較亮的漫射表面是因為它面向光源方向而比較亮。而較暗的漫射表面是因為切於照明方向而較暗。陰影出現在無法看到光源的表面點。照片出處：Mike Linksvayer(mlinksva on flickr)

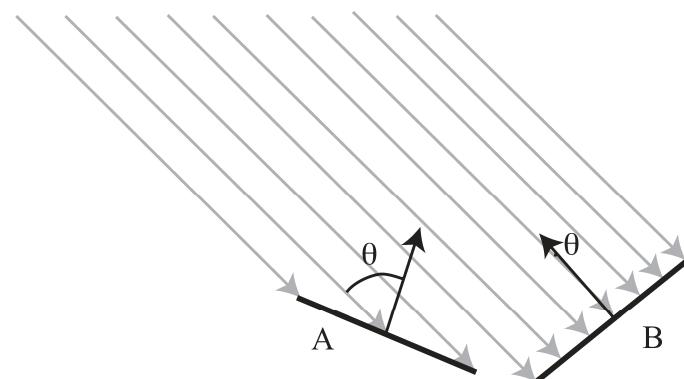


圖 24.5 兩塊平面均受到一遠處的點光源照射，其照射光線以灰色箭頭表示。A 區域傾斜於光源(θ 接近 90 度)且接收較少能量，因為其每單位面積上能接受到的光線較少。B 區域直接面對光源(θ 接近 0 度)，可以收集到較多能量

24.1.5 色彩

水果或果實是植物用來吸引動物協助其傳播種子的一種賄賂。樹木的果實成熟時，其會轉成紅色或黃色，而動物也演化成能夠偵測這些顏色的變化。抵達眼睛不同波長的光帶有不同的能量；這可以用一個頻譜能量密度函數來表示。人類眼睛可以感應到 380-750 nm 波長範圍的光線，其由三種不同的色彩接收細胞來感應，其峰值分別為 420 nm(藍色)，540 nm(綠色)，以及 570 nm(紅色)。人類眼睛僅能夠抓到全光譜能量密度函數中的一小部份——但是這已經足夠去區分水果是否已成熟。

三原色原理說明了對於任何頻譜能量密度，不論多麼複雜，都可以用三原色——通常是紅，綠，藍——的混合頻譜密度函數來組成，即使人類無法分辨出兩者的差異。這表示我們的電視以及電腦顯示器僅需要三種顏色(紅黃藍，R/G/B)元件即可。這也讓我們所討論的電腦視覺演算法更簡單了。每個表面可以用三個不同紅綠藍的反射率來建立模型。相同的，每個光源也可以用三個紅綠藍的強度來模擬。我們可以應用蘭式餘弦定律來得到三個黃綠藍像素值。這個模式也能正確的預測，相同表面在不同顏色光源下會產生不同的彩色區塊。事實上，人類觀察者相當擅於忽略不同色光的效應，並且能夠在白光下估測表面顏色，而這個效應稱之為**色彩一致**(color constancy)。現在已經有確實準確的顏色一致性演算法；你可以從你的照相機中找到一個簡單版本，其被稱為「自動白平衡」。注意如果我們想要用照相機來模擬一隻螳螂蝦，我們將需要 12 個不同顏色像素，來對應甲殼動物不同的 12 種顏色接受器。

24.2 初級影像處理運算

我們已經看到，光線是如何被場景中的物體反射，並形成一個由比如 500 萬個 3 位元像素所組成的影像。使用任何感測器，影像中都會有雜訊，此外在任何情況下都需要處理大量的資料。所以我們要如何開始分析這些資料？

在本節中我們將會研究三個有用的影像處理程序：邊緣檢測，紋理分析，以及光流計算。這些作法由於在一連串操作中的最前面，因此稱為「早期」或是「低階」操作。初級視覺運算的特徵是具有局部本質(它們可以在影像的某個部分上實行，而不必考慮在若干個像素以外的情況)和不需要知識：我們可以不需要考慮這些物體是否存在於場景當中來實行這些演算法。這使得低層次運算成為在平行處理的硬體中實作的不錯選擇——在圖像處理器(graphic processor unit, GPU)或眼睛。隨後我們會看到一個中階運算：將影像分割成數個區域。

24.2.1 邊緣檢測

邊緣(edge)是影像中的直線或曲線段，在它們附近的影像亮度有「顯著的」變化。邊緣檢測的目標是根據大量的、成百萬位元組的影像資料進行抽象，形成更緊湊、更概括的表示方式，如圖 24.6 中所示。這樣做的動機在於，影像中的邊緣輪廓與重要的場景輪廓相對應。在圖中，我們有三個深度不連續的例子，標為 1；二個表面法線不連續的例子，標為 2；一個反射不連續的例子，標為 3；一個亮度不連續(陰影)的例子，標為 4。邊緣檢測只關心影像，因此不區分場景中這些不同種類的不連續，不過後面的處理將進行區分。

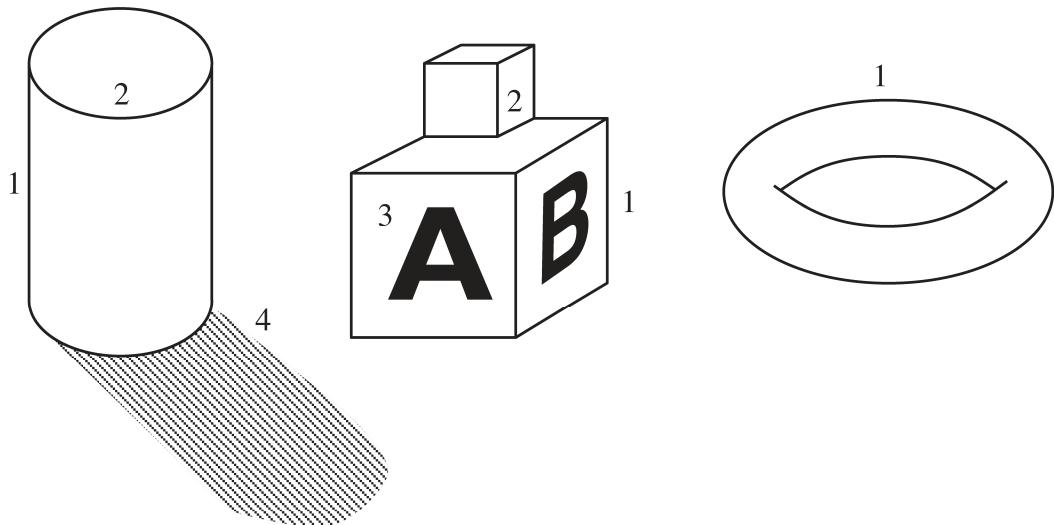


圖 24.6 不同類型的邊緣：(1) 深度不連續；(2) 表面方向不連續；(3) 反射不連續；(4) 亮度不連續(陰影)

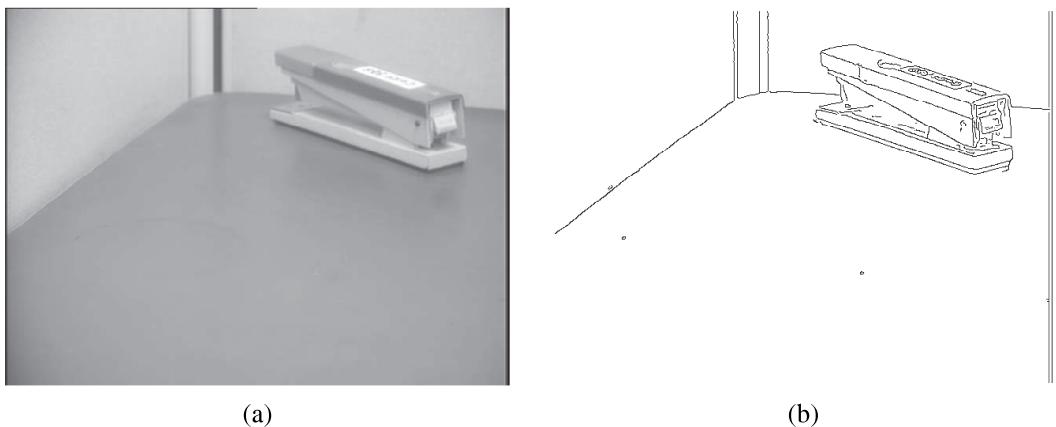


圖 24.7 (a) 一個訂書器的照片。(b) 根據(a)計算得到的邊緣

圖 24.7(a)顯示了場景中包含一個放在書桌上的訂書器的一幅影像，而(b)顯示了一個邊緣檢測演算法在該影像上的輸出。可以看到，這個輸出與理想的線條圖之間是有差異的。當沒有邊緣產生時，將會有缺口產生，而且有時候會有具有雜訊的邊緣產生，其並沒有對應到場景中任何顯著的物體。在後面的階段中將不得不改正這些錯誤。

我們如何在一幅影像中進行邊緣檢測？考慮沿著垂直於一條邊緣的一維截面的影像亮度曲線圖——例如，桌面的左邊緣和牆之間的那條邊緣線。這種曲線看起來如圖 24.8(上圖)所示。

因為邊緣對應著影像中亮度值發生劇烈變化的位置，所以一個天真的想法就是對影像進行梯度運算，然後尋找導數 $I'(x)$ 量級較大的位置。這個方法幾乎可行。在圖 24.8 中圖內，我們可以看到在 $x = 50$ 處確實有一個峰值，但是在其他位置(如 $x = 75$ 處)也會有其他峰值產生。導致這種情況出現的原因是影像中有雜訊。假設我們先將影像柔化，那麼這些偽峰值就會消失，就如同我們在最下圖中所看的一樣。

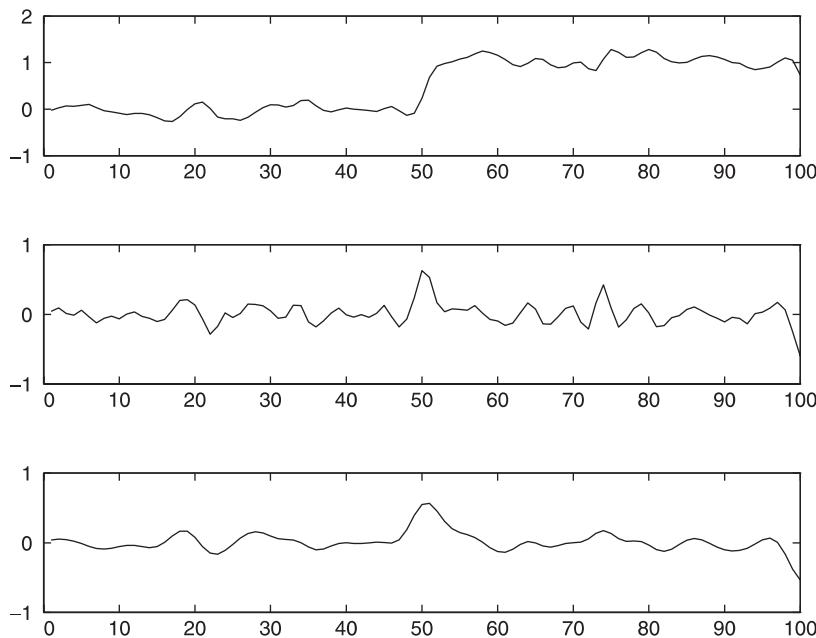


圖 24.8 上圖：在 $x = 50$ 處的邊緣的一維截面方向上的亮度曲線 $I(x)$ 。中圖：亮度的導數 $I'(x)$ 。該函數的較大取值對應於邊緣，不過該函數有雜訊。下圖：亮度的平滑化版本的導數 $(I * G_\sigma)'$ ，其能以一個步驟算出卷積 $I * G_\sigma$ 。位置 $x = 75$ 上的雜訊候選邊緣消失了

在 CCD 相機內，像素量測亮度的過程是透過一個吸收光子而釋放電子的物理機制；無可避免的會有量測上的統計波動——雜訊。雜訊可以用高斯機率分布來模擬，而且每個像素的雜訊都是獨立的。一種平滑影像的辦法是賦予每個像素點的值為其相鄰像素點的平均值。這樣可以傾向於消除較為極端的值。但是我們應該考慮多少個相鄰像素點——是一個、兩個，還是更多？一個合理的答案是利用加權平均法，將最近的點權重設為最高，依序隨著距離到最遠的像素點將權重降低。高斯濾波器便是使用這樣的方式(Photoshop 軟體的使用者可以透過使用高斯模糊選項來確認此點)。回顧一個標準差為 σ 及平均為 0 的高斯函數為

$$N_\sigma(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/2\sigma^2} \quad \text{一維情況下，或者}$$

$$N_\sigma(x, y) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2} \quad \text{二維情況下}$$

使用高斯濾波器可將所有 (x, y) 像素的強度 $I(x_0, y_0)$ ，以總和 $I(x, y)N_\sigma(d)$ 替換，其中 d 是從 (x_0, y_0) 到 (x, y) 之間的距離。我們常常使用此種加權總合，以至於其擁有專門的名稱和符號。我們稱函數 h 是兩個函數 f 和 g 的卷積(convolution，記作 $f * g$)，如果有

$$h(x) = (f * g)(x) = \sum_{u=-\infty}^{+\infty} f(u)g(x-u) \quad \text{一維情況下，或者}$$

$$h(x, y) = (f * g)(x, y) = \sum_{u=-\infty}^{+\infty} \sum_{v=-\infty}^{+\infty} f(u, v)g(x-u, y-v) \quad \text{二維情況下}$$

所以平滑函數透過將影像以高斯函數作卷積而得， $I * N_\sigma$ 。當 σ 取值為 1 個像素時，對於少量雜訊的平滑處理已經足夠了，而取值為 2 個像素時能夠對更大量的雜訊進行平滑，但是會損失某些細節。因為高斯函數的作用一有距離就快速減弱，我們可以將求和中的 $\pm\infty$ 替換為 $\pm 3\sigma$ 。

我們可以藉由合併平滑函數以及尋找邊緣功能成為一個演算，來將此計算最佳化。有如下定理：對任意函數 f 和 g ，它們卷積的導數 $(f * g)'$ 等於其中一個函數與另一個函數的導數的卷積，即 $f * (g)'$ 。所以與其對影像先平滑後求導，我們不如直接將影像與高斯平滑函數 N_σ 進行卷積。隨後我們將邊緣的峰值超過閥值的部份給標記出來。

這個演算法若要從一維截面應用到普通二維影像會有一個正規化動作。在二維平面上，邊緣可能具有任意的角度 θ 。讓我們將影像亮度視為一個純量函數，其中具有變數 x 及 y ，其梯度為一個向量。

$$\nabla I = \begin{pmatrix} \frac{\partial I}{\partial x} \\ \frac{\partial I}{\partial y} \end{pmatrix} = \begin{pmatrix} I_x \\ I_y \end{pmatrix}$$

影像位置中的邊緣，其亮度會有一個明顯的變化，因此某個邊緣點，其梯度後之量值 $\|\nabla I\|$ 應該會很大。而令人有個別興趣的是梯度方向

$$\frac{\nabla I}{\|\nabla I\|} = \begin{pmatrix} \cos \theta \\ \sin \theta \end{pmatrix}$$

上式將會給每個像素一個 $\theta = \theta(x, y)$ ，其定義了在該個像素的邊緣方向。

如在一維中，我們不計算 ∇I 以形成梯度，而是計算 $\nabla(I * N_\sigma)$ ，即以高斯對影像作卷積之平滑運算後之梯度。同樣的，捷徑等效於將影像對高斯函數部份微分做卷積。一旦我們計算出梯度值，我們可以藉著找出邊緣點和連結這些點來求出邊緣。要分辨出某個點是否是邊緣點，我們必須沿著梯度方向來尋找前後附近一段距離的點。若在這些點中的某一點其梯度值很大，那麼我們可以得到一個較好的邊緣點，透過稍微移動邊緣曲線。此外，若梯度量值太小，這個點便不會是個邊緣點。因此，在一邊緣點上，梯度大小是沿著梯度方向上的局部極大，而梯度大小是大於某個適合的閥值。

一旦我們用這種演算法標記出邊緣像素，下一階段就是把屬於同一條邊緣曲線的像素連接起來。假設任意兩個相鄰且方位一致的邊緣像素必屬於同一條邊緣曲線，就可以把它們連接起來。

24.2.2 紋理

在日常用語中，紋理是某個表面視覺上的感覺——你所看到的會引起你去想到如果你碰觸這個表面的感覺[「紋理」(texture)這個字和「紡織品」(textile)具有相同的字根]。在電腦視覺當中，紋理表示在一表面其空間上重複出現的形式，可以被視覺所感覺。範例包括建築物窗戶組成的圖案、毛衣上縫線排列、美洲豹皮膚上的花斑、草地上一片一片的草、海灘上的卵石，以及體育場中的人群。有時紋理排列具有明顯的週期特性，就像毛衣上縫線排列。而在其他的例子中，例如海灘上的卵石，這種規律性只有統計上的意義。

亮度是每個像素的性質，但是紋理的概念僅在多像素表面上才有意義。例如在某個區域中，我們可以計算出每個像素的方向，並且將此區域分類到一個方向的色階分佈圖中。在一堵牆中的磚塊紋理，會在色階分佈圖上擁有兩個峰值(一個垂直，一個水平)，然而在美洲豹毛皮上的斑點紋理將會擁有較為一致的方向分布。

圖 24.9 顯示了在不同的照明情況下，方向會有很大不同。這使得紋理成為辨認物體方向很重要的線索，因為例如邊緣等其他線索，在不同的照明情況下可能會導致不同的結果。

在具有紋理物體的影像中，邊緣檢測會因為平滑的物體而無法成功。這是因為最重要的邊緣可能會遺失在這些紋理元件中。就如同文獻上說的，我們可能會在條紋中找不到老虎。解決此問題的方法是找出紋理特性的差異，就如同我們找出亮度差異一樣。在老虎身上的某區域以及草地背景上的某區域會有非常不同的方位色彩分佈圖，這允許我們能夠找到兩者之間的邊界曲線。

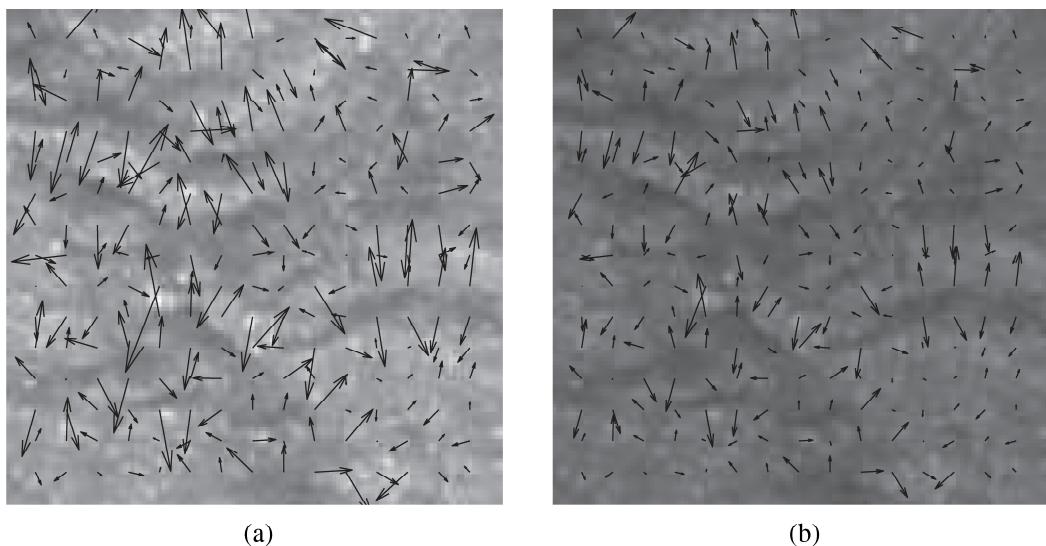


圖 24.9 兩張繃摺米紙的相同紋理影像，但不同照明度。圖上也顯示出梯度向量場(每八個像素)。請注意，當光線較暗時，所有的梯度向量均變短。這些向量並不會旋轉，因此梯度方向不會改變

24.2.3 光流

下一步，讓我們考慮當有一個錄影序列時會發生什麼事，而不僅是一個單一靜態影像。當在錄影中某物體移動時，或者當攝影機相對於一個物體在移動時，影像中的相對應顯著動作稱之為光流。光流描述了影像中紋理動作之方向與速度——賽車影像的光流會以每秒多少像素來計算，而非每小時多少英哩。光流包含許多關於場景結構的有用資訊。例如，從一輛行駛著的火車上所擷取下來的錄影當中可發現，遠處物體比近處物體的外表運動要慢得多。因此，外表運動的速率可以告訴我們一些關於距離的資訊。光流同時也可以讓我們確認動作。在圖 24.10(a)和(b)中，我們可看到一段從網球選手錄影中所擷取下來的兩幅影像。在(c)中我們可看到從這些影像中所計算出來的光流向量，顯示球拍和腿部移動速度最快。



圖 24.10 錄影序列中的兩幅影像。最右邊的圖片是從一張影像到另一張影像的位移所對應的光流場。注意到，網球拍及前腿的移動，是如何藉箭頭方向來捕捉。(來源：Thomas Brox)

在任何點 (x, y) 之光流向量場可以用其在 x 之分量 $v_x(x, y)$ 以及 y 方向之分量 $v_y(x, y)$ 。為了量出光流，我們需要找到相繼時間方塊架之間的對應點。一個最簡單的想法便是基於在附近的影像區域會擁有相似的強度形式。考慮一個在 t_0 時刻以像素 p 即 (x_0, y_0) 為中心的像素塊。這個像素塊要和在時刻 $t_0 + D_t$ 以各種候選像素為中心的像素塊進行比較，其中候選像素位於 $(x_0 + D_x, y_0 + D_y)$ 處。一種可能採用的相似性度量為差值平方和(sum of squared differences, SSD)：

$$SSD(D_x, D_y) = \sum_{(x, y)} (I(x, y, t) - I(x + D_x, y + D_y, t + D_t))^2$$

這裡， (x, y) 的取值範圍是以 (x_0, y_0) 為中心的塊內的像素。我們尋找一個使 SSD 最小的 (D_x, D_y) 。在 (x_0, y_0) 處的光流為 $(v_x, v_y) = (D_x/D_t, D_y/D_t)$ 。注意若此項需要工作，需要某些紋理或是場景有變化。而如果場景中是一堵均勻的白牆，則互相關函數對不同的候選比對點得到幾乎一樣的結果，這時的演算法退化為盲目猜測。量測光流的最佳演算法，當場景僅有部分紋理時就需要依賴額外限制的其他變異。

24.2.4 影像分割

分割是指將一個影像分割成數個具有類似像素區域的過程。每個影像像素都可以與某種視覺特性有關，諸如亮度、色彩和紋理。在一個物體中，或者是它的單獨一部分中，這些屬性的變化相對非常小，而穿過物體之間的邊界時，典型情況下這些屬性中的一個或多個會出現較大的變化。分割有兩個方式，其中一個是檢測這些區域的邊緣，另外一個是檢測區域本身(圖 24.11)。

穿過像素 (x, y) 的邊緣曲線會有一個方向角 θ ，因此檢測邊緣曲線的一個解決方法是如同機器學習分類問題一樣。基於附近的紋理，我們希望計算出確實在此像素上，延著某方向有一個邊緣曲線機率 $P_b(x, y, \theta)$ 。考慮一個圓型平面，其中心位於 (x, y) ，可以被方向為 θ 之直徑切成兩半部。若在 (x, y, θ) 點處有一個邊緣，這兩半部可以預期在它們的亮度、顏色、或是紋理方面有顯著的差異。Martin, Fowlkes 及 Malik(2004)對這兩半部，使用在亮度、顏色、以及紋理色階分佈圖上差異所求出的特徵，並且訓練一個分類器。為此它們使用一組自然影像圖庫，其中有人工手動標記「真實地面」邊緣，而此分類器的目標是儘可能準確的標記出人類所標記的邊緣。

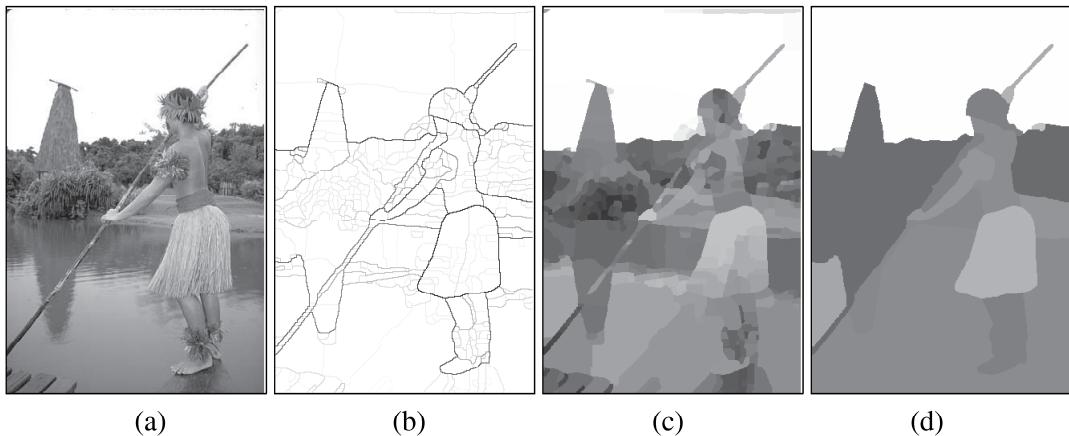


圖 24.11 (a) 原始影像。(b) 邊緣輪廓，當 P_b 值越高時，輪廓會越深。(c) 分割成數個區域，對應於影像較細緻的分割。各區域以其平均顏色作渲染。(d) 分割成數個區域，對應於影像的粗分割，產生更少的區域。(來源：Pablo Arbelaez, Michael Maire, Charles Fowlkes 與 Jitendra Malik)

透過此項技術所標記出來的邊緣，結果要比先前所提到的邊緣檢測方式所標示出來的效果要好。不過仍然有兩個限制：(1) 由閥值 $P_b(x, y, \theta)$ 所形成的邊緣像素將不一定可以形成封閉曲線，因此這個方式無法畫出區域。(2) 這個方法僅能使用在局部材質，並不能使用在整體一致性限制上。

另一個變通的方式，是嘗試將根據像素的亮度、顏色與紋理，「群組化」成數個區域。Shi 和 Malik(2000)把它描述為圖分割問題。圖的每個節點對應於像素，而每條邊對應於像素之間的連接。連接一對像素 i 和 j 的邊上的權值 W_{ij} 是基於這兩個像素在亮度、色彩、紋理等方面的相似度的。隨後我們可以找出最小並正規化的選擇條件來做出分割。簡而言之，圖分割的指標就是使跨組連接的權值總和最小，而組內連接的權值總和最大。

分割基本上僅靠著低階與局部屬性，如亮度及顏色等，我們不能預期是否能夠的推導出場景中所有物體的最終正確邊緣。要能夠找出可信的邊緣，我們需要在場景中物體的其他高階知識。這知識的表現法是熱門研究中的主題。一個廣為流行的策略是先對影像作過度分割，其中包含了數百個被稱為超級像素的同質性區域。從此，接下來基於知識的演算法便可以接手；處理數以百計的超級像素比起處理數以百萬計的原始像素要簡單的多。如何應用物體高階知識將是下一節的主題。

24.3 藉由外表之物體辨識

外表是一個物體看起來像什麼的速記。某些物體分類——比方說，棒球——在外表上的差異性不大；在同一分類中的物體在大多數環境中看起來一樣。在這個例子中，我們可以計算出一連串的特徵來描述影像的每個分類，其所可能包含的物體，之後以一個分類器來測試。

其他物體分類——例如房子，或是芭蕾舞者——可能會差異很大。房子可能具有不同的尺寸、顏色以及形狀，同時從不同的角度看起來會差異很大。一個舞者在不同姿勢，或是不同舞台燈光下看起來都不一樣。一個有用的抽象化是這樣說，某些物體是由局部圖樣所組成，這些圖樣很容易相對彼此移動。那麼我們可以這樣發現物體：觀察檢測器反應的局部色彩分佈圖，其顯示出某部份是否存在，但是卻壓抑了其所在處的細節。

用一個有經驗的分類器來測試每個影像的分類是一個重要且普遍的解決方式。當所有的人臉都直視照相機時，上述的方式可以運作的很好，但是當低解析度以及在很差的照明之下，所有的臉看起來會很相似。臉是圓形的，且相較於眼窩來說臉是較為明亮的；由於眼窩是凹陷的，因此是眼窩是暗的，而嘴巴和眉毛也都會是暗的。照明的變化也會導致這個型態有所變化，但是其變化的範圍是可以估計的。因此要偵測一幅影像中是否有人臉於其中成為可能。以前這是電腦圖學的一大挑戰，不過現在這個功能已經是所有平價數位相機的基本功能。

從此刻起，我們僅考慮鼻子朝向正面的人臉；我們會在後面章節處理臉轉動時的狀況。我們用一個固定大小的視窗掃描影像，計算出其特徵，並且將特徵送至分類器。這個策略有時被稱為滑動視窗。這些特徵必須要能分辨出陰影，同時要隨著照明變化而有所改變。其中一個重要的策略是建立梯度方向之特徵。另一個則是估計並校正在每個影像視窗中的照明狀況。要找出不同大小的人臉，可對較大或較小影像來重複掃描過程。最後，我們會對比例尺以及位置進行後處理，以便產生最終偵測結果。

後處理是很重要的，因為我們不可能剛好選擇到一個適合人臉的視窗大小(即使我們使用了各種不同的視窗)。因此，我們將會使用許多重疊的視窗，每個視窗將會回報是否有人臉符合。然而，若我們使用分類器可以回報反應強度(比方說，運用邏輯迴歸或是支援性向量機等)我們可以將在某區域這些部分重疊的符合項，推導出單一高品質的符合。這給了我們一個可以在不同位置與不同尺度操作的人臉偵測器。而為了尋找旋轉的臉，我們使用兩個步驟。我們可以訓練一個迴歸程序來估計在一個視窗中任何人臉的最佳方向。現在對於每個視窗，我們可以估計其方向，重新定位視窗，之後透過我們的分類器來測試是否視窗中一個正向我們的人臉。圖 24.12 顯示了上述討論架構的系統。

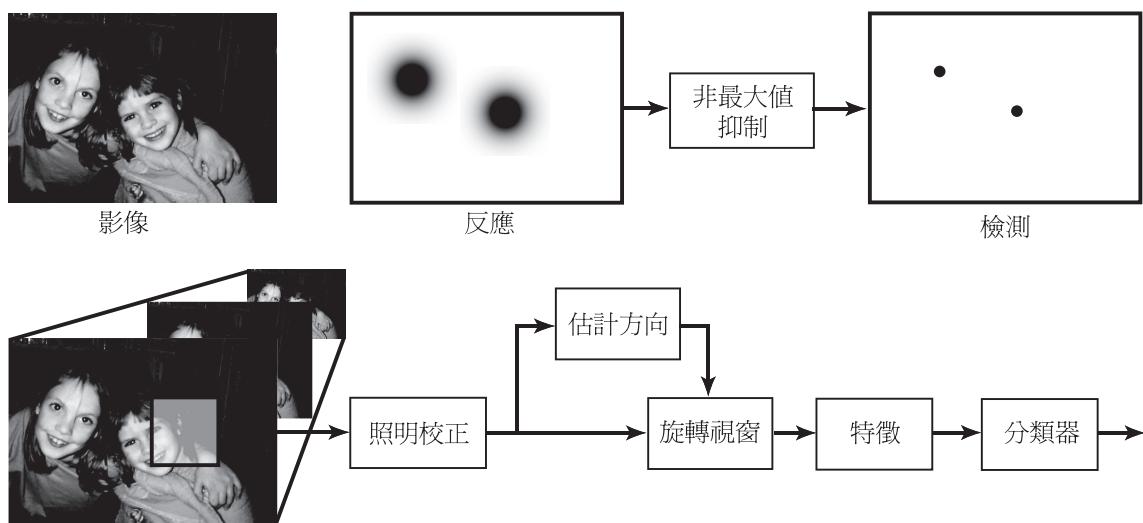
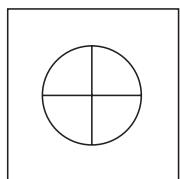


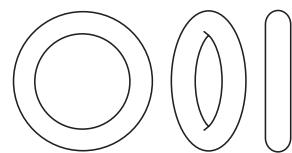
圖 24.12 人臉辨識系統可能會有差異，但是大多數仍照著此處所顯示的兩個架構。圖上半部，我們從影像到反應，然後應用非最大值抑制來得到最強的局部反應。其反應是由圖下半部的過程所得。我們以一個固定大小的視窗來掃描稍大或略小的影像，以便分別找到一個較小或較大的人臉。在視窗中的照明已經被校正過，且一個迴歸引擎(通常是一個神經網路)會預測人臉的方向。視窗被校正到這個方向，並且被傳到一個分類器。分類器的輸出隨後會進行後處理，來確保影像中每個位置僅有一張臉存在。

測試數據很容易取得。這裡有數個已標記的人臉影像數據資料組，而轉動的人臉視窗也很容易建立(只要從訓練資料組中旋轉視窗即可)。在此有個技巧是廣泛的使用原本的樣板視窗，然後藉由旋轉視窗方向、改變視窗中心位置、以及稍微改變視窗的比例尺，便可產生新的樣本。這是個得到較大資料組的簡單方式，同時也可忠實反應真實影像；這個方法通常可以大幅度的改進影像品質。根據上述方法所製作的人臉偵測器，通常對於正向的人臉均可以達到很好的辨識效果(對於側面效果則較差)。

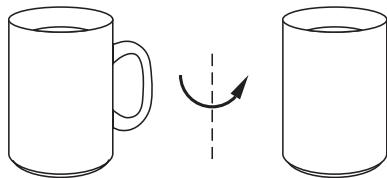
24.3.1 複雜外表與圖樣元素



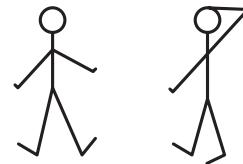
縮短透視



視角



阻擋



變形

圖 24.13 外表差異來源。首先，元件可能會透視縮短，如同左上方的圓形區塊。這個區域是斜著看，所以在影像中為橢圓。第二，從不同的方向來觀看物體時，其形狀會有戲劇性的改變，這個現象稱之為視角(aspect)。在圖右上方，是三個甜甜圈的三個不同視角。圖左下方的馬克杯把手則是，當馬克杯旋轉到一個地方時會消失，此種現象稱之為阻擋(occlusion)。在這個情況下，由於本體以及把手均屬於相同的馬克杯，我們有自我阻擋。最後，在圖右下方，有些物體能大幅變形

許多物體會比人臉產生更多更複雜的形式。這是因為有許多效應可以改變物體影像中的許多特徵。這些效應包括了(圖 24.13)：

- **透視縮短**，表示某個以傾斜角度觀看時會產生顯著失真。
- **視角**，其表示從不同角度觀看時，物體形狀會有差異。即使簡單如甜甜圈的物體也都會有好幾個視角；從側面看，看起來是個被壓扁的橢圓，但是從下方觀看，其變成一個環。
- **阻擋**，表示從某些觀看角度來看，某部份會被隱藏。物體可能會阻擋到別的物體，或是部份物體會阻擋到其他部分，此種稱之為自我阻擋。
- **變形**，物體內的自由度可以改變它的外觀。比方說，人們可以移動他們的手臂以及雙腳，而產生許多不同的身體組態。

然而，我們先前以不同比例搜尋以及搜尋不同位置的方法在此也可以適用。這是因為出現在影像內的某些結構也是來自於物體本身。比方說，一台車子的相片裡可能會顯示部份的頭燈、車門、輪子、車窗、車頂等，雖然它們在不同的圖片中會以不同的方式呈現。這建議要對帶有圖樣元素的物體進行模組化——收集組件。這些圖樣元素繞彼此移動，但是若大多數圖樣元素出現在大概對的位置時，那麼該物件就會出現。一個物體辨認器將會收集特徵，並且分辨是否有特徵元件存在，以及是否它們位於正確的位置。

最明顯的方法是，用出現在該處的圖樣元素之色彩分佈圖來表示影像視窗。這個方法並沒有辦法非常成功的運作，因為會有太多形式會和其他形式搞混。例如，如果有一個圖樣元素是色彩像素，法國、英國以及荷蘭國旗常常會搞混，因為它們都有相同的色彩分佈圖，即使這些顏色的配置有很大差異。將色階分佈圖簡化會產出非常有用的特徵。其重點在於保留其所顯示的空間細節；比方說，頭燈應該都是在車子的前方，而車輪應該是在車底。色階分佈圖為主的特徵已經廣泛的應用在許多辨認應用上；我們將會將它用在行人偵測上。

24.3.2 以 HOG 特徵來進行行人偵測

世界銀行估計每年車禍意外奪走一百二十萬人的生命，其中三分之二都是行人。這表示偵測行人是一個很重要的應用問題，因為車子若是可以自動偵測並且避免行人的話，就可以拯救許多生命。行人會穿著不同的衣著且以許多不同的外貌出現，不過在相當低的解析度時，行人可以具有一個相當顯著特徵的外表。最常使用的範例是走路時正面或是側面的影像。在這些案例中，我們通常看到一個「棒棒糖」形狀——軀幹比腿要寬，其在一個走路時的姿態——或者是一個「剪刀」形狀——在走路時雙腳擺動。我們期望能看到某些手和腿的證據，肩膀和頭的附近曲線也會變得容易看見且分辨。這表示，如果有一個小心建立的特徵架構，我們可以建立一個可用的移動視窗行人偵測。

在行人和背景之間通常不會有強烈對比，因此要表示影像視窗，最好是使用方向而非邊緣。行人會擺動它們的手臂和腿，因此我們應該使用色階分佈圖來壓抑在特徵中的空間細節。我們可將視窗分割成數個小部分，這些部份可以重疊，並且在每一部份內建立一個方向分佈圖。藉此可以產生一個特徵可以用來分辨是否頭肩曲線是在視窗的上方或是下方，但是若頭部輕微的移動將不會有所改變。

我們需要另一個方法來得到好的特徵。因為方向特徵並不會被照明天度所影響，我們並不能特別來處理高對比的邊緣。這表示在行人邊緣的獨特曲線，被當成是一個衣物上或是背景上細緻的紋理，因此這個訊號有可能在埋在雜訊當中。我們可以重建對比影像，是藉由計算具有權重之梯度方向，其反映了在相同區域中，其中一個梯度相較於其他梯度差異有多大。我們將用這個 $\|\nabla I_x\|$ 符號來表示影像中點 x 處之梯度量值， C 下標代表我們希望計算的那個區域， $w_{x,C}$ 下標代表我們對在 x 處的這個區域我們將會使用的權重。權重的自然選擇為

$$w_{x,C} = \frac{\|\nabla I_x\|}{\sum_{u \in C} \|\nabla I_u\|}$$

這比較了在區域中此梯度量值以及其他梯度的差異，因此若梯度遠大於區域中其他梯度，便會得到一個較大的權重。其所得到的特稱通常稱之為 **HOG 特徵**(Histogram Of Gradient orientation, 梯度方向分布圖)。

這個特徵結構是行人偵測和人臉偵測最主要差異之處。要不然的話，建立一個行人偵測器非常像建立一個人臉偵測器。偵測器會用一個視窗掃描影像，計算視窗內的特徵，並且用一個分類器來處理。而在輸出端需要使用非最大值抑制。在大多數應用當中，我們已經知道一般行人的大小與方向。比方說，在駕駛應用中照相機是固定在車上，我們預期主要會看到直立的行人，同時我們僅對於在附近的行人有興趣。數個行人資料組已經被公開，這些資料可以用來訓練分類器。

行人並不是我們唯一可以偵測的物體類型。在圖 24.15 我們看到類似的技巧可以應用在不同內容中來找出許多物體。

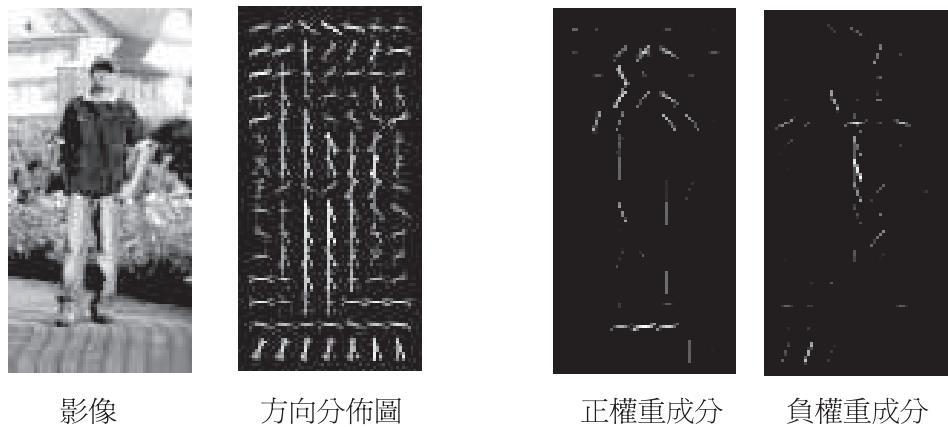


圖 24.14 局部方位分佈圖在辨認複雜物體時是個很有用的特徵。在圖左，是一個行人的影像。在圖左中，其顯示了局部方向分佈圖。我們隨後可以使用一個分類器，例如支援向量機，來為每個分佈圖找出其可最佳從非行人中區別出行人正例的權重。我們可以看到，正權重成分看起來像是一個人的外型。負權重成份較不清楚，他們顯示出非行人的所有圖樣。圖來源：Dalal and Triggs(2005) © IEEE



圖 24.15 圖示為另一物體辨認範例，其使用 SIFT(Scale Invariant Feature Transform)特徵法，一個較早期的 HOG 特徵法。在左邊，一個鞋子以及電話的影像用來做物體模型。在中間，是一個測試影像。在右端，鞋子與電話已經被如下偵測出：找出影像中的某些點，其 SIFT 特徵描述符合一個模型；計算該模型的姿勢估算；驗證該估算。若錯誤正例極少，通常會被驗證為一個強匹配。圖來源：Lowe(1999) © IEEE

24.4 重建三維世界

在本節中我們將說明如何從二維影像出發而得到場景的三維表示。基本的問題是：給予場景中的所有點會沿著一個光線抵達針孔，其投影到影像中相同的點，我們要如何回復成三維的資訊？有兩個想法可以幫助我們：

- 若我們擁有來自兩個(或以上)不同位置攝影機的影像，因此我們可以在場景中以三角定位方式找出該點的位置。
- 我們可以利用有關實際場景的背景知識並且將它加入影像當中。給定一個物體模型 $\mathbf{P}(\text{Scene})$ 及一個渲染模型 $\mathbf{P}(\text{Image} \mid \text{Scene})$ ，我們可以計算出一個後驗分布 $\mathbf{P}(\text{Scene} \mid \text{Image})$ 。

對於場景重建目前尚未有一個統一的理論。我們調查了八個最常使用的視覺線索：**動作**、**雙目立體**、**視覺**、**多重視角**、**紋理**、**明暗**、**輪廓**、以及**熟悉物體**。

24.4.1 運動視差

如果照相機在三維場景中有相對移動，則在影像中的明顯運動，也就是光流，可以同時得到照相機移動以及場景深度的兩個資訊。欲了解這個，我們陳述(但不加以證明)一個方程式，其將光流與觀看者的平移速度 \mathbf{T} 以及場景深度這兩者建立起關係。

光流場的成分為

$$v_x(x, y) = \frac{-T_x + xT_z}{Z(x, y)}, \quad v_y(x, y) = \frac{-T_y + yT_z}{Z(x, y)}$$

其中 $Z(x, y)$ 表示對應於影像上一點 (x, y) 的場景點的 z 座標。

光流的兩個分量 $v_x(x, y)$ 和 $v_y(x, y)$ ，在點 $x = T_x/T_z$ 以及 $y = T_y/T_z$ 處都等於零。這一點被稱為光流場的**輻輳點**(focus of expansion, FOE)。假設我們改變 x - y 平面的原點位置，使它處於輻輳點上，光流運算式將變成一種很簡單的形式。令 (x', y') 為新座標，其定義為 $x' = x - T_x/T_z$ ， $y' = y - T_y/T_z$ 。則

$$v_x(x', y') = \frac{x'T_z}{Z(x', y')}, \quad v_y(x', y') = \frac{y'T_z}{Z(x', y')}$$

注意這邊有一個尺度因素的歧異。若照相機移動速度為原本的兩倍，場景中的每個物體都是原本兩倍大，以及離照相機的距離有兩倍的話，光流場都會是一樣。但是我們仍然可以擷取出一些有用的資訊。

1. 假設有一隻蒼蠅正設法落在牆上，那麼它想要知道在當前速度下經過多長時間能夠接觸到牆。這個時間由 Z/T_z 紿出。注意，雖然暫態的光流場既不能提供距離 Z ，也不能提供速度分量 T_z ，但是它能夠提供二者的比值，因此可用來控制降落的過程。已經有許多實驗數據證明很多昆蟲或動物種類都會使用這個暗示。

2. 分別考慮在深度 Z_1 及 Z_2 的兩個點。我們或許不會知道其絕對值，但是我們可以透過考慮在這些點上光流的倒數，我們可以計算出深度比值 Z_1/Z_2 。這是運動視差很重要的特徵，也就是當我們坐在移動的火車或是汽車時會發現的現象，較遠的地標移動速度會較慢。

24.4.2 雙目立體視覺

大多數脊椎動物具有兩隻眼睛。在失去一隻眼睛的情況下，這是一種有益的冗餘，不過除此之外還有一些其他方面的好處。多數被捕食動物的眼睛長在頭的兩側，使它們具有更寬闊的視野。而捕食動物的眼睛則長在前面，使它們能夠利用**雙目立體視覺**(binocular stereopsis)。這個概念類似動作視差，除了不是使用隨時間變化的影像，我們使用空間中分開的兩個(以上)影像。因為場景中的一個給定特徵相對於每個影像平面的 z 軸的位置是不同的，所以當我們把兩幅影像重疊在一起時，兩幅影像中的影像特徵位置將會出現**視差**(disparity)。你可以在圖 24.16 中看到這一點，金字塔狀物體離我們最近的那一點在右邊影像中移到了左邊，而在左邊影像中移到了右邊。

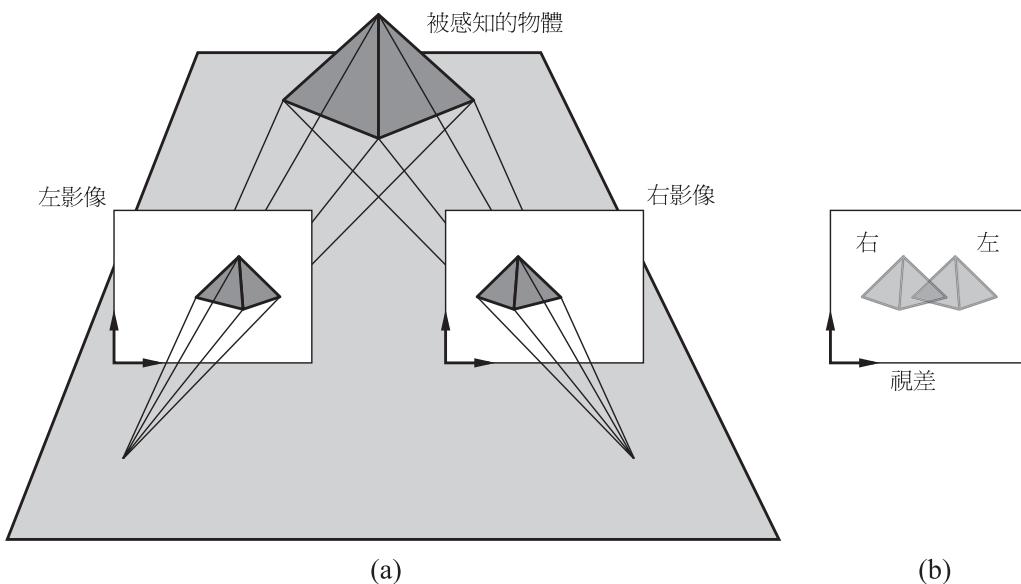


圖 24.16 將一個照相機平移成平行於影像平面，會造成影像特徵在照相機平面中移動。產生的位置視差是對於深度的一個暗示。若我們如同(b)中重疊左影像和右影像，我們可以看到視差

注意到，要量測視差時，我們需要解決對應性問題，也就是，對於左影像中的一個點，決定由相同場景點的投影所產生的右影像中的點。這很像我們之前在量測光流時所做的，最簡單的解決方式類似於基於比較像素的區域，在相對應的點中，並且使用平方差之和。實際上，我們使用更複雜的演算法，其可以利用額外的限制。

假定我們可以量測視差，要如何從視差推導出場景內的深度？我們將需要求出視差與深度之間的幾何關係。首先，我們考慮雙目(或兩個照相機)直視前方，即兩光軸彼此平行的情況。此時右側照相機與左側照相機之間的關係相當於沿 x 軸平移了一段距離 b ，稱為基線。我們可以使用上一節的光流方程式，若我們把它當作是起自作用時間 δt 的平移向量 \mathbf{T} ，且 $T_x = b/\delta t$, $T_y = T_z = 0$)。水平和垂直

視差可以由光流成分計算得知，也就是乘上時間 δt ， $H = v_x \delta t$ ， $V = v_y \delta t$ 。將這些等式代入後，我們可以得到結果 $H = b/Z$ ， $V = 0$ 。用語言表述，即水平視差等於基線與深度之比，而垂直視差等於零。假定我們已知 b ，我們可以量測 H 並且計算出深度 Z 。

人們通常看東西時會注視(fixate)；也就是說，兩眼的光軸交會於場景中的某一點。圖 24.17 顯示了兩隻眼睛注視點 P_0 的情況，它到兩眼連線中點距離為 Z 。為方便起見，我們計算角度視差，其單位是弧度。在注視點 P_0 的視差為零。對於在距離再遠 δZ 處的另外某點 P ，我們能夠計算出 P 在左右兩幅影像上的角度偏移，分別稱為 P_L 和 P_R 。如果左右兩邊各相對 P_0 偏移了一個角度 $\delta\theta/2$ ，那麼 P_L 和 P_R 之間的偏差，也就是 P 的視差，恰好等於 $\delta\theta$ 。從圖 24.17， $\tan\theta = \frac{b/2}{Z}$ 與 $\tan(\theta - \delta\theta/2) = \frac{b/2}{Z + \delta Z}$ ，但是對小角度而言， $\tan\theta \approx \theta$ ，因此

$$\delta\theta/2 = \frac{b/2}{Z} - \frac{b/2}{Z + \delta Z} \approx \frac{b\delta Z}{2Z^2}$$

此外，由於實際的視差為 $\delta\theta$ ，我們可得知：

$$\text{視差} = \frac{b\delta Z}{Z^2}$$

對於人類， b [雙眼間的基線(baseline)長度] 約等於 6 cm。設 Z 大約是 100 cm。若最小可分辨的 $\delta\theta$ (對應於像素尺寸)是 5 弧秒左右，由此給出 δZ 的值為 0.4 mm。若 $Z = 30$ cm，我們得到非常小的值 $\delta Z = 0.036$ mm。也就是說，在距離為 30 cm 時，人眼能夠分辨小到 0.036 mm 的深度變化，使我們能作穿針引線這樣精細的工作。

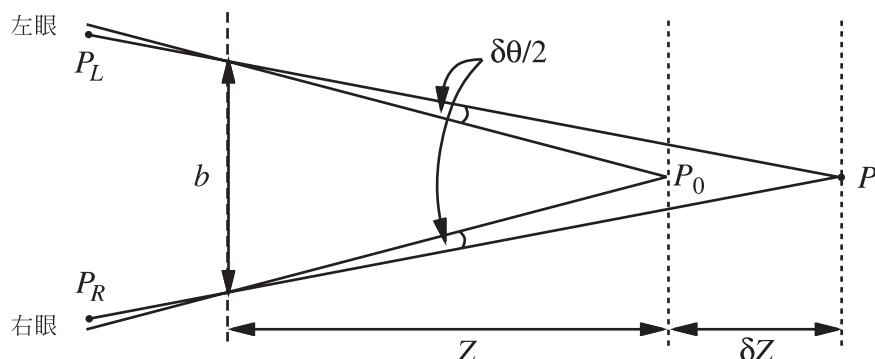


圖 24.17 立體視覺中視差和深度的關係。兩隻眼睛的投影中心相距 b ，而光軸相交於固定點 P_0 。在場景中的點 P 對兩隻眼睛的投影為 P_L 與 P_R 兩點。以角度表示，兩者間的視差為 $\delta\theta$ 。請參閱課本內文

24.4.3 多重視角

從光流或是立體視差中所看到的形狀是兩個一般架構中的例子，其從多個視角中找出深度。在電腦視覺中，沒有理由限制不可以對動作作微分或是僅使用兩台攝影機在交錯在注視點。因此，有許多技巧已經發展出來，其利用在多重視點(即使來自數百或數千台相機)中可得的資訊。以演算法來說，基本上有三個子問題需要解決：

- 對應性問題，亦即，辨別不同影像中的特徵，影像是三維世界裡相同特徵之投影。
- 相對方位問題，亦即，決定固定於不同相機的座標系統間的轉換(旋轉及平移)。
- 深度估計問題，亦即，決定世界中不同點的深度，其中影像平面至少可由兩個視點獲得。

對於對應性問題的穩健匹配方法(伴隨求解相對方位及場景深度的數值穩定演算法)是電腦視覺的成功例子之一。由 Tomasi 和 Kanade(1992)所進行的這個方面研究結果如圖 24.18 和 24.19 所示。

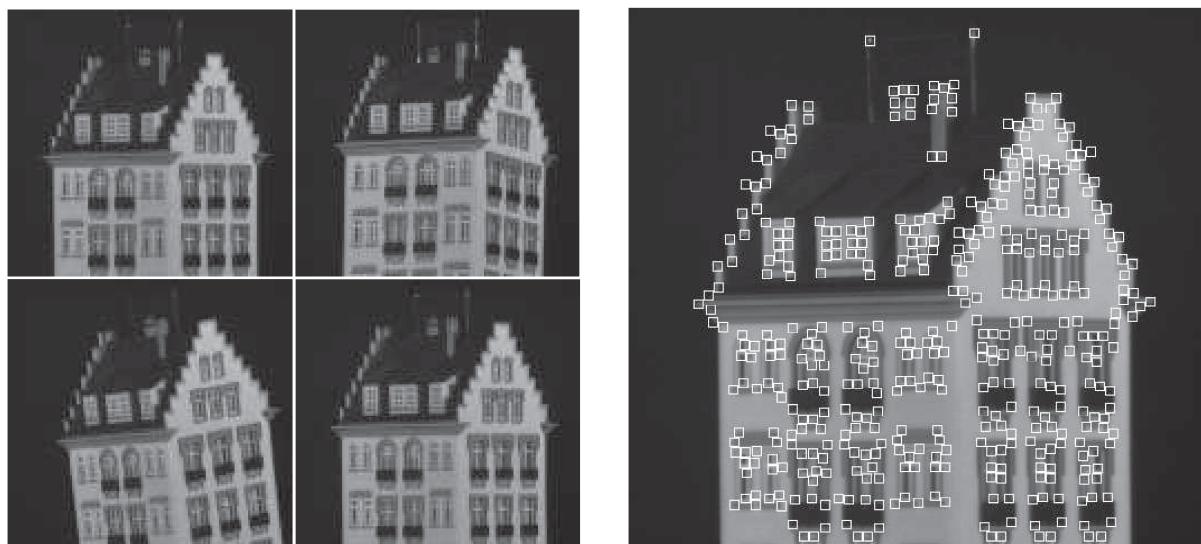


圖 24.18 (a) 一個影片序列中的四個碼框，其中攝影機相對於物體有移動和旋轉。(b) 序列的第一碼框中，小方塊突顯顯示了特徵檢測器找到的特徵(來源：Carlo Tomasi)

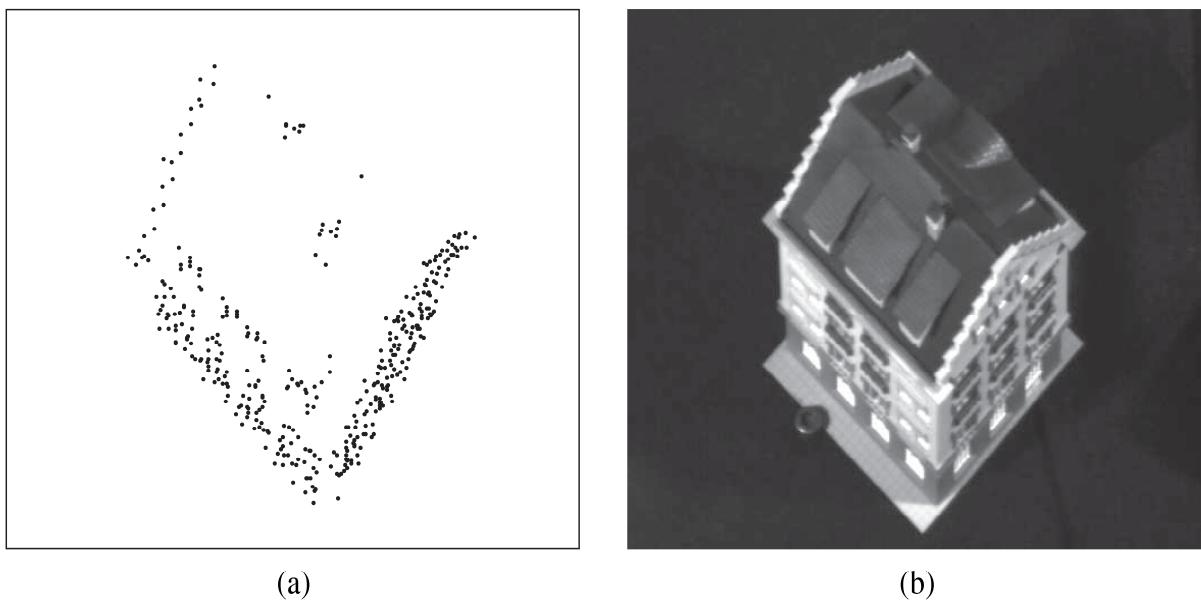


圖 24.19 (a) 圖 24.18 中影像特徵位置的三維重建(從上方看的俯視圖)。
(b) 實際的房屋，從同一個位置拍攝得到

24.4.4 紋理

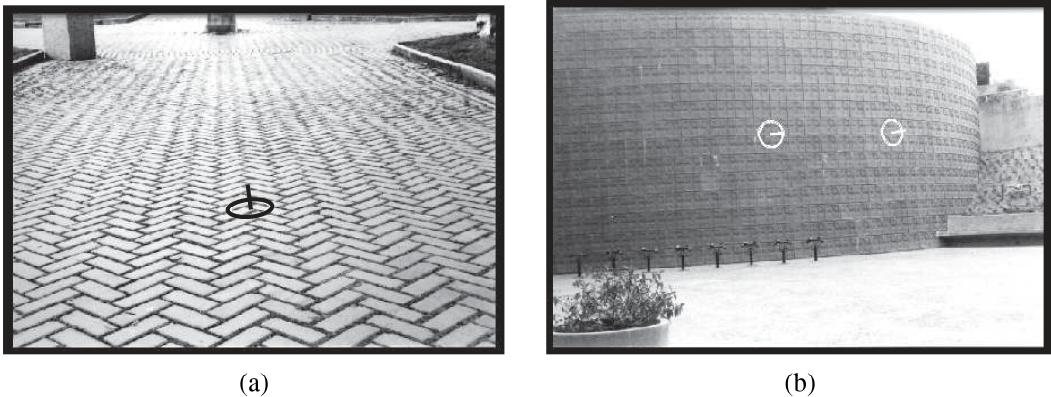


圖 24.20 (a) 一個具有紋理的場景。假設真實紋理的一致性可允許恢復表面方向。計算出的表面方向用疊加的白圓圈和指標表示，並且經過了變換，彷彿圓圈是畫在表面上的某一點處。(b) 從一個曲線表面的紋理作形狀恢復(這次是白圈與指標)[承蒙 Jitendra Malik 和 Ruth Rosenholtz(1994)允許使用這些影像]

先前我們看到如何使用紋理來分割物體。這也可以用來估計距離。在圖 24.20 中我們看到一個場景中的同質紋理，可以產生影像中不同的紋理元件，或稱為圖素。所有在(a)圖場景中的行人道磚塊都是一致的。它們在影像中會看起來有些差異，其主要原因有二：

1. 圖素到照相機的距離不同。距離較遠的物體看起來會比較小，其比例常數為 $1/Z$ 。
2. 圖素的透視縮短(foreshortening)程度不同。若所有圖素在地平面，則以一角度(偏離垂直更多)觀之的距離會更有透視縮短的效果。透視縮短效果會和 $\cos \sigma$ 成比例，其中 σ 是傾斜度，其為 Z 軸和圖素表面法向量 n 的夾角。

研究人員已經發展出不同的演算法，試著利用在投影圖素外表的變化作為決定表面法向量的一個基礎。然而，這些演算法的精準度以及可適用性並非處處適合，且如同使用多重視角一樣。

24.4.5 明暗

明暗——從場景中的物體表面上不同部分接收到的光強度的變化——是由場景的幾何特性和表面的反射特性決定的。在電腦圖學中，目標是根據場景的幾何特性和場景中物體的反射特性計算影像亮度 $I(x, y)$ 。而電腦視覺的目標則是相反的過程——也就是說，根據影像亮度 $I(x, y)$ 重新獲得幾何特性和反射特性。這已被證明是非常困難的，除非是在一些最簡單的情況下。

從 24.1.4 節的物理模型，我們知道若一個表面法向量朝向光源，該表面就會很亮；但是若是背光面，該表面會較暗。我們並不能遽下結論說，較暗的表面就是在背光面；相對的，有可能是因為較低的反照率。一般來說，反照率在影像中會很快速的改變，而明暗的變化是較為緩慢的，通常人類善於使用他們的觀察力來分辨一個較灰暗的表面，是由於低照度、表面方向，還是低反照率。為了簡化這個問題，讓我們假定已經知道每個表面點之反照率。不過還是很難去找出法向量，因為影像亮度是其中一個量測值，但是法向量有兩個未知參數，因此我們無法快速的解出法向量對於這個問題的重點在於附近的法向量是相似的，因為大多數的表面是平滑的——它們並不會有快速的變化。

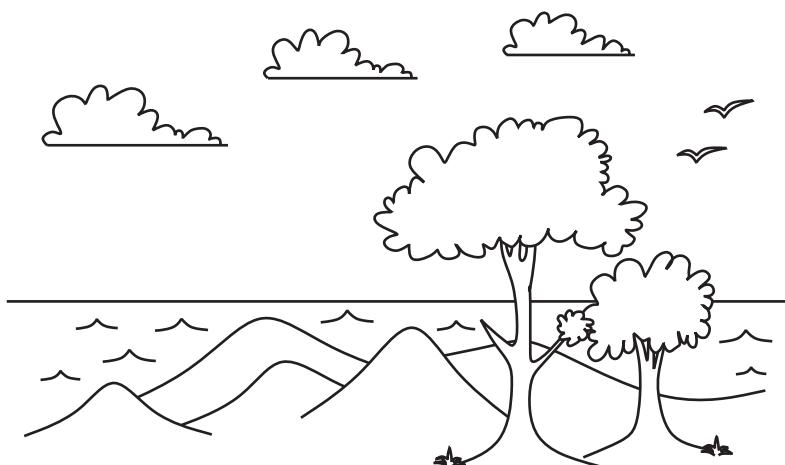
真正的難題在於處理交互反射。如果我們考慮一個典型的室內場景，例如辦公室中的一個物體，那麼物體表面就不只被光源照亮，場景中其他物體的反射光也很好地充當了次級光源。這些相互照明的效果是非常顯著的，同時使得要預測法向量與影像亮度之間的關係非常困難。兩個帶有相同法向量之表面可能會有完全不同的亮度，因為其中一個可能接受到來自一面全白牆壁的反射光，而另一面接收到較暗書櫥的反射光。儘管非常的困難，這個問題還是非常的重要。人類似乎可以忽略交互反射的效果，並且從明暗的情況感知到某個形狀，但是對於這個機制目前我們所知甚少且深感挫折。

24.4.6 輪廓

圖 24.21

一個令人回味的線條圖

(承蒙 Isha Malik 允許使用)



當我們看到類似圖 24.21 所示的線條圖時，會對其中的三維形狀和佈局有一個生動的理解。這是是如何做到的？在場景中熟悉物體的確認組合以及通用限制的應用如下：

- **相交的輪廓**，例如山丘的外型。輪廓的一邊是較靠近觀看者，另外一面則是較遠。類似區域凹陷以及對稱的特徵提供了線索來解決**圖-地面問題**——假定輪廓的某一邊是圖(較近)，而另一邊是地面(較遠)。而在一個相交的輪廓中，視線是垂直於場景中的表面。
- **T 型連結點**。當一個物體碰到另一個時，假定較近的物體是不透明時，較遠物體的輪廓會被擋住。此時會影像中會產生一個 T 型-連結點。
- **在地平面的位置**。人類就像許多具有地域性的動物一樣，常常居住在具有**地平面**的場景之中，而在此平面上會有許多不同位置的物體。因為重力的緣故，通常物體並不會浮在空中，而是穩定的放在地平面上，此外我們可以利用此觀看場景的特別幾何學。

讓我們開始處理，在地平面上具有不同高度以及不同位置物體之投影。假定眼睛或是相機，從地平面算起之高度為 h_c 。若有一高度為 δY 的物體位於地平面上，其底部位於 $(X, -hc, Z)$ 而頂端位於 $(X, \delta Y - hc, Z)$ 。底部投影到影像點為 $(fX/Z, -fh_c/Z)$ 而頂端投影至 $(fX/Z, f(\delta Y - hc)/Z)$ 。較近物體的底部(較小的 Z 值)投影到影像平面的較底部；較遠的物體期底部則較靠近地平線。

24.4.7 物體以及場景之幾何結構

一般成人頭部大約 9 英吋長。這表示若有人站在 43 英呎遠，他的頭從照相機看起來的視角大約是 1 度左右。若我們看到某個人的頭僅有半度，我們可經由貝式干涉推導出我們看到的正常人，其距離為 86 英呎遠，而非一個人她的頭是正常的一半大小。以上解釋提供了我們一個方法來檢查行人偵測器，以及一個用來估計物體距離的方法。比方說，所有行人都具有相同高度，並且他們趨向站在地平面上。若我們知道在影像中地平線的位置，我們便可以由行人距離照相機的間距排列出行人的位置。我們真的可以計算出來，因為我們知道行人的腿在何處，而在影像中若行人的腿較靠近地平線，表示他們距離照相機較遠的位置(圖 24.22)。離照相機較遠的行人在影像中看起來也會比較小。這表示我們可以將某些偵測器的反應排除——若一個感測器發現某個行人在影像中較大，而且又很靠近地平線，那個它就發現了一個異常的行人；這樣的行人不會存在，因此偵測器一定有問題。事實上，許多或大多數的影像視窗並不是可接受的行人視窗，也不需要傳給偵測器。

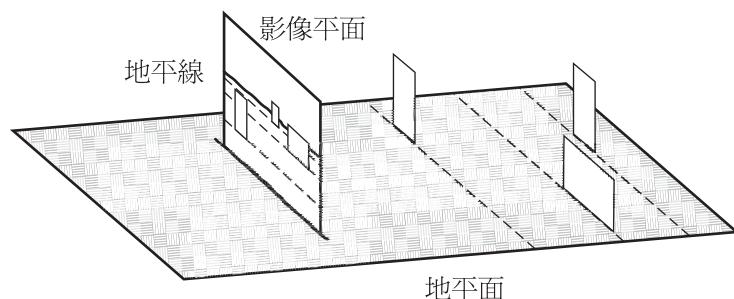


圖 24.22 行人站在地平面上的影像中，腿較靠近影像地平線的人，必定是距離較遠(頂圖)。這表示他們必定在影像中看起來較小(左下方圖)。這表示在影像中真實行人的大小與位置取決於他人還有地平線位置。要利用此點，我們需要對地平面作識別，其可用紋理恢復形狀的方法。由此資訊，以及從一些也可能出現的行人，我們可以恢復地平線，如中央影像所示。在右圖為，已知這個幾何背景下的可接受行人框。注意到，在場景中位置較高的行人必定較小。若不是如此，則他們為錯誤正例。影像出處：Hoiem *et al.*(2008) © IEEE。

要找出地平面有許多策略，包含大略找出一個地平線，在其上有很多藍色部分，或是使用表面方向法來估計紋理變形。另一個較為典雅的策略是利用我們的幾何限制反推而得到。若有數個行人在場景中且距離照相機不同位置，那麼一個合理且可靠的行人偵測器應該能夠估計出地平線的位置。這是因為行人的相對尺寸是找出地平線的一個重要線索。因此我們可以從偵測器找出地平線，並且利用這個來估計修正行人偵測器的錯誤。

若我們對於物體很熟悉，我們可以不只可以估計距離，因為在影像中物體看起來的樣子取決於其姿勢，也就是說其相對應於觀察者之位置與方向。這個特性有許多的應用。例如，在一個工業操縱作業中，機械手只有知道物體的姿態，才能夠把它拿起來。在剛體的情況下，不論是三維還是二維，這個問題都有一個基於校正方法(alignment method)的既簡單又清楚明確的解決方案。我們現在開始討論這種方法。

設物體是用 M 個特徵或不同的三維空間點 m_1, m_2, \dots, m_M ——也許是多面體物體的頂點——來表示的。它們都在一個對物體來說較為自然的坐標系中進行測量。那麼這些點受到一個未知的三維旋轉 \mathbf{R} 的影響，伴隨著平移一個未知量 t ，然後投影到影像平面上得到影像特徵點 p_1, p_2, \dots, p_N 。一般來說， $N \neq M$ ，因為有些模型點可能被遮住了，而且特徵檢測運算元會漏掉一些特徵(或者由於雜訊的原因會創造出錯誤的特徵)。我們可以表示為：

$$p_i = \Pi(\mathbf{R}m_i + \mathbf{t}) = Q(m_i)$$

對於一個三維模型點 m_i 以及對應的影像點 p_i ，其中， \mathbf{R} 是旋轉矩陣， \mathbf{t} 是平移量， Π 表示透視投影或者它的一種近似，例如比例正交投影。淨結果就是一個變換 Q ，將模型點 m_i 與影像點 p_i 對準。雖然我們最初並不知道 Q 是什麼，但是我們卻知道(對於剛體來說) Q 對所有的模型點一定是相同的。

已知 3 個模型點的三維座標與它們的二維投影，就可以求解 Q 。直觀上是這樣的：人們可以寫出將 p_i 和 m_i 座標聯繫起來的方程式。在這些方程式中，未知量對應於旋轉矩陣 \mathbf{R} 和平移向量 \mathbf{t} 的參數。如果我們有足夠的方程式，就應該能夠求解 Q 。我們不準備在這裡給出證明；我們只是陳述以下結論：

給定模型中不共線的三點 m_1, m_2 和 m_3 ，以及它們在影像平面上的比例正交投影 p_1, p_2 和 p_3 ，則恰好存在兩個從三維模型座標到二維影像座標的變換。

這兩個變換透過在影像附近的反射而相關，並能夠透過一個簡單的封閉形式的解進行計算。如果我們能在影像中辨識 3 個特徵對應的模型特徵，我們就能夠計算 Q ，即物體的姿態。

讓我們用數學語言對位置和方向進行描述。在以針孔為原點，光軸(圖 24.2)為 Z 軸的坐標系中，場景中一點 P 的位置可以用由 3 個數值表示的座標(X, Y, Z)刻畫。我們所能得到的是該點到影像上透視投影座標(x, y)。這樣就確定了一條從針孔發出透過 P 點的射線。這兩點之間的距離是未知的。名詞「方向」含有兩重含義：

1. 物體作為一個整體的方向：

這可以用物體坐標系相對於照相機坐標系的三維旋轉量來描述。

2. 在 P 點處物體表面的方向：

這可以用物體表面單位法向量 \mathbf{n} 來描述——它是指明與物體表面垂直的方向的向量。通常我們用變數傾角(slant)和斜角(tilt)來表示表面方向。傾角是 Z 軸和 \mathbf{n} 之間的角度。斜角是 X 軸和 \mathbf{n} 在影像平面上的投影之間的角度。

當照相機相對於物體運動時，物體的距離和方向都在改變。只有物體的形狀(shape)是不變的。如果該物體是個立方體，那麼無論怎麼運動它還是立方體。若干世紀以來，幾何學家曾想方設法對形狀進行形式化描述，其基本的概念是在某些變換群下——例如旋轉和平移的組合，保持不變的屬性即為形狀。其困難在於，需要找到一種對全部形狀的表示方法，它應該足夠通用，可以適用於真實世界中形形色色的物體——而不只是諸如圓柱體、圓錐體和球體之類的簡單形式——同時又易於從視覺輸入中發現。對表面的局部刻畫問題的理解，則要深入得多。本質上，可以從曲率的角度來完成：當在表面上向不同方向運動時，表面法向量是如何變化的。對於平面來說，根本不存在任何變化。對於圓柱體來說，在平行於軸線的方向上沒有變化，而在垂直於軸線的方向上，法向量將以反比於圓柱體半徑的速率旋轉，諸如此類。這些都是被稱為梯度幾何學的學科所研究的課題。

物體的形狀與一些操縱任務(例如確定物體可以被抓住的部位)有關，不過它最重要的用途是物體識別，其中幾何形狀與色彩、紋理一起提供了最有效的提示，使我們能夠辨識物體，以及對影像內容按已知類別進行分類，等等。

24.5 從結構資訊中進行物體辨識

在影像中，將一個盒子放在行人旁邊，將可以避免人們開車撞到行人。我們已經看到我們可以藉由收集方向所提供的證據，以及使用分佈圖方法，來壓縮可能令人覺得困擾的空間細節。若我們想要知道更多有關誰在做什麼，我們必須要知道他們的手臂，雙腿，身體，以及頭部在影像內的位置。個別身體部位是差異很大的，以至於很困難用一個移動視窗法來偵測，因為它們的顏色和紋理有著極大的差異，同時它們在視窗中也僅佔一小部分。一般而言，前臂以及小腿會小到僅有兩到三個像素。這些身體部位通常不會單獨出現，此外其所聯結的地方也是非常的重要，因為容易找到的部份可以告訴我們要到哪裡去找那些不容易找到的部份。

在影像中找出人體的位置在視覺中是一個很重要的工作，因為身體的形狀通常會表示人現在正在做什麼。一個稱之為可變形範例的模型可以告訴我們那個組態是可被接受的：手肘可以彎曲，但是你的頭絕對不可能和腿連在一起。一個人最簡單的變形範例是前臂連接到上臂，上臂連接到軀幹，依此類推。在此有許多種不同的模型：比方說，我們可以表示出以下事實：左上臂與右上臂擁有相同的顏色和紋理，左腿和右腿也是。即使有這麼多的模型，不過要完全可以找出來還是相當困難。

24.5.1 身體的幾何形狀：找出雙手與雙腿

從現在開始，我們假定我們已經知道一個人的身體看起來像什麼(也就是說，我們已經知道這個人衣著的顏色以及紋理)。我們可以將身體的幾何形狀，以一個具有十一個部份的樹來表示(左右上下肢、軀幹、臉部、頭頂的頭髮)，並且每個部份都以長方形表示。我們假設左下臂的位置與方向(姿態)是獨立於其他所有部份，以及左上臂的姿態；左上臂的位置與方向(姿態)是獨立於其他所有部份以及軀幹；並且將以上的假設推廣到所有部份，包括雙腿、臉部以及頭髮等。這樣的模型通常被稱為「紙板人」模型。這些模型可以形成一個樹，其根通常位於軀幹部份。我們會以此影像來搜尋最佳符合紙板人的部分，並且使用干涉法對一個以樹為結構的貝式網(請參閱 14 章)。

在此有兩個評估此組態的依據。首先，一個影像方塊應該要看起來像是它的區塊。此時，我們會保留模糊空間來解釋什麼是準確，但是我們假定我們有一個函數 ϕ_i ，可以用來說明這個影像方塊有多麼符合身體區塊。對於每一對相關區段，我們有另一個函數 ψ ，其對於一對影像矩形間的關係與從身體區段中所預期者的匹配性進行評分。每個區塊之間的關連性可以形成一個樹，因此每個區塊僅會有一個母區塊，我們可以將其表示為 $\psi_{i,\text{pa}(i)}$ 。若關聯性很高，則所有的函數值會增大，所以我們可以把它當成是一個對數(log)的機率關係。若有某個區塊特別符合，其將會分配影像方塊 m_i 到身體區塊 i 為

$$\sum_{i \in \text{segments}} \phi_i(m_i) + \sum_{i \in \text{segments}} \psi_{i,\text{pa}(i)}(m_i, m_{\text{pa}(i)})$$

動態程式可以找出最佳解，因為其關係模型是一個樹。

這對於連續空間內的搜尋不太方便，我們需要將影像方塊空間數位化。我們透過將固定尺寸(這個尺寸可以根據不同的區塊而有所不同)的長方形位置與方向給數位化來達成。因為腳踝以及膝蓋是不同的，我們需要分辨出一個方塊和旋轉 180 度後的方塊。我們可以看到許多小影像方塊所堆疊出來的大方塊組，其以不同位置以及方向中所切出來的。每個區塊都會有一個堆疊。我們必須要找出每個區塊方塊的最佳配置位置。這會是個漫長的過程，因為有許多影像方塊，對於我們已經擁有的模型來說，若有 M 個影像方塊，選擇對的軀幹將會是 $O(M^6)$ 。然而，若選擇適合的 ψ 則會有許多加速的方法，而且這個方式是很實際的(圖 24.23)。這個模型通常被稱為**圖像式結構模型**。

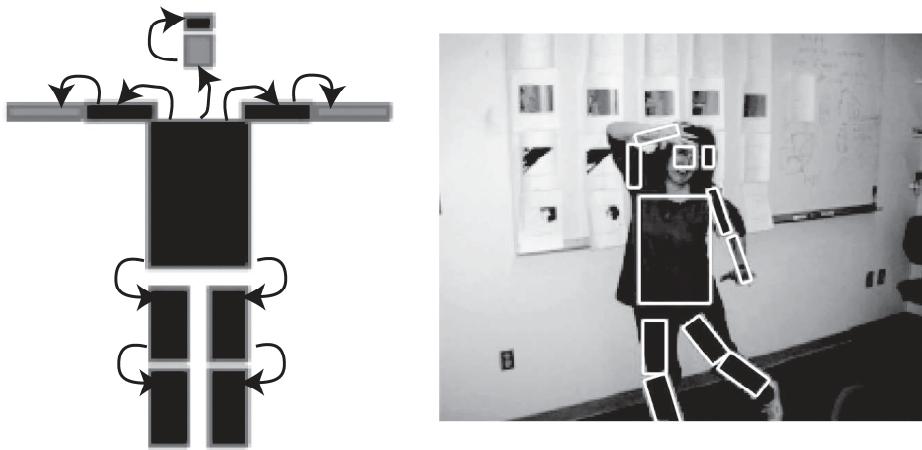


圖 24.23 一個圖像式結構模型，用以評估影像方塊組與一個紙板人(圖左)間的匹配性，方法是評估身體區塊及影像區塊之間的外觀相似性，以及評估影像區塊之間的空間關聯性。一般來說，若影像區塊和另外一個比對的區域有大概相似的外表及位置，就會有較好的匹配。這個外表模型會使用各個區塊的平均顏色，包括頭髮、頭部、軀幹、上下肢等。相對應的關係以前頭表示。在右邊，對於某個特別影像的最佳匹配，是使用動態編程所得。這個匹配是一個合理的身體組態估計。圖來源：Felzenszwalb 與 Huttenlocher (2000) © IEEE

請回想我們的假設，我們知道我們需要了解這個人看起來像什麼。假設我們在單一影像中找到一個符合的人，那麼對於評分區塊是否符合的最有用特徵將會是顏色。紋理特徵在大多數的案子中將沒有辦法工作，因為在寬鬆衣服上的縐褶會產生強烈的明暗變化，其會影響影像的紋理。這些模式都非常的強烈，並且足以干擾衣服的真正紋理。在當前的工作中， ψ 一般會反映出對於區段末端要合理地靠在一起的需求，但通常對角度沒有限制。一般來說，我們不需要知道一個人看起來怎麼樣，但是必須要建立一個區塊外表的模型。我們稱這個人看起來如何的描述叫做**外表模型**。若我們必須要報告在單一影像中一個人的組態，我們可以一開始用一個比較粗造的外表模型，利用此模型來估計其組態，之後重新估計其外表，隨後可重複進行這個過程。在錄影片中，我們擁有同一個人的很多影像，而這個部份可以發現其外表。

24.5.2 一致外表：追蹤在影片中的人

追蹤在影片中的人是一個重要的實際問題。若我們能夠可信的回報在影片中手臂、雙腿、軀幹、以及頭部位置，我們就可以建立更優秀的遊戲介面以及監視系統。濾波方法對於處理這個問題並沒有太多成功結果，因為人們可能會突然加速並且快速移動。這表示在一個 30 Hz 的影片中，影像 i 中身體的組態並不會完全等於在影像 $i + 1$ 中的組態。目前來說，最有效的方式是利用每個影像中的外表改變的非常慢的這個事實。若我們可以從影片中找出一個人的外表模型，那麼我們就可以在圖像結果模型中使用這個資訊來偵測影片中每個影像內的人。我們可以將這些位置以時間關係連結在一起並且產生一個軌跡。

在此有數個方法來找出一個好的外表模型。我們將影片視為一組數目眾多的圖像，其中有我們想要追蹤的人。我們可以利用這組圖來尋找外表模型，並且解釋許多圖片。這將會透過偵測在每個影像中的身體部位，並且使用每個區塊擁有平行邊緣的事實。這樣的偵測器將不會特別準確，但是我們希望找到的區塊是很特別的。它們至少會在影片中的大多數影像中出現一次；這樣的區塊可以透過將偵測器的回應做群組化而被找到。最好是從軀幹開始，因為它很大然後軀幹偵測器是比較可以被相信的。當我們擁有軀幹外表模型後，大腿部分的區塊應該是接近軀幹的，依序延伸到其他的部分。這個部分會產生一個外表模型，但是它未必是正確的，因為如果人們出現在一個靠近固定背景的地方，其區塊偵測器會產生許多假陽性反應。另一個方式是估計影片中的重複影像出現次數，並且重新估計組態以及外表；我們隨後可以看到是否某個外表模型可以說明許多影像。另一個方法，實際上較為可信賴的是應用一個偵測器來對某個固定的組態是否存在於所有的影像中。一個好的組態選擇，是其是否很容易偵測出可信賴，以及當有人出現在組態中會有一個很大的機會找到，即使在一個短的序列當中(側面行走是一個好的選擇)。我們可以微調偵測器的靈敏度，讓它擁有較低的假陽性率，因此我們知道若真的有一個人存在，它可以真正偵測出；還有由於我們可以確認出他們的軀幹、手臂、雙腿以及頭部，因此我們知道這些區塊看起來像甚麼。

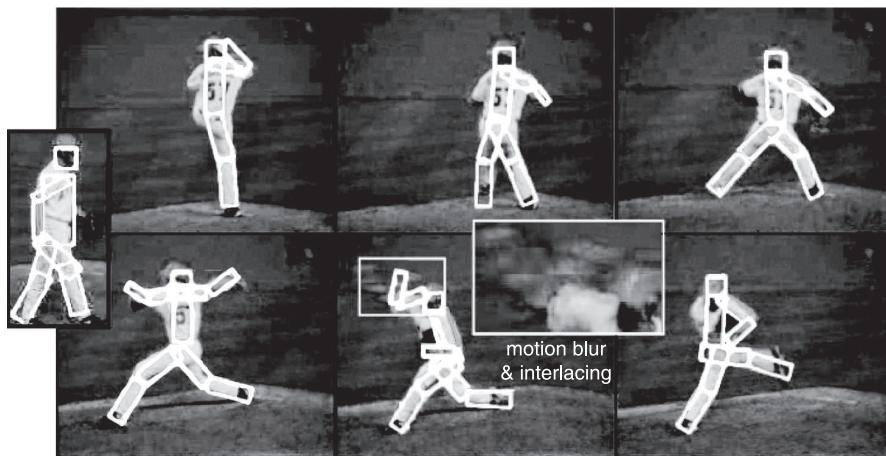
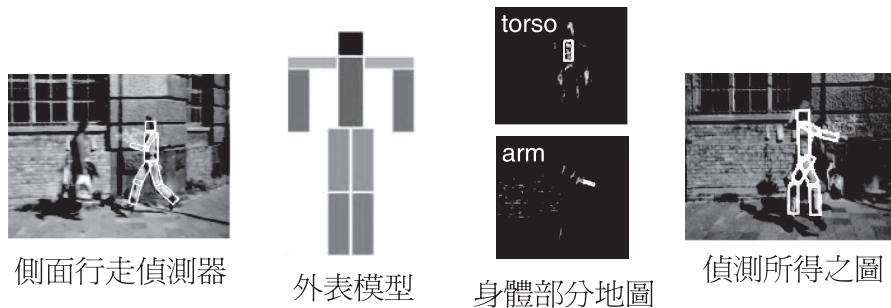


圖 24.24 我們可以用一個圖像式結構模型來追蹤移動的人，方法是先獲得一個外表模型，之後應用此模型。要得到外表模型，我們掃描這個影像而得到一個側面行走姿勢。偵測器不需要非常準確，但應該產生很少的錯誤正例。從偵測器的反應，我們可以讀出在每個身體區塊的像素，以及其他不在區塊的像素。這使得對每個身體部分的外表建立出分辨性模型成為可能，此外這些將會與被追蹤之人的圖像式結構模型作緊密聯繫。最後，我們能藉由在一每一碼框對模型作偵測而進行可靠追蹤。如同影像的下半部的碼框所示，這個方法可以追蹤複雜又快速變化的物體組態，儘管有動態模糊(motion blur)造成的影像訊號劣化。圖來源：Ramanan *et al.* (2007) © IEEE

24.6 使用視覺

若視覺系統可以分析影片並且了解人們在做什麼，我們將能夠：設計建築物以及公共廠所來收集並且使用這些數據，了解人們在公共區域做些什麼；建立更多更準確、更安全、並且更不侵入的監視系統；建立電腦運動評論員；並且建立人與電腦的介面，來觀看人們並且對他們的行為做出反應這些反應介面的應用，包括了電腦遊戲中告訴玩家起來以及在系統中移動，以便節省能源來管理建築物中的熱能和光線，依據他們的位置以及動作。

有些問題已經被充分了解。若人們在影片的影像中相對應很小，而且背景是穩定的，那麼要偵測出人們是很簡單的，因為我們僅需要從原有的影像中減掉背景影像即可。若絕對值差異很大，背景減去法會將此像素視為一個前景像素；透過對時間連結前景像素，我們可以得到軌跡。

某些已經具有完整結構的行為，例如跳芭蕾舞、體操、打太極拳等會有特殊字彙來描述動作。當在一個簡單的背景前進行這些動作時，這些動作的影片我們很容易分析。背景減去法可以定義出主要移動的區域，此外我們也可以建立 HOG 特徵(追蹤光流而非方向)來將資料送到分類器。我們可以利用行人偵測器的某種變形，來偵測動作的一致性模式，其中方向特徵會被收集到色階分佈圖桶中，隨著時間與空間的變化(圖 24.25)。



圖 24.25 某些複雜的人類動作會產生一致的外表和運動。比方說，喝水這個動作包含了將手移到臉的前方。前三個影像是對喝水的正確偵測；第四個是一個錯誤正例(廚師正看著咖啡壺內，並不是在喝水)。圖來源：Laptev 與 Perez(2007) © IEEE

更多普遍問題仍然維持著開放(沒有固定答案)。最大的研究問題在於要連結對身體與附近物體的觀察，以及移動人們的意圖和目標。其中一個困難是，我們缺乏簡單的字彙來描述人類動作。動作有時候很像顏色，一般來說人們會傾向於他們知道很多動作的名稱，但是卻無法產生一長串列表的字彙來說明。有許多證據可以說明這些動作合併——比方說，你可以在提款機領錢時同時喝一杯奶昔——但是我們不確定身體的哪一部分在工作，或是有多少部分在工作，或是如何工作。第二個困難在於我們不知道當動作產生時哪一個特徵會顯現出來。比方說，當我們知道某一個人非常靠近 ATM 提款機時，大概可以知道他要去使用提款機。第三個困難是要去瞭解訓練與實際測試的數據是不值得相信的。比方說，我們不能主張一個行人偵測器會比較安全，只因為其在一群大的資料庫內表現良好，因為資料庫一定會忽略掉某些重要，但不常發生的現象(比方說，騎腳踏車的人)。我們並不希望我們的自動駕駛會撞上行人，只因為行人碰巧在做些不尋常的動作。

24.6.1 文字與圖片

許多網頁提供了影像收集以供觀賞。我們要如何找到我們想要的影像？假定使用者輸入文字查詢，例如「自行車競賽」。某些影像內會擁有關鍵字或是標題，或者來自網頁內靠近圖片所包含的文字。對於這些，影像檢索的工作可能像是文字檢索：忽略掉影像並且試著去找到符合查詢字詞的影像內文字(請參閱第 22.3 節)。

然而，關鍵字通常是不完整的。比方說，一隻在街上遊玩的貓圖片，會被標記為「貓」與「街道」。但是通常很容易會忘記注意到「垃圾桶」或「魚骨頭」。因此，替一個影像加入額外合適的關鍵字是一件很有趣的工作(它可能已經有許多關鍵字在其中)。

進行這個工作最直接的做法是，我們擁有一組已經正確標記的範例影像，而且我們希望去標記某些測試影像。這個問題有時候被稱為自動註解。最合適的解決方案是使用最近區域方法。我們可以發現訓練影像是在特徵空間矩陣中最靠近測試影像，並且回報它們的標記。

這個問題的另一個版本包含了在測試影像中預測要在哪個區域附上那個標記。在此我們將不知道對於訓練資料，那個區域會產生標記。我們可以使用預測最大化的一個版本，來猜測文字與區域之間的初始對應性，並由此估計出一個較好的分解到區域中，以此類推。

24.6.2 由許多視角進行重建

雙目視覺可以工作是因為對每個點，我們擁有四個量測，其中包含三個未知自由度。這四個量測值分別為從每個觀看角度(x, y)位置，以及未知自由度在場景中的(x, y, z)座標值。這個較為原始的論點正確的提出了是有幾何上的限制的，為了預防許多點從可接受的符合中得到。許多影像組的點應該可以毫不模糊的顯示他們的位置。

我們不用老是需要另一個影像，來得到另一組的點的第二個視角。若我們相信原始點組來自一個熟悉的 3D 剛性物體，那麼我們或許可以擁有一個物體模型作為資訊來源。若這個物體模型是由一組 3D 的點所組合而成，或是由物體的一組影像組合而成的，若我們可以建立每個點的對應點，那麼我們就可以求出照相機的相關參數，並且找出原始影像中的點。這是非常有力的資訊。我們可以使用它們來評估我們原本從物體模型所推導出來點的假設是否正確。我們可以藉由使用某些點來計算出照相機的參數，進而找出投影模型，以及檢查是否這些影像點在附近。

我們已經勾勒出一個科技，目前已經發展的非常完整。目前這個科技可以處理非正交視角的影像；要處理僅在某些僅能在某些視角才能看到的點；處理特性(如焦距)未知的照相機；以及利用不同的複雜搜尋對於合適對應；以及對數量眾多的點與視角所得到的影像進行重建。若在影像中，已經部份準確的知道點位置且視角為合理的，那我們可以獲得高解析度照相機以及點的資訊。這些應用包括了

- **建立模型：**

比方說，一個人可以建立一個模型系統，其可以從一串影片序列中重塑一個物體，並且產生一個細微三維網格，並且應用在電腦圖學以及虛擬實境的應用中。類似這個的模型現在可以從明顯相當沒希望的影像組來建立。比方說，圖 24.26 顯示了透過從網路上找到的影像，來重建自由女神的模型。

- **將移動作匹配：**

要將電腦圖學中的特徵應用在實際影片中，我們需要知道實際影片中，其拍攝之照相機如何移動，因此我們才可以正確的給予特徵。

- **路徑重建：**

可移動機器人需要知道它們走過哪些地方。若它們在一個充滿剛性物體的世界裡移動，那麼重建以及保留照相機資訊是獲得路徑的一個方式。

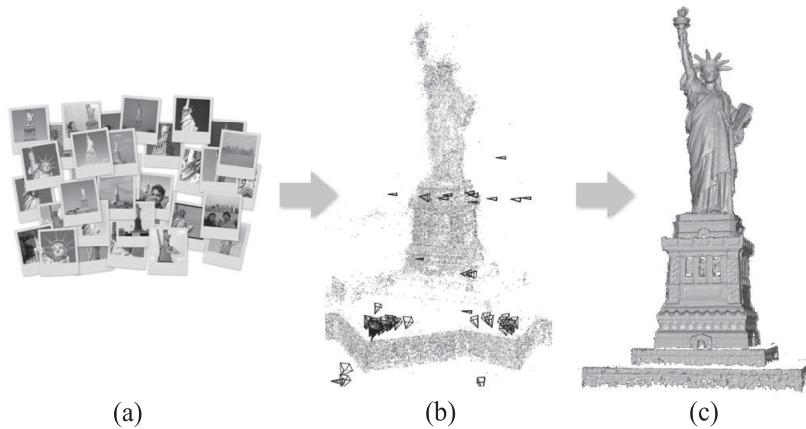


圖 24.26 多視點重建技術現在的水準已經非常先進。此圖顯示了由 Michael Goesele 及華盛頓大學、達姆城科技大學(TU Darmstadt)、與微軟研究所等同行所一起開發建立出來的系統。從一大群使用者所拍攝並且張貼在網際網路上的自由女神像紀念館照片集，(a) 它們的系統可以決定這些照片的拍攝角度，如(b) 所示的黑色小金字塔，而(c)中則展示了完整的 3D 重建結構。

24.6.3 使用視覺來控制動作

視覺的一個主要應用是為操縱物體——拾起、抓住、轉動等等——和避障導航提供資訊。利用視覺完成這些目標的能力，對於動物視覺系統來說是最基本不過的。在許多情況下，如果視覺系統從可獲得的光線場中擷取的僅僅是動物指導其行為所需的資訊，那麼這個視覺系統是最小限度的。很可能，現代視覺系統是從早期原始生物體進化而來的，這些生物體利用身體一端的感光點指引它們自己朝向(或離開)光源的方向。我們在第 24.4 節中看到，蒼蠅使用一個非常簡單的光流檢測系統來降落到牆上。一個經典的研究，《青蛙的雙眼揭示了青蛙大腦的哪些東西》(*What the Frog's Eye Tells the Frog's Brain*)(Lettvin 等人，1959)，對一隻青蛙進行了觀察：「如果它周圍的食物不移動的話，它就會餓死。它對食物的選擇只取決於大小和運動」。

讓我們考慮一個在高速公路上自動車輛駕駛的視覺系統駕駛員面對的任務如下：

1. 橫向控制——確保車輛安全地保持在它的車道內，或者在需要時平穩地換道。
2. 縱向控制——確保和前面車輛之間有一個安全的車距。
3. 障礙物避讓——監視相鄰車道的車輛，並準備好當它們中的某一輛決定換道時應做出避讓動作。

司機要解決的問題在於產生適合的轉向、加速和制動行動，以最好地完成這些任務。

對於橫向控制，需要保持對汽車與車道的相對位置和方向的表示。我們可以使用邊緣檢測演算法來找出相對應於車道標誌部分的邊緣。然後我們可以用光滑曲線擬合這些邊緣部分。這些曲線的參數攜帶有關於汽車的橫向位置、它相對於車道前進的方向，以及車道曲率等資訊。這些資訊，再加上關於汽車的動態資訊，就是駕駛控制系統所需的全部資訊。若我們擁有更詳細的道路地圖，那麼視覺系統更可以用來確認我們的位置(並且用來觀看是否有不在地圖上的障礙物)。

對於縱向控制，需要知道到前方車輛的距離。這可以利用雙目立體視覺或光流來完成。利用這些技術，視覺控制的汽車現在能夠以高速公路上的速度可靠疾駛。

可移動機器人探索不同室內與室外環境也是最常被拿來研究的例子。其中一個特別問題為如何替機器人定位其在環境中的位置，現在已經有不錯的解決方案。在 Sarnoff 的研究團隊已經開發出一套系統，其藉由兩台攝影機以三維方式向前觀看軌道特徵點，並且使用這些點來重建機器人在環境中的位置。事實上，它們有兩套立體視覺照相機系統，一個往前看一個往後看——這給予了機器人更大的功能，也就是它可以穿過陰影、全白的牆壁，以及類似的環境。通常不可能在前面或後面完全沒有任何特徵出現。當然這個情況還是有可能發生，因此使用了一個具有慣性動作單元(Imperial motion unit)的後援系統，其類似人類內耳中感覺加速度的機制。藉由整合這些加速度，我們便可以追蹤其在位置上的變化。結合從視覺以及 IMU 來的資料，是一個機率證據的融合，其可以使用一些濾波技巧來解決，例如卡式濾波器，我們已經在本書的其他部份討論過。

使用視覺里程計(估計在位置上的變化)，就如同在里程計的其他問題中，會有一個有關「漂浮」的問題，位置誤差會隨著時間而累積。解決此問題的方案是使用地標來提供絕對的位置：當機器人通過地圖上的某點時，它可以正確的調整其位置之估計。運用這些技術可以將定位的準確度提升到數公分等級。

駕駛的例子很清楚地說明了一點：對於某個特定任務，並不需要從一幅影像中找到原則上能夠找到的所有資訊。人們不需要得到每輛車的確切形狀，不需要為公路邊的草地表面求解從紋理到形狀的問題，等等。取而代之，一個視覺系統應只計算對於完成任務是需要什麼。

24.7 總結

雖然感知看起來對人類來說是一種不費力氣的活動，它卻需要大量的複雜計算。視覺的目標是為諸如操縱、導航和物體識別等任務擷取所需的資訊。

- 成像(image formation)過程在它的幾何和實體方面是為人熟知的。給定一個三維場景的描述，我們可以很容易地從某個任意的照相機位置製作出它的一幅圖片(圖形學問題)。逆轉這個過程，從一幅影像得到關於場景的描述卻很困難。
- 為了擷取操縱、導航和識別等任務所必需的視覺資訊，不得不構建中間表示形式。初期視覺影像處理(image processing)演算法從影像中擷取原始特徵，諸如邊緣和區域。
- 影像中有一些提示資訊使人們能夠獲得關於場景的三維資訊：運動、立體視覺、紋理、明暗和輪廓分析。為了提供近乎無歧義的解譯，這些提示資訊中的每一個都依賴於實際場景的背景假設。
- 完全通用的物體識別是一個非常難的問題。我們討論了基於亮度和基於特徵的方法。我們還介紹了一個簡單的姿態估計演算法。其他的可能性是存在的。

●參考文獻與歷史的註釋 BIBLIOGRAPHICAL AND HISTORICAL NOTES

眼睛大約在寒武紀大爆發(約 5.3 億年前)時期開始發展，很明顯的在每個祖先身上都有出現。從那時起，無止盡的變化已經在不同的生物上發展，但是都位在相同的基因上 Pax-6，其調控各種不同動物眼睛的發展，如人類、老鼠、果蠅。

對理解人類視覺的系統化嘗試可以追溯到古代。Euclid(約西元前 300 年)論述了自然透視——與三維世界中每一點 P 相聯繫的映對，射線 OP 方向連接了投影中心 O 與點 P 。他很瞭解運動視差的概念。早在古羅馬時代就已經開始在藝術作品當中使用透視技巧，其證據在 Pompeii(西元 79 年)的廢墟中，但是其大約失傳了大約 1300 年。對透視投影的數學上的理解(這裡是指在投影到平面上的上下文中)在 15 世紀文藝復興時期的義大利產生了下一步重大發展。Brunelleschi(1413)通常被認為創作了最早的基於三維場景正確幾何投影關係的畫。1435 年，Alberti 整理了這些規則，從而激發了幾代藝術家的靈感，他們的藝術成就至今仍令我們歎為觀止。尤其值得一提的是李奧納多·達·芬奇(Leonardo da Vinci)和 Albrecht Dürer 對透視科學(他們當時就這麼稱呼它的)的發展。達·芬奇在 15 世紀後期關於光線和陰影的相互作用(明暗對照法，chiaroscuro)、陰影的本影和半影區以及空間透視的描述仍值得一讀——參見 Kemp 的譯文(1989)。Stork(2004)利用電腦視覺技術，分析了文藝復興時期的不同藝術創作。

雖然希臘人認識到了透視，但他們卻令人感到好奇地被眼睛在視覺中的作用所迷惑。亞里斯多德認為眼睛是會發出射線的裝置，相當於雷射測距儀的工作模式。10 世紀的 Alhazen 等阿拉伯科學家消除了這種錯誤觀點。Alhazen 同時也開發了暗室(camera obscura)(拉丁文 camera 是指「房間」或「腔室」)，也就是有一個帶有針孔的房間，其會將影像投影在對面牆上。當然，影像是反的，於是導致了無休止的困惑。如果認為眼睛也是這樣的影像裝置，那我們怎樣看到正確方向的影像呢？這個謎團困擾著那個時代最偉大的頭腦(包括達文西)。Kepler 首先提出眼睛的透鏡會將影像聚焦在視網膜上，而 Descartes 使用手術將牛眼取出，並且驗證了 Kepler 的想法。在此當然還是會有疑惑，為何我們所看到的影像不是上下顛倒？今天我們知道這僅是一個如何正確擷取視網膜上資料的問題。

在 20 世紀上半葉，視覺方面最重要的研究結果是由 Max Wertheimer 領導的完形(Gestalt)心理學派取得的。他們指出感知組織的重要性：對一個人類觀察者而言，影像並不是一堆光感測器(電腦圖學中的像素)輸出的集合；而是被整合成一致的群體。對於尋找區域及曲線的電腦視覺其想法原點，可以追溯到這個洞察。完形心理學派同時也注意到「圖片-地面」的現象——也就是在世界上分開兩個影像區域的輪廓是在不同深度，若是在較近的區域內的物體，便會被歸類為「圖片」，而非較遠區域的「地面」。有關根據其在場景中之顯著性來分類影像曲線電腦視覺的問題，已經被視為是這個想法的一般解。

二戰後的一段時期以重建活動為顯著特徵。最為重要的當屬 J. J. Gibson(1950, 1979)的工作，他指出了光流和紋理梯度對於估計表面的傾角和斜角等環境變數的重要性。他重新強調了刺激的重要性和豐富性。Gibson 強調了主動觀察者的角色，其自主導向之動作會擷取有關外界環境之資訊。

1960 年代起開始有電腦視覺研究的出現。Roberts(1963)在 MIT 的論文是此領域中最早的發表著作，介紹了一些重要概念，例如邊緣偵測與模型導向匹配。在此有個傳說，Marvin Minsky 把「解決」電腦視覺這個問題指派給一個研究生作為暑假計畫。根據 Minsky 的說法，這個傳說是錯的——事實上是指派給一個大學生。但是實際上他是一個非常優秀的大學生，Gerald Jay Sussman(現在已經是 MIT 的教授)，此外這個工作並不是去「解決」視覺，而是調查視覺的某些部份。

在 1960 與 1970 年代，有關這方面的進展很慢，其阻礙主要是來自於缺乏計算以及儲存資源。低階的視覺處理受到很多注目。廣泛使用的 Canny 邊緣檢測法是 Canny(1986)提出。基於多尺度、多方位的影像濾波而找出紋理邊界的技術可追溯至如 Malik 及 Perona(1990)的工作。結合多個線索——亮度、紋理以及顏色——在一個學習架構中找出邊界曲線，是由 Martin，Fowlkes 及 Malik(2004)所提出，其明顯的改進工作效率。

找出局部區域一致亮度、顏色以及紋理的最接近的問題，會導致他自己形成公式來找出最佳部分，變成一個最佳化的問題。有三個頂尖的例子，分別 Geman 與 Geman(1984)所提出的 Markov Random Fields 逼近法，Mumford 與 Shah(1989)所提出的變異性公式，以及 Shi 與 Malik(2000)所提出的正規化切割。

在 1960 年代、1970 年代與 1980 年代的大部分時間，有兩個進行影像辨識的範例，其由對於所感知為何的不同觀點所主導，而成為主要問題。有關物體辨認的電腦視覺研究主要注重在三維物體投影成二維影像的議題。校正的想法也是 Roberts 首先引入的，後來重新出現在 20 世紀 80 年代 Lowe(1987)以及 Huttenlocher 和 Ullman(1990)的工作中。此外，另一受歡迎的方式主要是以體積原函數來描述形狀，例如用**正規化圓柱體**，其由 Tom Binford(1971)所提出，證明了非常的受歡迎。

相對的，形式確認社群則不認為三維轉二維的方面的問題非常重要顯著。具有推動性的例子存在於諸如光學字元識別和手寫體郵遞區號識別等領域中，所關心的首要問題是對一個類別中物體的典型變化特點進行學習，並將它們與其他類別區分開。各種方法的比較參見 LeCun 等人(1995)。

在 1990 年代晚期，這兩個範例開始匯集，兩個部份都採用了機率模型以及學習技術，也就是人工智慧很重要的與很受歡迎的方法。在此有兩條研究主軸，許多人對此問題著力甚多。其中一個是人臉偵測，例如 Rowley，Baluja 與 Kanade(1996)，以及 Viola 與 Jones(2002b)其證明了形式辨認技巧是明顯很重要以及有用的技巧。另一個部份是開發點描述器，其能夠藉由部分物體來重建特徵向量。這是由 Schmid 與 Mohr(1996)最先開始探討。Lowe(2004)的 SIFT 描述器已經被廣泛的使用。HOG 描述器則是來自於 Dalal 與 Triggs(2005)。

Ullman(1979)以及 Longuet-Higgins(1981)在早期進行從多重影像重建工作中具有相當大的影響力。Tomasi 和 Kanade(1992)的工作減輕了對於從運動獲得的結構穩定性的顧慮，他們證明了，使用多碼框下，形狀能夠得到相當精確的恢復。在 1990 年代，隨著電腦速度與儲存容量的進步，動作分析已經有許多新的應用。已被證實特別流行的是建立真實世界場景的幾何模型，用於透過電腦圖學技術進行繪製，其指導想法是重建演算法，例如 Debevec、Taylor 和 Malik(1996)所提出的演算法。Hartley 和 Zisserman(2000)和 Faugeras 等人(2001)所著的書提供了關於多視圖幾何學的全面論述。

對於單一影像，從陰影中要找出形狀最早是由 Horn(1970)所開始進行，以及 Horn 及 Brooks(1989)開始了一個對過去一段時間所發表的論文之廣泛調查。Gibson(1950)是首先提出紋理梯度作為找出形狀的線索，而 Garding(1992)和 Malik 與 Rosenholtz(1997)則是首次對於曲線表面進行一個完整的分析。相交輪廓的數學計算，以及較佳了解視覺事件在平滑彎曲物體上之投影，主要是來自於 Koenderink 以及 van Doorn，其在 Koenderink(1990)的 Solid Shape 中找到了一個廣泛解。在最近幾年，注意力已經轉移到從一影像中找出形狀與表面作為一個機率推論問題，其幾何方面線索並沒有明確的被模組化，反而在一學習架構中很含蓄的使用。其中一個好的代表是 Hoiem, Efros 與 Hebert(2008)的研究成果。

對於人類視覺有興趣的讀者，Palmer(1999)的著作提供了最完整的說明；Bruce *et al.*(2003)的著作則是一份較短的教科書。由 Hubel(1988)與 Rock(1984)所撰寫的書則是分別友善的介紹了神經生理學以及感知。David Marr 的書《視覺》(Vision)(Marr, 1982)在把電腦視覺聯繫於心理物理學和神經生物學當中，扮演了一個歷史性角色。然而他的許多特殊模型並未受到時間的測試，其理論感知可以經由分析在資訊面、電腦面以及實用面等各層面上仍然非常光亮。

對於電腦視覺方面，最完整的教科書是 Forsyth 與 Ponce 所撰寫的(2002)。Trucco 與 Verri(1998)有一篇較短的書籍。Horn(1986)與 Faugeras(1993)則是兩本較舊以及依然有用的教科書。

《IEEE 模式分析與機器智慧會刊》(IEEE Transactions on Pattern Analysis and Machine Intelligence)和《電腦視覺國際期刊》(International Journal of Computer Vision)是電腦視覺方面的兩種主要期刊。電腦視覺會議包括 ICCV(International Conference on Computer Vision，國際電腦視覺會議)、CVPR(Computer Vision and Pattern Recognition，電腦視覺與模式識別)，和 ECCV(European Conference on Computer Vision，歐洲電腦視覺會議)。以機器學習元素進行的研究也發表在神經資訊處理系統(Neural Information Processing Systems, NIPS)會議，此外，利用電腦圖學的介面研究工作常出現在 ACM SIGGRAPH(圖學特殊利益團體，Special Interest Group in Graphics)會議。

❖ 習題 EXERCISES

- 24.1 在一棵枝繁葉茂的樹木蔭影下，可以看到許多光點。奇怪的是，它們看上去都是圓形的。為何？畢竟陽光穿過的樹葉之間的縫隙不太可能是圓形的。
- 24.2 考慮一個半徑為 r ，軸線沿 y 軸方向的無限長圓柱體。該圓柱體具有蘭氏面，一個沿 z 軸正方向的照相機對它進行觀察。如果圓柱體被 x 軸正方向上無窮遠處的一個點光源照亮，那麼你期望在影像中能看到什麼？請畫出在投影影像中等亮度區域的輪廓。這些等亮度線之間的間隔是均勻的嗎？
- 24.3 考慮一個半徑為 r ，軸線沿 y 軸方向的無限長圓柱體。該圓柱體具有蘭氏面，一個沿 z 軸正方向的照相機對它進行觀察。如果圓柱體被 x 軸正方向上無窮遠處的一個點光源照亮，那麼你期望在影像中能看到什麼？請畫出在投影影像中等亮度區域的輪廓。這些等亮度線之間的間隔是均勻的嗎？

- 24.4** 一幅影像中的邊緣可以對應於場景中的多個事件。考慮圖 24.4，並假設它是一幅反映真實三維場景的圖片。從影像中辨別出 10 條不同的亮度邊緣，並說出每一條邊緣對應的不連續是(a)深度，(b) 表面法線，(c) 反射，(d) 光照方面的。
- 24.5** 考慮一個用於繪製地形圖的立體照相系統。它由兩個 CCD 照相機組成，每個照相機具有所使用的鏡頭具有焦距 16 cm，而焦點固定在無窮遠。對於左邊影像上的對應點 (u_1, v_1) 和右圖上的對應點 (u_2, v_2) ，可得 $v_1 = v_2$ ，因為兩影像平面的 x 軸都平行於外極線(epipolar line)——即物體到相機的線。兩個照相機的光軸相互平行。它們之間的基線長度為 1 米。
- 如果測量的最近距離為 16 m，那麼能夠產生的最大視差是多少(用像素表示)？
 - 對 16 m 距離處進行測量，由像素分佈所導致的解析度範圍是多少？
 - 什麼距離對應於一個像素的視差？
- 24.6** 下面這些話哪些是正確的？哪些是錯誤的？
- 在立體影像中尋找對應點是尋找立體深度過程中最容易的階段。
 - 在同一個場景的立體視圖中，兩個照相機相距越遠，對深度進行計算的精度越高。
 - 場景中長度相等的線段投影到影像中的長度總是相等的。
 - 影像中的直線必然對應於場景中的直線。
- 24.7** (圖來源：Pietro Perona)。圖 24.27 表示了位於 X 和 Y 的兩個照相機正在觀察一個場景。畫出從每個照相機中看到的影像，假設所有已命名的點都在同一個水準面上。關於點 A、B、C、D、E 到照相機基線的相對距離，根據這兩幅影像可以得出什麼結論？根據是什麼？

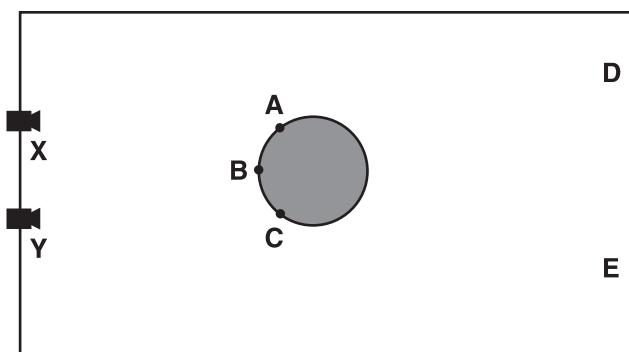


圖 24.27 圖示為雙相機視覺系統的俯視圖，其正在觀察一個瓶子(後面有牆壁)

