Multimedia Content Analysis Homework 4

GMM-based Color Image Segmentation

WEI-CHIH CHANG Q36111281

April 27, 2023

Abstract

使用高斯混合模型完成三種設想實驗並調整其中的參數，從設想實驗中得到訓練集對模型預測結果的影響，並另外討論三種參數：轉換方式、mixture 數量和共變異數類型對模型預測的影響。

1. Related Work

為了得到彩色影像的分割，我使用三種方法將影像轉換到特徵空間，接著使用 EM 演算法估算特徵在特徵空間中的分布，透過高斯分布判斷特徵的局域性結構。

藉由 EM 演算法，我們能判斷出特徵的局域性結構，同時，辨識出的結構數量根據事先設計的聚類數決定，而聚類數則代表高斯混合模型中的 mixture 數量。mixture 的數量越多，越能精確分辨特徵空間的細部局域性結構，但也有著將相同語意的特徵分割的風險。也因為聚類數需要事先設計，因此該演算法不能算是完全無監督的圖像分割方法。同時對於高斯混合模型而言，特徵空間中不同基底之間的共變異數設計方法也是相當重要的，而在在 sklearn 預建好的 mixture.GaussianMixture()函式中有四種不同的共變異數型態，在這篇報告中我也將對其進行實驗並討論之。

綜合上述分析，每個設想實驗中都有三個可調控的參數，分別是 1) 轉換方式 t、2) mixture 數量 nc 以及 3) 共變異數類型 ct。

2. Scenario 1

在 Scenario 1 中，我使用 soccer1.jpg 來訓練高斯混合模型，並對同一張圖片進行類別的推論，以此來辨識足球轉播畫面中場地的部份以及非場地的部分。切割效能是以 pixel accuracy 表示，該方式是將推論結果和 ground truth 進行逐像素的比較，藉此得到分割的精確度。

2.1 Three method to transform from image to feature

一張彩色圖片的最小單位是像素，而平面彩色圖片的像素又由 x、y 兩個軸的座標和代表三種顏色的 RGB 組成，因此每一個像素都值觀地包含五種特徵。也就是說，將這五種特徵投影到上述提及的五種特徵向量，即可將圖片轉換至特徵空間。

第一種轉換方式是將圖片投影到 RGB 三種顏色的特徵向量上，因此特徵空間中的特徵僅保留每個像素的顏色特徵。透過 GMM 的方式聚類，可以將顏色依據語意分類，以足球場的例子為例，根據結果可以看出足球比賽圖片被分成是足球場地的顏色和不是足球場的顏色兩種語意，精確度為 91.8%。

第二種轉換方式則是將圖片投影到包含 x、y 軸以及 RGB 三種顏色的特徵向量上，因此特徵空間中的特徵保留平面彩色圖片像素的所有特徵。在排除顏色和座標位置之間有特徵耦合的情況，透過 GMM 的方式聚類，可以將不同座標的顏色依據語意分類。也就是說如果一張圖包含綠色的草皮和藍色的天空，天空中有一顆綠色的氣球，用第一種轉換方式會將綠色的草皮和綠色的氣球視為相同語意；但透過第二種轉換方式，就會將綠色的草地和綠色的氣球分割為不同語意，這都是因為第二種轉換方式包含座標的特徵。

一樣以足球場的例子為例，因為聚類的標籤結果沒有依照該類別的數量進行排序，而是根據高斯函數的初始化位置決定，因此雖然兩種轉換方式所得到的語意接近，但標籤結果相反，不果我們只要對標籤之間進行映射的正規化，即可得到相同語意的標籤。根據轉換方式二，其更規化後精確度為 97.6%。

轉換方式三是考慮 RGB 三個顏色特徵代表相同語意，這次實驗是將其設計成代表亮暗的語意，因此我們先將三維的 RGB 結果轉換成一維的灰階特徵後，再將圖片轉換到特徵空間，直接藉由圖片判讀，會認為轉換方法三正規化後的精確度一定最低，但其實際正規化後精確度為 93.1%。

會有這樣的結果發生其一是因為 pixel accuracy 是根據像素逐一比對，因此儘管語意辨識大致正確，但若分割的位置與 ground truth 有些為偏移就會造成精確率大幅下降。其二是在 ground truth 中，人像的辨識較為精確，但轉換方式一和轉換方式二在人像辨識上採取較寬容的方式，因此在場上有十三個人的情況上，儘管每個人像的辨識造成 0.5%的誤差，但所累積出的誤差也不容小覷，而轉換方式三則在人像辨識上較為精準，因此儘管場地有大區塊的辨識錯誤但所產生誤差也沒有想像中的嚴重。

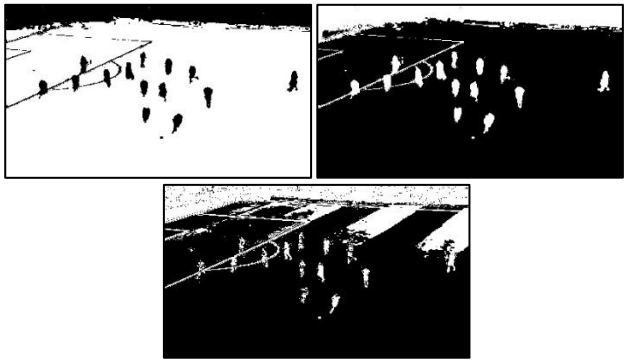

figure 1. Scenario 1 中，不同像素轉換到特徵空間的方法，nc=2，ct=0。1)左上，將像素的顏色轉換到特徵空間；2)右上，將像素的位置和顏色轉換到特徵空間；3)中下，將像素的位置和灰階轉換到特徵空間。

2.2 Effect of different amounts of mixture on performance

在高斯混合模型中，每個高斯 mixture 代表一種語意，例如在 nc=10 時，該模型可以將圖片分成十種語意，從下圖可以發現，該模型居然可以分割數字的語意，也就是能夠分辨特徵空間的細部結構，然而該模型卻將場地分割成多個不同的語意，將相同語意的特徵分割。

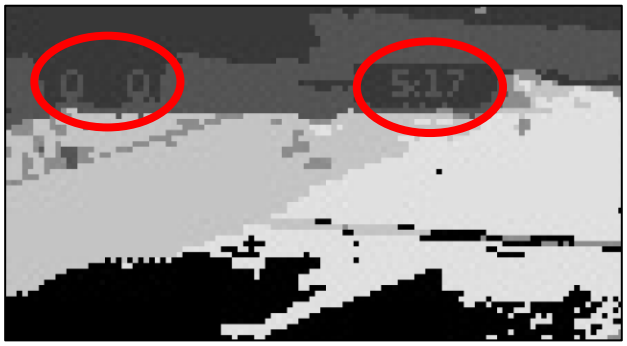

figure 2. 在 t=1，ct=0，nc=10 時，語意分割的細細節

隨著高斯 mixture 的數量增加，足球員的分割越來越精確。然而從 nc=4 開始，模型將足球場分割成多個不同語意，從原始圖片我們可以推論，模型是將深色的草皮和淺色的草皮分割成不同語意，甚至是將養護草皮所導致的分界線視為語意分割的標準。

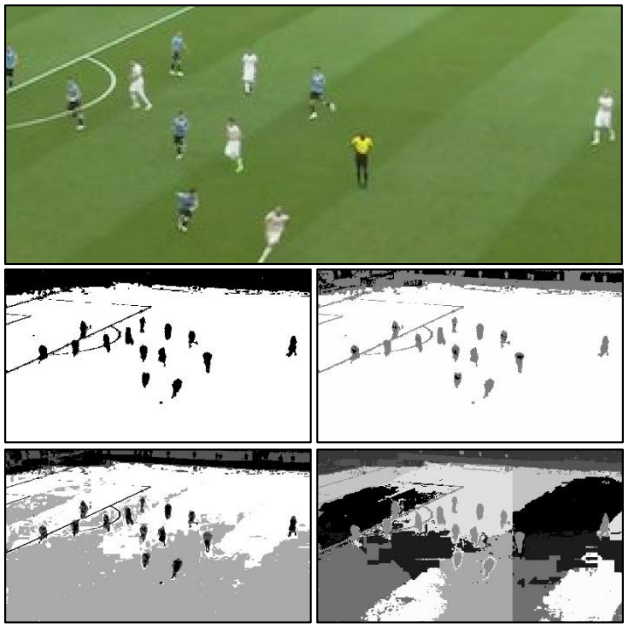

figure 3. Scenario 1 中，不同 mixture 數對語意分割的影響，t=1，ct=0。1)上，原始圖片場地中養護草皮導致的分界線；2)中左，mixture 數=2；3)中右，mixture 數=3；3)左下，mixture 數=4；3)右下，mixture 數=10。

2.3 Effect of covariance type on performance

在 sklearn 預建好的 mixture.GaussianMixture()函式中有四種不同的共變異數型態，分別是‘full’，‘tied’，‘diag’，‘spherical’四種。‘full’指每個分量具有各自不同的標準共變異數矩陣；‘tied’ 指每個分量具有相同的標準共變異數矩陣；‘diag’ 指每個分量具有各自不同的對角共變異數矩陣；‘spherical’指每個分量具有各自的單一變異數。

從下圖我們可以知道，‘full’的語意分割效果最好；‘tied’則因為具有相同的標準，導致其分割的標準是根據像素位置；‘diag’和‘spherical’都將電視轉播的跑馬燈從圖片中分割出來，而‘diag’的 分割傾向於顏色的特徵，‘spherical’則傾向於位置的特徵。

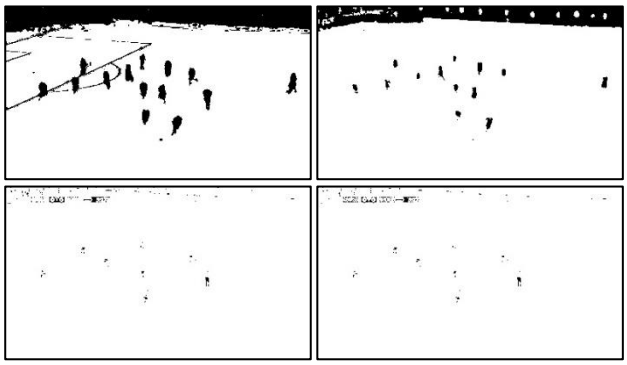

figure 4. Scenario 1 中，不同共變異數類型對高斯混合模型分割效果影響，t=1，nc=2。1)左上，ct=‘full’；2)右上，ct=‘tied’；3)左下，ct=‘diag’；4)右下，ct=‘spherical’。

3. Scenario 2

Scenario 2 我是使用 soccer1.jpg 來訓練高斯混合模型，接著對 soccer2.jpg 進行類別的推論，以此來辨識足球轉播畫面中場地的部份以及非場地的部分。

下圖是不同特徵空間的轉換方法對影像分割的影響。在訓練資料不足的情況下，轉換一和轉換二都只能勉強辨識出場地中深色的草皮。精確度分別為 38%和 36.2%。然而轉換方式三的精確度卻高達 70.2%，這可能是因為轉換方式三的特徵向量分別是像素位置和亮暗程度，因此分割的結果不會因為 soccer2.jpg 中有大面積的陰影而受到影響。

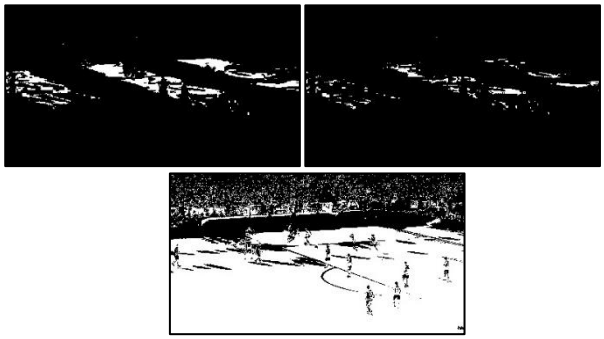

figure 5. Scenario 2 中，不同像素轉換到特徵空間的方法，nc=2，ct=0。1)左上，將像素的顏色轉換到特徵空間；2)右上，將像素的位置和顏色轉換到特徵空間；3)中下，將像素的位置和灰階轉換到特徵空間。

增加 mixture 數量並沒辦法提升足球場地的辨識能力，在這個設想實驗中還是只能夠辨識足球場地中深色的部分，但可以明顯提升人像的辨識能力。

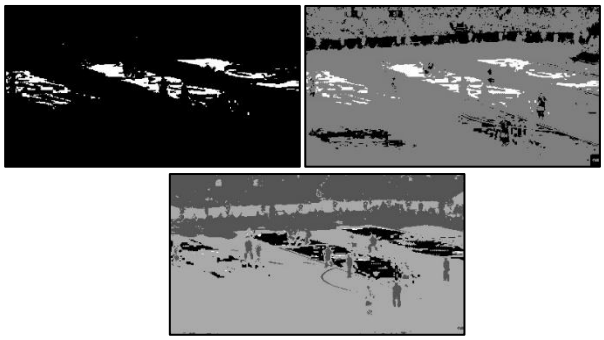

figure 6. Scenario 2 中，不同 mixture 數對語意分割的影響，t=0，ct=0。1)左上，mixture 數=2；2)右上，mixture 數=3；3)中下，mixture 數=4。

因為 'tied'具有標準的共變異數矩陣，因此在面對沒看過的資料時，其精確度仍能達到 89%； 'full'仍能辨識出場地中深色的部分； 'diag' 和 'spherical' 能辨識出跑馬燈的部分，但場地中的其他零星噪點則無法解釋。

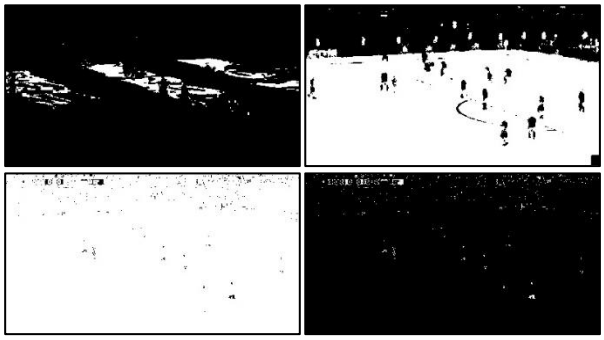

figure 7. Scenario 2 中，不同共變異數類型對高斯混合模型分割效果影響，t=1，nc=2。1)左上，ct= 'full' ；2)右上，ct= 'tied' ；3)左下，ct= 'diag' ；4)右下，ct= 'spherical' 。

4. Scenario 3

在訓練資料更多的情況下，soccer1.jpg 三種特徵空間轉換方式正規化後的精確度分別是 91.7%、92%、96%；soccer2.jpg 三種特徵空間轉換方式正規化後的精確度分別是 93%、93.6%、66.6%。可以發現精確度都有小幅度的下降，這可能是因為僅靠兩張圖不能夠代表足球場地這個語意，但兩張圖特徵空間的聚類結果卻會互相干擾。不過有五項精確度的數值都在九成以上，最差的雖然只有六成，然而卻能分割出更多的細節。

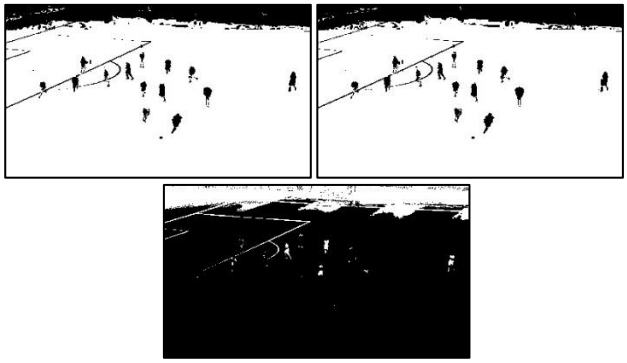

figure 8. Scenario 3-1 中，不同像素轉換到特徵空間的方法，nc=2，ct=0。1)左上，將像素的顏色轉換到特徵空間；2)右上，將像素的位置和顏色轉換到特徵空間；3)中下，將像素的位置和灰階轉換到特徵空間。

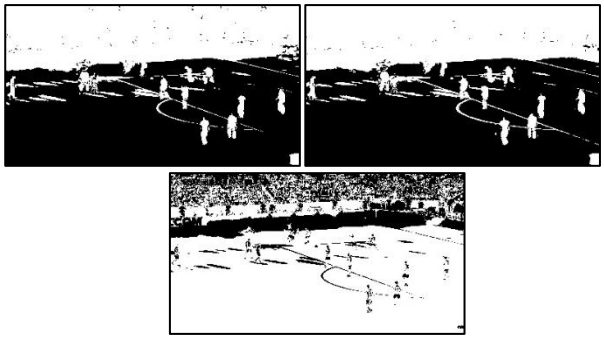

figure 9. Scenario 3-2 中，不同像素轉換到特徵空間的方法，nc=2，ct=0。1)左上，將像素的顏色轉換到特徵空間；2)右上，將像素的位置和顏色轉換到特徵空間；3)中下，將像素的位置和灰階轉換到特徵空間。

更多的訓練資料使得 mixture 數增加的同時，語意的耦合程度下降，分割出的 segmentation 語意變得更為明確。

figure 10. Scenario 3-1 中，不同 mixture 數對語意分割的影響，t=0，ct=0。1)左上，mixture 數=2；2)右上，mixture 數=3；3)中下，mixture 數=4。

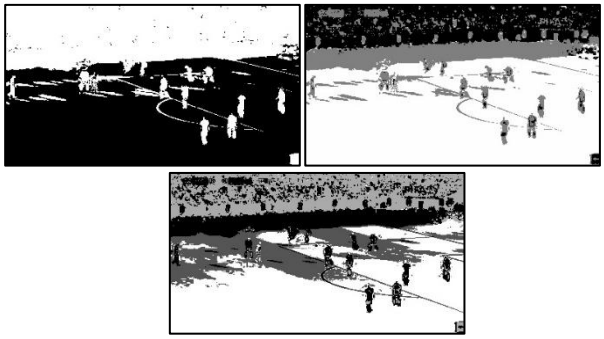

figure 11. Scenario 3-2 中，不同 mixture 數對語意分割的影響，t=0，ct=0。1)左上，mixture 數=2；2)右上，mixture 數=3；3)中下，mixture 數=4。

儘管訓練資料增加，‘full’和‘tied’都是較好的共變異數類型。‘diag’和‘spherical’不利於語意分割任務的進行。

figure 12. Scenario 3-1 中，不同共變異數類型對高斯混合模型分割效果影響，t=1，nc=2。1)左上，ct=‘full’；2)右上，ct=‘tied’；3)左下，ct=‘diag’；4)右下，ct=‘spherical’。

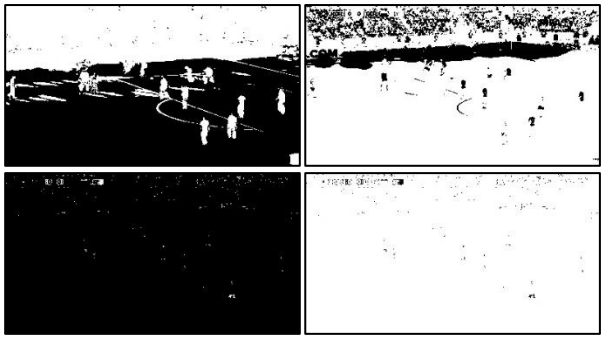

figure 13. Scenario 3-1 中，不同共變異數類型對高斯混合模型分割效果影響，t=1，nc=2。1)左上，ct=‘full’；2)右上，ct=‘tied’；3)左下，ct=‘diag’；4)右下，ct=‘spherical’。

5. Performance

黃底的部分是因為，這兩個聚類方法傾向將數據全猜測為同一類，因此在數據不平衡的情境下，全猜測為同一類的精確度會比隨機猜測來的高。

綜合所有實驗，我們可以發現包含座標的轉換方式是較好的特徵空間轉換方式。變異數型態則是‘full’有最好的效果，但是在面對完全沒學習過的目標時，‘tied’則是可以嘗試的便異數型態。增加 mixture 數量雖然能增加細節的分割效能，但是需要搭配更多的訓練資料。

Scenario 1		Scenario 2		Scenario 3	
名稱|精確率%		名稱|精確率%		名稱|精確率%	
t2_c2_ct0	98	t2_c2_ct1	89	t3_c2_ct0	96
t3_c2_ct0	93	t2_c2_ct1	80	t1_c2_ct0	93
t1_c2_ct1	93	t2_c2_ct3	72	t2_c2_ct0	92
t1_c2_ct0	92	t2_c2_ct2	70	t1_c2_ct0	91

6. Conclusion

這次報告完成單張彩色圖片的圖片分割，並且進行各種參數之間變化的實驗，得到各種情境下較好的參數結果，未來可以將模型應用在影片上，或著是改良多 mixture 的 GMM 模型，使分割的類別能自動與真實語言的語意連結。

參考文獻

[1] Yiming Wu, Xiangyu Yang and Kap Luk Chan, "Unsupervised color image segmentation based on Gaussian mixture model," Fourth International Conference on Information, Communications and Signal Processing, 2003 and the Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint, Singapore, 2003, pp. 541-544 Vol.1, doi: 10.1109/ICICS.2003.1292511.