行政院國家科學委員會專題研究計畫　成果報告

子計畫二：多層次視訊物件分割

計畫類別：整合型計畫
計畫編號：NSC91-2213-E-002-127-
執行期間：91 年 08 月 01 日至 92 年 07 月 31 日
執行單位：國立臺灣大學資訊工程學系暨研究所

計畫主持人：洪一平

計畫參與人員：蔡玉寶、蘇敬仁、蔡尚儒

報告類型：完整報告

處理方式：本計畫可公開查詢

中　華　民　國 92 年 10 月 27 日

行政院國家科學委員會專題研究計畫成果報告

視訊中多物體之偵測、分割與追蹤(2/2)－子計畫二

多層次視訊物件分割
Multi-Scale Segmentation of Video Objects

計畫編號：NSC 91-2213-E-002-127
執行期限：91 年 8 月 1 日至 92 年 7 月 31 日
主持人：洪一平　國立台灣大學資訊工程學系
共同主持人：陳祝嵩　中央研究院資訊科學研究所
計畫參與人員：蔡玉寶、蘇敬仁、蔡尚儒

一、中文摘要

本計劃的目的在於探討多層次視訊物件分割的相關理論與技術。在去年的計畫中我們將傳統的分水嶺分割法由單張影像之分割推展到視訊物件之分割，其中重點在於發展三維時空容積的分水嶺分割技術。在本年度的計畫中，我們改進去年所發展的三維時空容積的分水嶺分割技術，並擴展到多層次的視訊物件分割技術。首先，使用多層次分水嶺分割法將視訊影片切割成多階層的分水嶺區域。針對每一個階層我們建立一個時空相鄰區域圖，再利用最大事後機率的方法來取得一個較佳的分割。我們的方法是從最粗略的階層依序往最細緻的階層進行切割。而較粗略階層所切割的結果會被拿來當做是較細緻階層的最佳化之初始值。本計劃所發展出來的系統允許使用者介入修改切割結果。使用者只需少量的修改，系統會自動地將修改後的正確資訊傳遞到整個視訊影片進行切割結果的修正，以降低人為的介入，並達到更佳的物件分割效果。

關鍵詞：視訊物件、影像分割、貝氏法、馬可夫隨機場、分水嶺演算法

Abstract

The goal of this project is to develop multi-scale video object segmentation. In the first year, we extended the traditional watershed segmentation method, which dealt with only one single image, to video object segmentation, which dealt with a video sequence. We focused on developing a watershed segmentation technique for dealing with three-dimensional spatio-temporal volume. In the second year, we improved the first-year method and extended to multi-scale video object segmentation. First, we partition each image frame into a set of multi-scale watershed regions. Then, we used all the watershed regions at each resolution scale to build a Spatio-Temporal Region Adjacency Graph (ST-RAG) – one ST-RAG for each resolution scale. Next, we performed the MAP labeling process, starting from the ST-RAG of the coarsest scale, and gradually proceeding to the finer scale. At the finest scale, if the segmented result produced by the MAP labeling is not consistent with the user's expectation, he can simply adjust the segmentation result in a couple of frames and restart the process of MAP labeling to propagate the modified result to all frames.

Keywords: Video Object, Image Segmentation, Bayesian Approach,

Markov Random Field, and Watershed Algorithm

二、緣由與目的：

影像分割在影像處理及電腦視覺的領域上一直是一個很基本而且很重要的研究主題，其主要的目的是希望能把影像中具有物理意義的物件給分割出來。由於數位攝影機的快速發展以及 MPEG-4 標準的制定，使得視訊物件的分割研究越來越受到重視。在本年度的計劃中，我們改進去年所發展三維分水嶺分割演算法，並將之擴展到多層次的視訊物件分割技術，以加快視訊物件分割的速度。

由於視訊物件的多樣性，要讓電腦自動且精確地把物件分割出來並不是一件容易的事。目前全自動的視訊物件分割都是假設視訊物件是具備有某種特定性質，方能得到較為精確的結果(例如在[1][2]中所提到的方法)。在本計劃中，我們提供一個互動式使用者界面，可以把使用者在少數影像上修改過的切割結果傳遞到所有的影像上。目的在於讓使用者可以在做最少的修改情況下得到最多最好的修正效果。

三、結果與討論：

在本年度的計畫中， 我們對去年計畫中所提出的分割方法做了些修改，並將之擴展到多層次的視訊物件分割。新的方法可分為兩個階段：「前處理階段」 和「標記階段」，如圖 1 所示。

在前處理階段中，我們先為輸入影像建立不同解析度的多階層影像，接著利用分水嶺分割法把每一階層的影像切割成分水嶺區域，然後再估算每個分水嶺區域的運動向量，最後則為每一階層的分水嶺區域建立出一個「時空相鄰區域圖」。

在標記階段中，系統會從最粗略的階層到最精細的階層分別使用 MAP 方法來求得一個較佳的切割結果。每個較粗略的階層的切割結果都將會是其下一個較精細階層的初始狀態。最粗略的初始值則可由使用者給定，或由某些全自動方法求得（但是無法保證精確度）。當進行完最精細層的 MAP 分割後，使用者可以檢查分割結果，如果有些地方不如使用者預期，使用者可以透過系統所提供之界面來修正。一旦使用者在少量的影像進行修改後，系統會自動地重跑 MAP 分割，把正確的分割資訊傳遞到其它影像中。

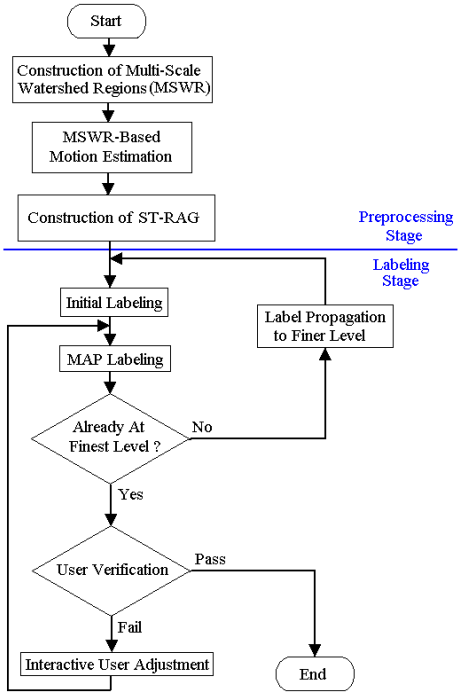

圖 1. 系統流程

以下是前處理階段及標記階段的詳細說明。

1. 前處理階段

在這個階段中，我們先為輸入視訊影像建立不同解析度的影像階層(image pyramid)，再利用分水嶺分割演算法[4]將每個階層的每張影像分割成分水嶺區域，我們稱之為「多階層分水嶺區域」(multi-scale watershed region)。 接著我們以每個分水嶺區域為樣本，在前後的視訊框(frame)中估

算其運動向量，以建立起分水嶺區域之間的時間域(time domain)關係。而在同一視訊框中，每個分水嶺區域與其相接鄰的分水嶺區域則存在著空間域 (spatial domain) 的關係。利用時域與空域的關係，我們可以為每一階層的所有分水嶺區域建立一個時空相鄰區域圖，稱之為 ST-RAG (*Spatio-Temporal Region Adjacency Graph*)。其中 ST-RAG 上每一個節點即代表一個分水嶺區域，節點和節點間，則依據視訊在空間及時間的關係來決定是否產生連結。當兩個分水嶺區域在同一個視訊框(frame)中，且兩者相鄰時，即產生空間的連結；當兩個分水嶺區域位於相鄰的視訊框上，且經由運動向量的投影有重疊時，即產生時間的連結。

2. 標記階段

在這個階段中，我們要對 ST-RAGs 上的每一個節點給予一個標記。這裡所謂的標記是指將節點所對應的分水嶺區域分類為「前景」、「背景」或者是「不確定」區域。詳細的方法敘述如下：

a. 初始標記

在前處理階段建立好 ST-RAGs 後，接下來在標記階段中將為 ST-RAGs 中的每一個節點賦予一個初始標記。

初始標記可分為三類，其定義如下：

$$K = \{"F","B","U"\}$$

其中 F、B、U 分別代表前景的區域、背景的區域和未確定的區域。 初始標記的給定可分成兩類： 第一種只有發生在最低解析度的 ST-RAG 上。此類初始標記可由使用者給定，或採用某些自動的方法，在關鍵視訊框(key frames)設定。 第二種則是較精細解析度的 ST-RAG 標記，可由其上一階層中較粗略解析度的 ST-RAG 標記傳遞衍生而得之(label propagation from coarser level)。

(i) Label Propagation from Key Frames

Label propagation from key frames 可分為自動和半自動兩種。我們去年發展之方法[5]即屬於自動的方法：把視訊中的每一個視訊框都當作是關鍵視訊框，利用相同的物體具有相同的運動向量的假設自動地進行視訊切割。而分割完的結果就可以拿來當作 ST-RAG 的初始標記。

半自動的方法則是以一互動式工具對視訊中幾個關鍵視訊框先做分割[6][7]。關鍵視訊框分割好後其所對應的 ST-RAG 的節點亦已完成初始標記，對於其他非關鍵視訊框的節點，則必須由關鍵視訊框傳遞法來取得。其方法解說如下：對於一個未確定標記的區域，我們先定義兩個變數 l_b 和 l_f。其中，l_b 表示前方之最近關鍵視訊框的初始標記，l_f 表示後方之最近關鍵視訊框的初始標記。先針對 l_b，將未確定標記區域中所有的像素依區域運動向量投影至 l_b 上，如果投影後大部分的區域都落在標記為 F(或 B)的地方，則未確定標記區域亦標記為 F(或 B)，否則則標記為 U；同樣的針對 l_f 亦執行相同的動作。

(ii) Label Propagation from Coarser Level

低解析度(較粗略)的 ST-RAG 在完成 MAP 最佳化分割後，其標記將傳遞至高解析度(較精細)的 ST-RAG 。舉個例子，一個在低解析度時被標記為 F 的區域，這區域在高解析度時被細分為三個區域，則這三個區域也會被標記為 F。

b. MAP Labeling

在 initial labeling 之後，我們要進一步的將標記的結果最佳化，所謂的最佳化是將 ST-RAG 中被分類為 U 的標記，進一步的劃分為 F 或 B。在這裡我們是採用最大事後機率(MAP)的方法來進行最佳化。利用貝氏法則(Bayes rule)可以將最大事後機率分解為觀察機率項(observation term)和事前機率項(prior term)，其數學表示法如下：

$$\max_{L} P(L \mid M) \propto \max_{L} P(M \mid L)P(L)$$

where

$M = \{m_i \mid i \in V\}$: the observed motion features

$L = \{ l_i \mid l_i \in K, i \in \mathrm{V}\}$: unknown labels

在這裡我們是以高斯分佈來建立觀察機率項，而以馬可夫隨機場(Markov Random Field)來建立事前機率項。接下來將分別介紹觀察機率項和事前機率項：

(i) 觀察機率項(Observation Term)

我們是以高斯分佈來建立觀察機率項。在計算高斯分佈時會利用到分水嶺區域的運動向量。由於區域的運動向量有前後兩個方向，所以會得到兩個高斯分佈，分別對應到兩個觀察機率，這兩個觀察機率的乘積就是最後我們所要的觀察機率。詳細的做法為：對某一區域 R，先找出與其相鄰且標記相同之區域集合。針對這些區域的集合，我們依照其面積以及其到區域 R 的距離，給予不同的權重。因此我們可以計算出區域 R 的鄰居之運動向量的高斯分佈，如方程式(1)所示。有了平均值和變異數之後，我們可以依方程式(2)來計算區域 R 的觀察機率項。

$$\mu_i(k) = \frac{\sum_{j \in S_i(k)} \omega_{ij} * m_j}{\sum_{j \in S_i(k)} \omega_{ij}},$$

$$\Sigma_i(k) = \frac{\sum_{j \in S_i(k)} \omega_{ij} * [m_j - \mu_i(k)][m_j - \mu_i(k)]^t}{\sum \omega_i} \qquad (1)$$

$$\omega_{ij} = \frac{\mathrm{Area}_j}{\mathrm{BlockDistance}_{ij}}$$

$k \in \{"F","B"\}$

m_i : estimated motion vector of region i

$S_i(k)$: the set of neighboring regions of region i that has the same label k with region i

$$P(m_i \mid \{l_i = k\} \cup \bar{L}_i) = \begin{cases} \frac{1}{K_A} \exp\left[-\frac{1}{2}[m_i - \mu_i("F")]^t \Sigma_i^{-1}("F")[m_i - \mu_i("F")]\right] & ; k = "F" \\ \frac{1}{K_B} \exp\left[-\frac{1}{2}[m_i - \mu_i("B")]^t \Sigma_i^{-1}("B")[m_i - \mu_i("B")]\right] & ; k = "B" \\ \frac{1}{K_u} & ; k = "U" \end{cases} \qquad (2)$$

where $\bar{L}_i = L - \{l_i\}$ and K_F, K_B and K_U are normalization factors for label "F", "B" and "U".

(ii) 事前機率項(Prior Term)

我們使用馬可夫隨機場來建立事前機率項，其公式如下：

$$P(L) = \frac{e^{-\frac{1}{T}U(L)}}{Z} \qquad (3)$$

其中 $U(L) = \sum_{i \in V} \sum_{j \in N_i} V_c(i, j, l_i, l_j)$，$Z$ 是一個正規化的常數。T 反比於目前的解析度，當解析度越低則 T 越大。$U(L)$ 是所有可能的 cliques C 的位能總合。在這裡我們只使用 2-cliques。

在定義 2-cliques 的位能方程式之前，我們要先定義空間和時間的相似量(similarity measure)S_s 和 S_t。空間的相似量 S_s 的定義如方程式(4)：

$$S_s = \begin{cases} H_1(2e^{-\delta H_2} - 1) & \text{where } v_a \text{ and } v_b \text{ are reliable} \\ H_1 & \text{otherwise} \end{cases}$$

$$H_1 = \max(\frac{l_{a,b}}{c_a}, \frac{l_{a,b}}{c_b})$$

$$H_2 = \begin{cases} |v_a - v_b| & \text{where } v_a \text{ and } v_b \text{ are reliable} \\ \infty & \text{otherwise} \end{cases} \quad (4)$$

其中，a 和 b 表兩個空間相鄰區域，$l_{a,b}$ 為 a 和 b 間的距離長度，c_a、c_b 是指 a、b 區域的輪廓長度，v_a 、v_b 是指 a、b 區域的運動向量，δ 是兩個運動向量間差異的權重因子。

時間的相似量 S_t 的定義如方程式(5)。

$$S_t = \max(\frac{r_{x,y}}{Area_x}, \frac{r_{x,y}}{Area_y}) \quad (5)$$

x 和 y 表示兩個時間相鄰區域。$r_{x,y}$ 為 x 和 y 重疊的面積。

依據 S_s 和 S_t 我們定義出事前機率的位能方程式(6)和(7)，其中區域 i 和 j 可以為空間域上的相鄰區域，亦可為時間域上的相鄰區域。

$$V_c(i,j,l_i,l_j) = \begin{cases} k_1 S_s^{\,5} & , l_i = l_j \text{ and } l_i, l_j \in \{"F","B"\} \\ -k_1 S_s^{\,5} & , l_i \neq l_j \text{ and } l_i, l_j \in \{"F","B"\} \\ k_1 \min(S_s^{\,4} - k_a, -\frac{k_a}{2}), & \text{otherwise} \end{cases}$$

k_a : factor for "U" labels

k_1 : normalization factor

(6)

$$V_c(i,j,l_i,l_j) = \begin{cases} k_2 S_t^{\,5} & , l_i = l_j \text{ and } l_i, l_j \in \{"F","B"\} \\ -k_{21} S_t^{\,5} & , l_i \neq l_j \text{ and } l_i, l_j \in \{"F","B"\} \\ k_2 \min(S_t^{\,4} - k_b, -\frac{k_b}{2}), & \text{otherwise} \end{cases}$$

k_b : factor for "U" labels

k_2 : normalization factor

(7)

必須注意的是，S_s 的範圍介於-1 和 1 之間，而 S_t 的範圍介於 0 和 1 之間。當 S_s 或 S_t 值很大時，表示兩個相鄰區域很相似，反之則表示極不相似。

在定義好觀察機率項和事前機率項之後，我們是採 ICM (Iterative Conditional Modes) 演算法[8]來求取最大事後機率(MAP)。

c. 實驗結果

圖 1 和圖 2 所示為採用"Akiyo"和"Foreman"這兩個視訊影片來做實驗的結果。

在"Akiyo"的影片中，以我們的方法全自動分割(未有人為的介入修改)，就可以得到良好的分割結果。

在"Foreman"的影片中，由於 foreman 的頭盔和建築物的牆顏色太相近了，所以全自動的方法不能得到良好的分割結果，而必須以人工介入的方式，對幾個影像的分割結果進行修正。圖 2 是以人工介入修正後，重新進行三維分水嶺分割的結果。

四、計畫成果自評

在本年度的計畫中，我們修改了去年所發展的三維分水嶺分割技術，並將其擴展至多層次的結構，採用多層次的結構可以大幅降低影像分割時所需要的計算時間。

初步的實驗顯示在多數狀況下，我們的方法可以成功地從視訊影片中「自動」抽取出與背景有不同運動的物件。但是如果前景物件的運動不明顯，或前景與背景的顏色分界不明顯時，我們的系統如使用全自動地分割，則分割出來的結果會比較不精確，必須要有人工介入，才能獲得較良好的分割結果。

雖然在視訊監控的應用領域中全自動的要求是必要的，不過對分割精確度的要求比較低。所以我們目前的全自動方法所獲得的結果，應可滿足此類應用的精度要求。但是未來在速度上仍須改進，以滿足即時的需求。

此計畫所發展的視訊物件切割技術除了可應用在視訊監控系統之外，在多媒體方面也有很大的應用潛力。例如在數位典藏的虛擬展示應用上，因為影像型物體相較於傳統的幾何型

物體，可以大為提高物體展示的逼真度，因此是一個很受重視的虛擬展示技術。 然而影像型物體所需的「背景去除」（即物件分割）對分割精確度之要求比較高，且其對自動化的需求不是那麼高，因此可以利用我們的方法，再加上少許的人工修改，就可以得到較高品質的切割方法。將來我們期望能針對影像型物體的特性，找到一個全自動的方法來進行切割，進一步提高本方法之實用性。

參考文獻

[1] D. Wang, "A Multiscale Gradient Algorithm for Image Segmentation Using Watersheds". *Pattern Recognition*, 30(12):2043-2052, 1997.

[2] D. Wang, "Unsupervised Video Segmentation Based on Watersheds and Temporal Tracking," *IEEE Trans. Circuit Systems Video Technology,* 8(5):539-546, 1998.

[3] F. Meyer and S. Beucher, "Morphological Segmentation," *J. Visual Commu. Image Representation*, 1:21-46, 1990.

[4] L. Vincent and P. Soille, "Watersheds in Digital Spaces: An Efficient Algorithm Based on Immersion Simulations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 13, no. 6, pp. 583-598, 1991.

[5] Y.-P. Hung, Y.-P. Tsai, C.-C. Lai, "A Bayesian Approach to Video Object Segmentation via Merging 3D Watershed Volumes," *Proceedings of International Conference on Pattern Recognition (ICPR02)*, Quebec, Canada, Vol. 3, August 2002.

[6] Y.-P. Hung and Y.-P. Tsai, "Trail-Dependent Intelligent Scissors Based on Multi-Scale Image Segmentation," *Proceedings of Fifth Asian Conference on Computer Vision*, pp. 539-544, Melbourne, January 2002

[7] E. N. Mortensen and W. A. Barrett, "Toboggan-Based Intelligent Scissors with a Four Parameter Edge Model," *Proceedings of IEEE: Computer Vision and Pattern Recognition* (CVPR '99), Vol. II, pp. 452-458, June 1999.

[8] S. Z. Li, Markov Random Field Modeling in Computer Vision, Springer-Verlag, Tokyo, 1995

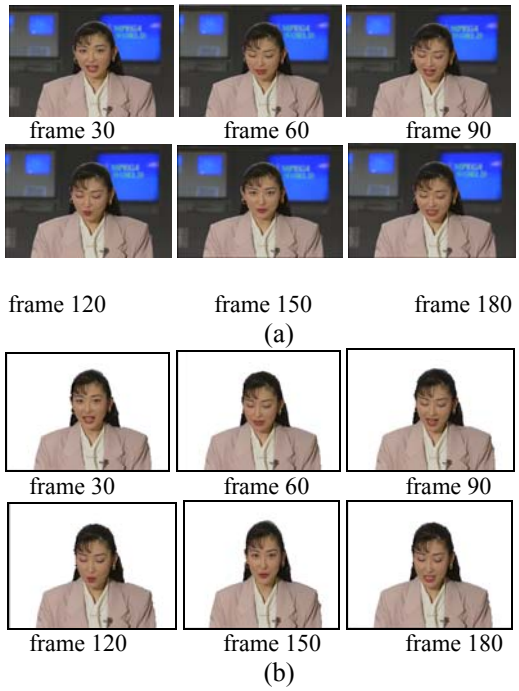

圖 1 Segmentation results of the "Akiyo" sequence without any user adjustment. (a)Original images (b) Segmentation results.

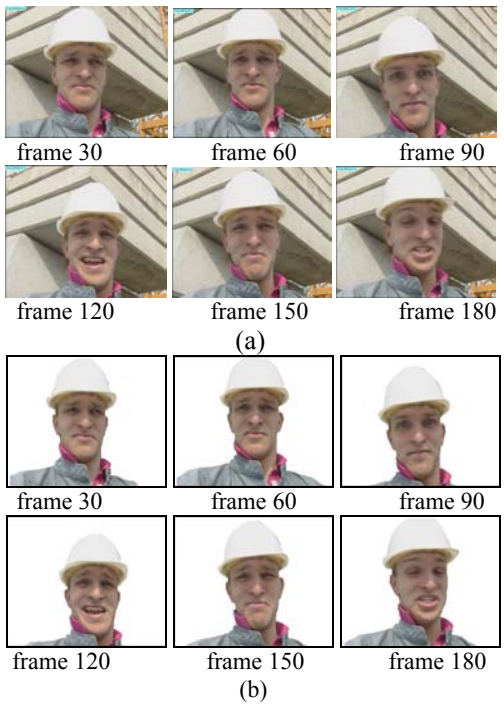

圖 2 An example of segmentation result of the "Foreman" sequence. (a) Original images. (b) Segmentation results.