國 立 交 通 大 學

資訊科學與工程研究所

碩 士 論 文

由影片自動生成「Spot the Differences」遊戲

Automatically Generating a “Spot the Differences” Game from a Video

研 究 生：林立偉

指導教授：林文杰　教授

中 華 民 國 一 百 年 九 月

由影片自動生成「Spot the Differences」遊戲

Automatically Generating a “Spot the Differences” Game from a Video

研 究 生：林立偉　　Student：Li-Wei Lin

指導教授：林文杰　　Advisor：Wen-Chieh Lin

國 立 交 通 大 學

資 訊 科 學 與 工 程 研 究 所

碩 士 論 文

A Thesis

Submitted to Institute of Computer Science of and Engineering

College of Computer Science

National Chiao Tung University

in Partial Fulfillment of the Requirements

for the Degree of Master

in

Computer Science

September 2011

Hsinchu, Taiwan, Republic of China

中華民國一百年九月

由影片自動生成「Spot the Differences」遊戲

研究生：林立偉　　指導教授：林文杰　博士

國立交通大學

資訊科學與工程研究所

摘要

本論文提出一個方法，可利用一段固定視角的影片產生一組「Spot the Differences」遊戲，而我們也是第一個提出自動產生此遊戲的概念。所謂的「Spot the Differences」，也就是一般俗稱「大家來找碴」，是一種以兩張相似的圖片為組合的謎題，遊戲目的是玩家必須找出其中所有的相異之處，然而「Spot the Differences」這遊戲最令人感到有趣的地方在於，玩家有時會對劇烈的變化視而不見，這種現象便是所謂的「變盲」（Change blindness）。我們提出一個有效的方法，除了能有效提高變盲發生率，並且也以物體為基礎定義了「物體視覺注意程度」以及「物體易發意識程度」，前者提供了選擇物體當作相異處的準則，後者則判斷了擺放位置的優劣，而我們便是以這三項作為製作遊戲的基準。最後透過實驗的結果佐證了這三項對於人眼搜尋及辨識相異處的重要性。

Automatically Generating a “Spot the Differences” Game from a Video

Student: Li-Wei Lin Advisor: Wen-Chieh Lin

Institute of Computer Science and Engineering

National Chiao Tung University

Abstract

In this thesis, we propose an approach to generate a “Spot the Differences” game from a video automatically, and it is the first time that this kind of idea be introduced in computer science. “Spot the differences” is a name given to a puzzle where two similar image create from on image are shown side by side, and the player need to find all the differences between them. The most interesting part of this puzzle game is that players sometimes ignore the obvious difference while playing. This phenomenon is called “change blindness”. In order to generate a challenging “Spot the Differences” game, we propose an approach not only increasing the occurrence rate of change blindness but also define the two important terms in this thesis-“the attention rate of object, called ATL” and “the awareness rate of object, called AWL.” Based on the definition of ATL and AWL, we choose which object to be difference and decide the position where to place. In the last, we design two experiments to prove the correctness’ of our assumptions and approach.

致謝

感謝我的父母在我求學路上的信任與支持，讓我能隨自己的興趣及意願學習。感謝我的指導教授林文杰教授兩年來的耐心教導，使我得以了解做研究的態度就如同做人一般，凡事認真並且對自己負責。感謝摯愛姚小姐從我決定報考研究所開始的包容與照顧，感謝GPL以及CAIG實驗室的學長、學姊、學弟、學妹和同學對於本論文的意見及幫助，感謝所有撥空前來參與實驗的交大同學們，感謝你們，因為有你們才能讓這篇論文得以完成。

目錄

圖片目錄

一、緒論

視覺感官上的刺激對於人的日常生活有著相當大的影響力，人類容易受到新的場景、移動物體、變換中的顏色或造型等等外顯的視覺資訊的吸引，並也樂於關注在容易引起視覺注意力（visual attention）的物件。人類喜好接受挑戰，並享受面對難題的刺激感，也因此有許許多多的人投入各式各樣的競技比賽以及形形色色的益智遊戲。同時具備以上特性的遊戲便是「Spot the Differences」。

所謂的「Spot the Differences」，也就是一般俗稱「大家來找碴」，是一種以兩張相似的圖片為組合的謎題，遊戲目的是玩家必須找出其中所有的相異之處。除了是全世界相當受歡迎的遊戲外，也顯示出人類對於找尋相異處的高度興趣。玩家在進行遊戲的過程中需要仔細觀察、比對兩張圖片的細節，因此需要視覺注意力（visual attention）的配合以及視覺短暫記憶（visual short term memory）來儲存配對圖片的局部細節影像以供對照。然而「Spot the Differences」這遊戲最令人感到有趣的地方在於，玩家有時會對劇烈的變化視而不見（failures of visual awareness），而此種現象由Rensink等人[44]在1997年定義為「變盲」（change blindness）。

「變盲」泛指觀察者對於偵測視線內巨大的更動有著令人意外的困難，而這種現象不如名稱所表達是與眼睛有關，而是因為視覺意識（visual awareness）方面的失敗所

引起。在進行「Spot the Differences」的遊戲當中，玩家會以較能引起視覺注意力的地方找起，進而認知這是否是相異處，然後繼續尋找下一個，當玩家疲於關注在較低階層的視覺刺激時，容易忽略景象周遭的變化。另一個產生變盲的原因在於，玩家往往只能專注在單一或少數的物件，因此當視覺場景是紊亂且具有多種目標物時，會使得在物體辨識、知覺以及視覺方面的短期記憶有所限制，而這往往讓他們無法去回想先前的物體狀態。

「Spot the Differences」早期多見於孩童的遊戲書或是書報雜誌，內容以手繪圖為主，近年由於數位影像編輯的普及和便利，以現實照片為底所製作的謎題出現大量在網路及行動通訊平台上。即使這是個人氣相當高的遊戲，目前為止仍然沒有一種簡便的工具提供想製作「Spot the Differences」的人去利用，另外也沒有一種策略讓製作者去參考，究竟要如何依照難易程度來選擇擺放相異處的位置。利用人類視覺的錯亂所產生的影像或遊戲有很多，例如「Camouflage image」和「Necker Cube」，要產生優秀的結果往往需要大量的經驗去累積，而也有若干研究已提出不錯的方法能夠產生接近經驗者的成果[6, 30]。而「Spot the Differences」也是需要一定的製作經驗才可能產生具有一定難度及保留圖片完整性的作品。

在這篇論文，我們提出了一種能自動產生「Spot the Differences」的方法，使用者輸入一段固定視角的影片，我們的方法可根據難易度的需求來產生出一組相似但具有一定挑戰性的圖片，並可依照不同目的替換其中的計算模組來達到最佳化，也可以當作一種簡易的工具讓製造者能夠輕易的使用。為了選擇並擺放相異處，我們也提出一種分析方法計算物體與所處背景之間的相互視覺影響，藉此來定義一個物體的易偵測度方式，而這也是我們最大的貢獻。最後經由一般人直接進行我們所製作出的「Spot the Differences」的實驗結果不但顯示出此論文的方法產生的結果是具有一定難度，並且也證實了我們對於相異處被偵測率的計量方式是合理且有效的，而這也是我們最大的貢獻。

二、相關研究

本論文的研究目的在於能由一段影片自動生成一組「Spot the Differences」，因此本章節主要探討玩家在進行遊戲時所引發的一些視覺現象以及有關的視覺研究；第一小節介紹何謂「變盲」，以及「視覺注意力」與其關聯性，第二小節導出「視覺顯著區域表示圖」（saliency map）的理論根據和計算模組，第三節提出了「偵測相異處」（change detection）和「視覺意識」（visual awareness）對於本論文研究方向的影響性，最後章節則試著討論我們的做法與其他「影片精簡化」的異同處。

2.1 變盲

視覺系統（visual system）對於人類感知方面扮演著重要的角色；當一個未知、複雜的景象映入眼簾，對視網膜造成神經刺激後將訊息傳送至大腦，爾後經過中樞神經一段簡單、近乎全自動的處理，對於整體結構以及內含元素便可產生大致的意識（awareness），而這一切往往只需要簡單的一瞥，即便是相當複雜的場景也可於短暫時間內有相當程度的了解。雖然視覺是人類對於處理外在信息最主要的管道，但不幸的是，我們所見的大部份都是幻象[33]。透過適當的人為操作往往可使觀察者忽視巨大且醒目的物件。例如，著名的「看不見的大猩猩」實驗（Invisible gorilla test）[47]，當觀

察者凝視著傳球的人群的同時，會完全忽略從中央走過了一隻由人假扮的大猩猩。這是所謂的「不注意視盲」（inattention blindness），人對於突然出現在凝視畫面中央，即便是相當明顯但預想不到的目標視而不見[29]。類似情況尚有：沒有注意到與你交談的人途中有過交換身分[48]、無法發現巨大物體的消失或出現[34, 44]這些相關的視覺屏障（visual failures），都可屬於「變盲」（change blindness）的型別[43]，而「Spot the Differences」便是以產生變盲現象來達到欺騙玩家的目的。

早期研究變盲主要是以觀看條件當作控制變因來引起所謂的「誘發式變盲」（induced change blindness），而依據控制變因又大略分為兩種方向；第一種關注於視覺記憶的影響（visual memory），給予受測者一排的圖片或字母序列，爾後顯示一段「間歇刺激間隔」（interstimulus interval, ISI）來阻隔觀察的延續（ISI通常為一個黑屏或是灰屏中心點有個矯正視角的標誌，例如「＋」、「・」），接著再顯示有一個元素遭更改的序列，研究發現「ISI」存在時間越長，受測者對於偵測相異的能力就越低落[36, 38]。另一派的學者將焦點放在眼睛的移動，研究受測者在對圖片做掃視（saccade）的過程中，偵測相異處的能力[4, 7, 31]，不論是圖片、電影場景抑或是文字，皆可發現偵測率相當低，除非相異處是發生在掃視時關注的物體[7]。在這個階段的方向仍侷限在「實驗室內部」，受測者所關注的多為符號或字母，在’96年有個研究指出受測者在眼睛移動時，無法發現真實照片中的巨大改變[12]，比如說有一半左右的受測者對於頭部互換的兩相鄰牛仔完全沒有注意到，這研究讓整個趨勢往更多元方向前進。

變盲領域另一個重大的發展在於視覺注意力對於發現變異處的重要性；雖然誘發式視盲的方向分為利用「ISI」阻隔觀察的連續性，以及著眼於掃視造成的焦點轉移，但後者在眼球掃視途中，視網膜接受到的是圖片模糊化後的影像，其實也可視為一種「ISI」，爾後也由Resink等人所進行的四個實驗來證實：影響這兩種誘發式變盲的實驗中，左右受測者發現到相異處的關鍵在於視覺注意力（visual attention）[44]。

Resink所採取的方式是類似利用一段ISI來阻隔觀察連續性，但只以一個純粹的灰色全屏幕而沒有矯正用標誌，並使用一組原始和修改過後的真實照片當作比對目標，

而這實驗稱之為「忽隱忽現」（flicker）實驗，並利用兩組受測者做交差測試，結果發現較易引起視覺注意力的相異處也容易被偵測，反之則否[44]。而Kelly等人將實驗方法作更進一步的延伸，方法為將照片以上下顛倒呈現，發現若是相異處與照片場景有較高關連性以及關注度（比如熱氣球競賽中的「熱氣球本體」，相比較則為「熱氣球表面的塗裝」），偵測率有相當程度的滑落[21]。

2.2 視覺注意力與視覺顯著區域表示圖

由上節後段所提出的論點可得知：場景中較易受到關注的物體，受測者對其分析、比對的優先權會較高，那又是甚麼影響物體被關注到的可能性？換言之，人眼對於視野內的哪些元素會優先去觀察？以這個方向去探討的科學家中，Bela Julesz博士是最早利用「視覺搜尋」（visual search）的概念來探討視覺系統處理外在刺激的架構[18, 19, 20]。他的研究指出，單一符號放置在一群相異卻又相似的元素群裡，有些容易跳脫出來吸引人們的注意力，而有些會隱沒於其中；比如字母「L」處在越多的字母「T」當中，偵測反應時間會越長，而處在「+」裡頭，反應時間都相當的短暫。以視覺搜尋為基底，爾後更發展出所謂的「注意力特徵整合論」（feature integration theory）[52]，研究發現具有某些特定特徵的物體放在一群相異物體；例如下頁圖1要找出平行線段，無論物體數量的多寡，具有特徵的物體（平行）是可在短時間被發現，而具有這種特性的視覺搜尋方式稱為「平行搜尋」或是「特徵搜尋」（parallel search or feature search）[1]；相對的，以下頁圖2為例，若是結合方向性與顏色這兩種特徵，會發現平行線段便不再是那麼容易被發現的，而這種則稱為「組合搜尋」（conjunction search）[52]。

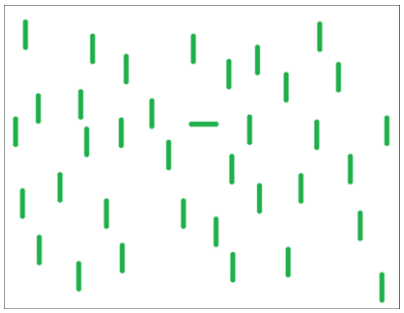

圖 1
特徵搜尋（feature search）

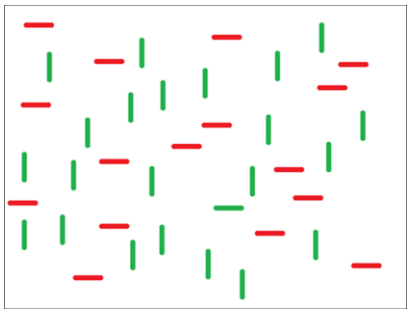

圖 2
組合搜尋（conjunction search）

視覺系統在進行所謂「特徵搜尋」的速度是相當快，並且能搶先在注意力被視野內物體所吸引過去之前，迫使視線往某些點位看去。而根據這種單純依靠低階層級特徵來吸引注意力的特性，衍生出一種「視覺注意力的可計算模組」（Computational model of visual attention）[15, 23]，計算出的結果為視覺顯著區域表示圖（saliency map）。視覺顯著區域表示圖的應用面相當廣泛，除了計算何處的視覺刺激較為強烈，也可預測人類視野的轉移[16]及物體辨識[51, 53]。這種由底而上的計算模組，雖符合人眼對於視覺刺激的理念且適用性高，但真實世界中有些情況需考量到高層級視知覺的影響力，而在本論文裡我們提出了一種計算模組結合高層級的視覺資訊及視覺顯著區域表示圖，藉此判斷物體的顯著程度和被意識率。

2.3 物體辨識與視覺意識

前兩小節提到，對於受測者偵測到相異處的關鍵在於有沒有將注意力放在該處，但是要將注意力轉移至察覺相異處的階段是需要視覺意識的產生，而這也由Fukuba學者等人利用高磁場功能性磁振造影儀（fMRI, functional magnetic resonance imageing）證實，進行「Spot the Differences Game」過程中，與一般人最大的差異在引起視覺意識的腦神經區域有大量的活動[9]。要如何從視覺刺激的階段進入意識到有相異的存在？換言之，「視覺注意力」以及「視覺意識」中間的關聯性是以甚麼形態存在。以「變盲」和「不注意視盲」的觀點來看，意識必須要觸發注意力才能產生[29, 44]，雖有

研究指出，觀察者能同時對一區域內物體產生視覺注意，卻只會察覺到一個相異處[43]，但在「dual-task」實驗中可發現，受測者不需要注意力的協助便可意識到視野邊緣一閃即逝的物體[3]，更有研究指出，視覺意識是可以發生在選擇性注意力（attentional selection）階段以前[26]。對於意識分離在注意力之外，尤其是「dual-task」實驗的現象，有學者指出當視覺神經元受到非當前注意目標的刺激後，會進行一種稱為「前饋」（feed-forward）的行為來取得它們在中樞神經的選擇權[24]。但這種現象是會受到一些狀況的干擾而不顯著，在「dual-task」實驗中，視野邊緣一閃即逝的圖片若非明顯的物體，而是單純無意義的低層級視覺刺激，受測者的感知程度會有顯著的下降[28]。

2-4 影片精簡化

我們的做法也可視為將影片精簡化（video synopsis），分析整段影片後盡可能放入越多的具有高視覺注意力物體在最後的圖像裡，我們可以合理的認為這些物體除了符合影片場景的意義，也代表影片裡較為關鍵片段。早期影片精簡化主要用途在於將影片內過於緩慢的觀測行為加快速度以利於比較及分析，但純粹的快轉有時會讓影片失真並喪失一些重要的畫面。因此有相關研究試著保留具有較高的行為發生的畫面，盡可能不漏掉會引起興趣的片段[32, 37]，或是挑出數段簡短並能代表影片內容的片段作接合的動作[49]。另一種做法是去除了時間軸，將影片精簡成數張關鍵畫面[22, 55]。以上的製作過程中，皆不會對影片內容物做編輯修改的行為，有別於這兩個方向，另有類似馬賽克拼貼的做法將影片中的物體合成一張圖片[14, 35, 39]。而最早以物體為精簡化考量目標的做法是由Acha等人以及另一組研究團隊，Kang等人在'06年提出[13, 40, 42]，他們試著保留影片中移動物體的完整軌跡，讓原本處於不同時間軸的物體放置在同一畫面內，減少畫面內無移動物體處以確實的縮短影片長度。

三、研究內容與方法

這個章節裡會詳細介紹本論文的做法，首先在第一小節提出演算法所採取策略以及整體架構，接著分項細說各步驟，第二小節說明如何處理輸入的影片資訊，接著在第三節細說對於擷取出的移動物體所採取分析過程，第四節挑選圖片配對以及作為相異處之物體，而後在第五節說明相異處的類型及依據。

3-1 策略以及演算架構

玩家在進行「Spot the Differences」最主要行為是「偵測相異處」（change detection），但因為發生「變盲」（change blindness）現象使得偵測上有所困難。若要提高遊戲的難度，使得玩家無法於短時間內找出相異處，需要達成兩個最主要目標：

一、降低相異處的被偵測率。

二、提高變盲的發生頻率。

有若干研究已提出一些建議，例如：以人工方式分類相異處的難易[44]、延長「間歇刺激間隔」（ISI）的持續時間[36, 38]、將觀看圖片上下顛倒[21]，但並不實用；好比「ISI」無法真正實做於遊戲系統裡，而上下顛倒會失去遊戲性，因此在這裡我們提

出了三種實際且有效的策略：

一、相異處應不引起較大的視覺注意。

二、根據「視覺意識」（visual awareness）對於偵測相異處的重要性，讓相異處與圖片上視線熱點的距離不要太近。

三、當場景紊亂，玩家容易受低層級的視覺刺激吸引而忽略「相異處」的存在。

為了兼具遊戲性及困難度，我們不使用背景裡的材質貼圖或是晃動樹葉及雜草等一般人會認為無意義的地方來當作挑選相異處的依據，而是以完整並與整個場景有關連性的物體為主要挑選原則，因此從輸入影片擷取出移動的物體能夠符合我們的要求。以影片當作輸入有許多優點：場景內容豐富度比單張照片高、物體擷取較準確和方便、利於後製，且近年來高畫質錄影技術的普級化讓我們相信利用影片會是最好的選擇。

我們的做法分成前半階段的資料處理階段，以及後半的分析與製作，當進入後半階段使用者的控制（難度、相異處數量）進來參予決策；流程示意圖於下頁圖3，文字敘述如下：首先將輸入影片轉換成間隔一段時間擷取出的鏡頭畫面組，取得所有鏡頭畫面的視覺顯著區域表示圖，接著分離出移動物體的資訊，與所在鏡頭的視覺顯著區域表示圖作結合，建立一組有關移動物體的清單，裡頭儲存除了基本資訊（鏡頭畫面代碼、顏色…等），還有最重要的「物體視覺注意程度」以供後製時選擇主畫面以及擺放相異處時做參考。後製階段除須符合使用者一開始設定的條件還會考量到編輯後的自然度，我們希望避免出來的成果有過多的人工效果。在這兩個基準底下並根據先前提出的三個策略，接著從移動物體與所屬鏡頭畫面來決定主要畫面，並挑出一些適合做相似化的副畫面，透過比較主畫面與副畫面之間的視覺顯著區域表示圖、色差、移動物體配置，最後得到最終的配對。

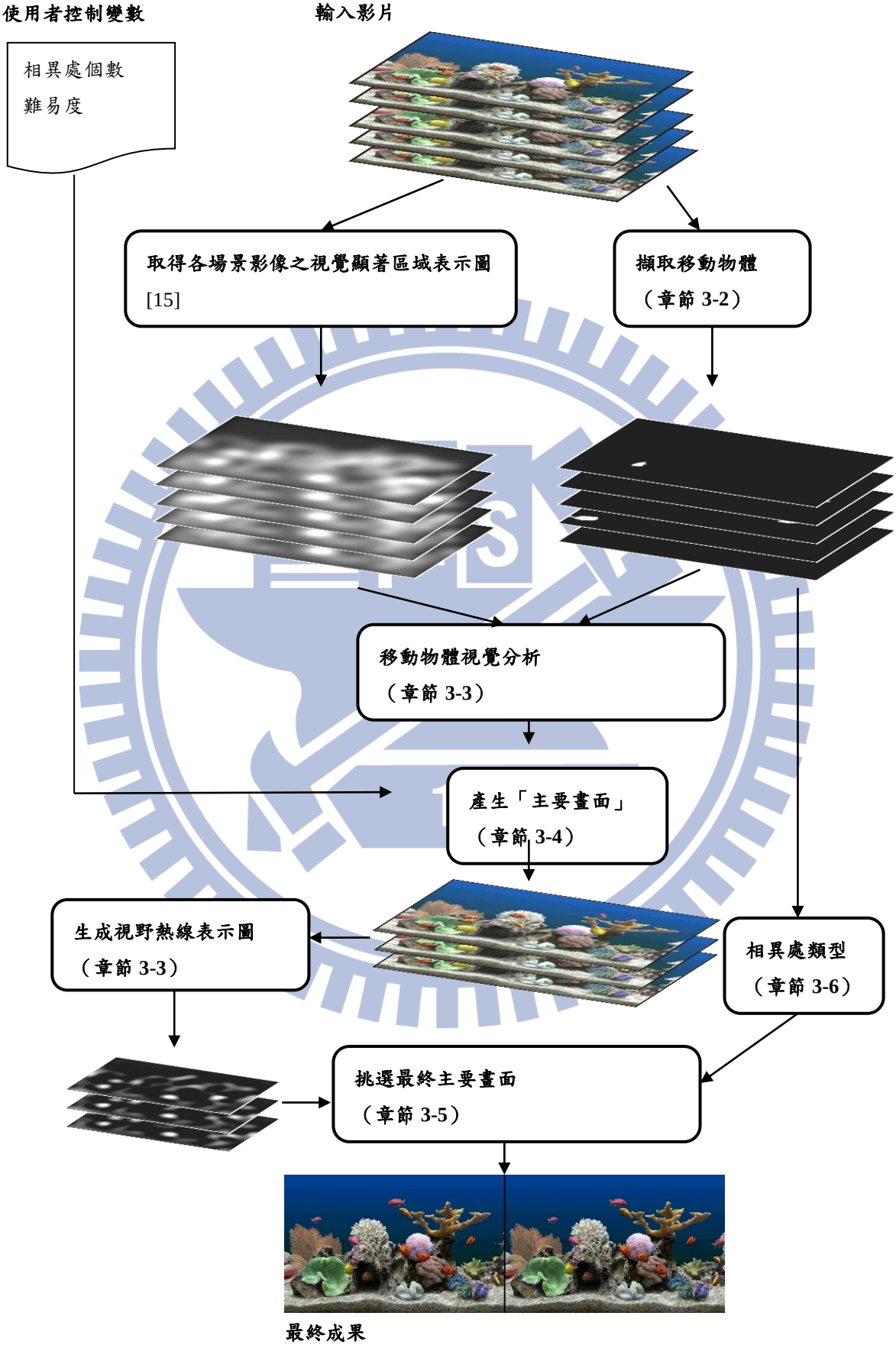
使用者控制變數
輸入影片
相異處個數
難易度
取得各場景影像之視覺顯著區域表示圖
[15]
擷取移動物體
（章節 3-2）
移動物體視覺分析
（章節 3-3）
產生「主要畫面」
（章節 3-4）
生成視野熱線表示圖
（章節 3-3）
相異處類型
（章節 3-6）
挑選最終主要畫面
（章節 3-5）
最終成果

圖 3　本論文作法流程圖。

3-2 擷取移動物體

為了產生優秀的「Spot the Differences」，我們必須先面對一般人工製作時所要面對的幾種情況：由於相異處是場景中的物體，並且須與場景有一定的關連性，因此為了後製時的方便性以及成品的合理性並降低切割不完整所造成的偵測誤會，物體辨識在前置階段相當重要。另外背景填補也是需要去克服的部分，因為相異處對物體做移動或是消除的動作，這時移除後的區域需要回填正確且完整背景資訊，否則也會產生偵測上的誤解以及增加最終成品的不自然處。我們物體偵測參考自Yael Pritch等人於'08年提出的方法，並依照我們的需求做適當的更改及優化[41]

一般靜態圖片的物體偵測方法的分野在於有無資料庫的支援，圖片透過資料庫做訓練及學習的成果可以近乎於人工標示[11, 46, 54]，但往往受限制於資料庫的內容完整性以及物體標示的精確度，並且需要相當長的計算時間。相對來說，無資料庫的做法的優勢便是較短的計算時間、高可攜性、對於特定狀況的辨識能力有時會較佳[5]。當輸入資訊為影片而非靜態圖片時，最大差異在於場景資訊具有空間以及時間上的連續性，而我們的做法便是結合靜態圖片無資料庫作法以及動態影片所帶來的優勢，可以有效的偵測出場景中的移動物體。

相對於移動物體，影片中的背景資訊是不變的，因此利用空間的連續性，我們的物體辨識方法首先由背景建構開始。在視角固定的影片中，短時間內背景的變化量基本上是微小到可以忽略，變化較明顯的情況通常是發生在環境光照變化、背景大型物件的改變，而這些需要較長時間的記錄才有較高的機率發生，並且可以切割簡化為分段的短期影片來解決。至於短期影片的背景建構，一般快速且有效的方法是考量整段時間內背景資訊的分佈狀況，根據前面的假設，背景資訊通常是不會有所變動並且出現頻率相對較高；比如給定一個像素位置，透過統計以及計算分析一段時間內的顏色值，可判斷具有較高比率的顏色便是背景資訊。而這裡我們對於每張鏡頭畫面都會建立各自的背景圖，盡可能的減少自然環境的不確定性，對於畫面中每個位置的像素的顏色值做統計分析，單純取中位數當作所求的背景資訊。

圖 4

擷取移動物體流程：由左至右為原圖、原圖之背景圖、去除背景後的移動物體資訊。

一旦決定了影片背景，便可假定每張鏡頭畫面中不同於背景資訊的便是所求的移動物體，這裡我們使用簡化版本的「Background cut」[50]方法來將移動物體從背景中擷取出來。傳統去背景的計算方式是比較當前圖片與背景的色差，並假設一個門檻值以藉此判斷是否為前景物，結果優劣取決於門檻值設定與輸入影像的適應性以及色差的判斷方式，但單純判斷色差是無法有效處理大部份的狀況。「Background cut」除了色差的判斷，另考量到圖片與背景梯度（gradient）差異性，梯度不一致的區域，即使顏色相近但有極大可能是前景物所造成的運動邊界（motion boundaries），最後利用「min-cut」演算法來取得切割前景物與背景的最佳解，以下詳述做法。

令I為當前處理的鏡頭畫面，B為該鏡頭的背景，N為I內所有相鄰像素的配對集合。給定一方程式f來標示像素r為前景物體（$f(r)=1$）或是背景資訊（$f(r)=0$），利用Gibbs的能量式[2]得到最佳解：

$$E_o(f) = \sum_{r\in I} E_1(f(r)) + \lambda \sum_{(r,s)\in N} E_2(f(r), f(s)) \qquad (1)$$

$E_1(f(r))$是處理色差的能量式，$E_2(f(r), f(s))$計算相鄰像素間的梯度及對比度的便化量，λ則是使用者控制用變數。

給定$d_r = \|I(r) - B(r)\|$為計算鏡頭畫面I與背景B的色差，而E_1根據像素r的標籤為前景（$f(r)=1$）或是背景（$f(r)=0$）而有不同的計算方式：

$$E_1(1) = \begin{cases} 0, & d_r > k_1 \\ k_1 - d_r, & \text{otherwise} \end{cases} \qquad (2)$$

$$E_1(0) = \begin{cases} \infty, & d_r > k_2 \\ d_r - k_1, & k_2 > d_r > k_1 \\ 0, & otherwise \end{cases} \quad (3)$$

算式(2)代表意義為假定像素r與背景間色差大過一個門檻值，便可合理認為將像素r認為前景物的標記是正確的，因此其能量逞罰為零，相反的未達門檻值，能量逞罰則是與門檻值的差異。算式(3)第一項的代表意義為，若像素r被標記為應當是背景資訊，與背景色差應當很小，只能允許光影變化所帶來的微小差異，因此當色差過大，便可判定此標記（0）絕對是錯誤的，對於能量式的逞罰設定是無限大。而算式(3)餘下第二和第三項次與算式(2)所要表達意義相似，當被標記為背景資訊的像素與背景色差小於門檻值，給定的能量逞罰為零，反之則是與門檻值的差異。k_1以及k_2是由使用者定義的的門檻值，這裡我們參考了[41]的做法，各設定為$30/255$和$60/255$，但是可以根據使用者的輸入影像特性做適當的微調，比如影片內光影變化劇烈，門檻值的設定可以調高避免判斷式過於敏感。

計算E_2的目的在於單純以顏色作為區分前後景的依據時，難免會因為前景物部分顏色與背景物的過於相近，或是背景光影變化這兩種因素使得結果過於破碎，此時可從兩個方向下手：前景物本身的輪廓線，以及與背景該處梯度值的差異性。前景物輪廓部分的梯度變化量理應較高，但有可能發生將背景區塊視作前景物的延伸，因此除了考量前景物所在畫面內的梯度變化以外，可與純背景圖的梯度值做比較，可以合理假設兩者差距較大時才會是前景物的輪廓線。E_2計算方式參考自[46]：

$$E_2(f(r), f(s)) = \delta(f(r) - f(s)) \cdot \exp(-\beta d_{rs}) \quad (4)$$

其中$\beta = 2 < \|(I(r) - I(s))\|^2 >^{-1}$是權重係數（weighting factor），$<\cdot>$會去計算整張圖所有相鄰像素色差平均值。$d_{rs}$則是當前處裡的鏡頭畫面梯度值，並依據背景圖梯度值做遞減，詳細如下頁所述：

$$d_{rs} = \|I(r) - I(s)\|^2 \cdot \frac{1}{1 + (\frac{\|B(r) - B(s)\|}{K})^2 \exp(\frac{-z_{rs}^2}{\sigma_z})} \quad (5)$$

等式右方第一項次為判斷前景物的輪廓，後項次則是與背景梯度做比較，其中z_{rs}是計算前景物與背景的差異性：

$$z_{rs} = \max(\|I(r) - B(r)\|, \|I(s) - B(s)\|) \quad (6)$$

用意在避免前景物輪廓恰好位於背景的梯度變化量大的區域，藉由判斷輪廓線內外是否有與背景色差較大的部分來避免這種情況的誤判。K以及σ_z根據[50]的建議，分別設定為5和10。

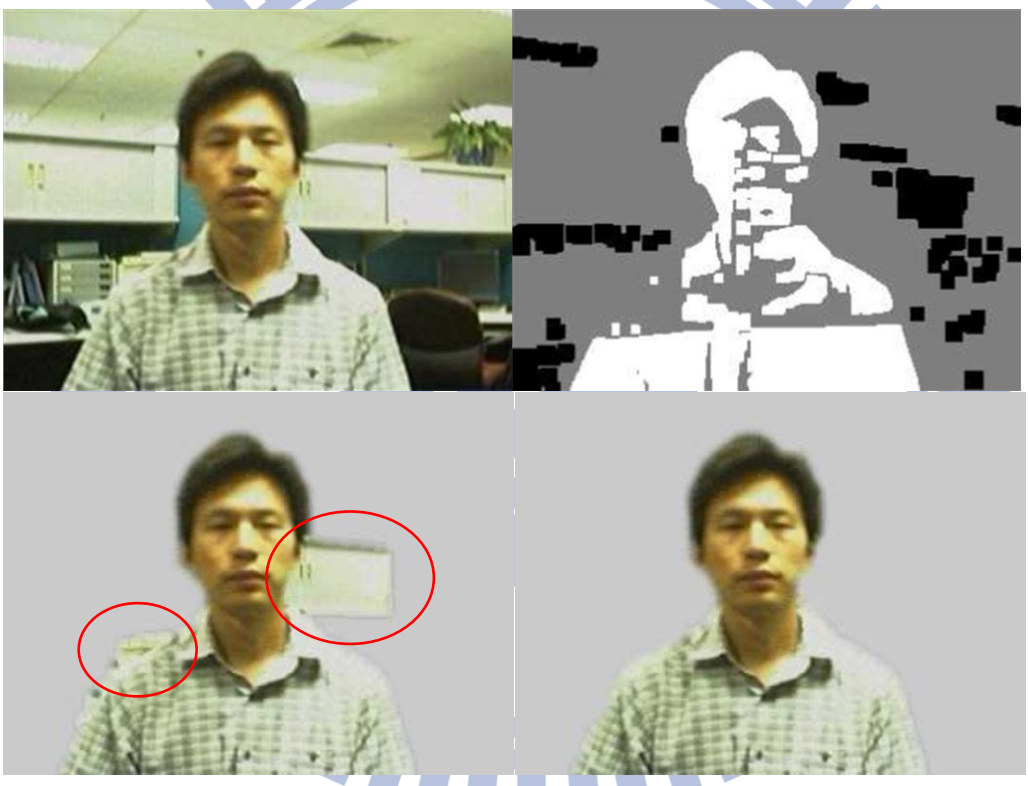

圖 5（此圖片組取自[50]）

左上為影片序列裡其中一格畫面，右上為單純比較前景與背景的色差，會選擇讓此階段只定義部分前景區域的原因在於，接下來判斷輪廓時會做回填的動作，左下圖則為考慮畫面內梯度變化，可看出紅圈內為誤判，在比較前後景間梯度值差異得到右下的最終結果，紅圈部分已被認定為背景區塊。

3-3 分析物體與背景間視覺影響現象

我們挑選物體做為相異處是以先前所提出的策略一及策略二為準則，因此必須明確量化這兩種狀態才能比較出不同物體間做為相異處的適當性。在這個小節裡，會定義以下兩種特徵：移動物體相較於所處背景的物體視覺注意力，我們稱之為「物體視覺注意程度」（the attention level of an object），以及背景資訊如何誘導觀察者察覺到非注視目標的物體，並制定出「物體易發意識程度」（the awareness level of an object）。量化「物體視覺注意程度」以及「物體易發意識程度」的方式是由建構低層級視覺刺激開始，接著和物體資訊做結合，並考量到與環境間的相互關係。

在此先說明我們是利用何種計算視覺顯著程度的模組；「視覺顯著區域表示圖」（saliency map）的計算方法有很多種，由於我們主要輸入影片內容較為紊亂，大多沒有明顯的目標物體，計算的方式傾向全域性的視覺顯著指標，因此使用的是[15]的方法，並利用來計算移動物體的顯著程度。有些視覺顯著區域表示圖的對於輸入圖片的量化方法有考慮到高層級的特徵，例如環境中物體的存在性[10, 51]，雖然這些方法對於醒目物體的偵測能力相當優秀，但輸入圖片中場景往往限制在較單純的情況，若是圖片情景較複雜則以上提出的三種方法相差不遠，因為計算方式都是基於同一原則：視覺注意力是由低階層的視覺刺激所引起，著重在色彩、對比等底層特徵。在我們的演算法模組裡這部分是可以做替換的，對於某些特定輸入可以利用不同的視覺顯著區域計算模組，但普遍來說並沒有Itti等人提出的模組來的泛用。

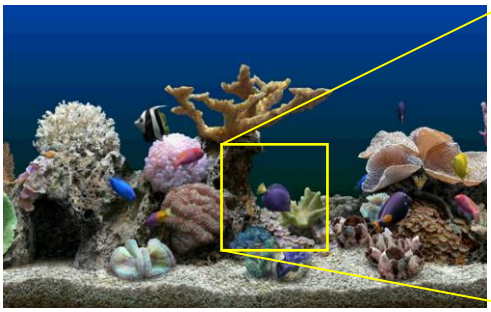

圖 6

圖 7

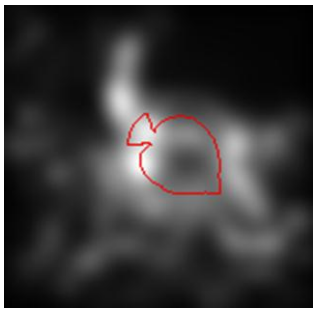

圖 8

圖 6 為當前處理之物體所在畫面 I，圖 7 中的魚為移動物體 O，圖 8 為視覺顯著區域表示圖 S 的局部區域，圖中紅線所包圍區域為物體 O 的區域遮罩 R。以圖 6 中間偏下方的魚為例，與顯著區域表示圖 8 比對可發現大部分的視覺刺激較為低落，但人眼仍會被尾鰭部分高視覺刺激而吸引過去，進而意識到整體，因此在定義此條魚的視覺顯著值 ATL 會以尾鰭周圍為主。

傳統視覺顯著區域表示圖所計算的視覺注意力是全域性，主要目的在研究一般人眼受到視覺刺激程度的強弱[16, 17, 27]。低層級視覺注意力對於辨識物體有相當程度的幫助[45]，無論物體於畫面所占面積比例，具有高視覺注意程度的物體也較容易吸引觀察者的目光[8]，因此有時在計量視覺顯著程度是必須考量到高層級的視知覺（visual perception）[10]，但目前卻沒有一種系統去明確定義物體與背景之間的視覺影響力，包括低層級視覺刺激所引起的注意程度，以及視覺意識所引發物體偵測。而在這小節餘下的部分我們會先說明如何制定物體的「物體視覺注意程度」，視覺刺激對於物體識別的重要性，接著定義「物體易發意識程度」。

當物體具有一部分高視覺刺激的區域，即便剩餘部分是相對較低，對於觀察者而言會因為吸引過去的第一眼而查覺到該物體[53]。因此我們定義物體的視覺注意程度值是取其所包含區域內視覺顯著值統計上的高標，藉此避開大部份數值低的區域（實例請見本頁圖6至圖8）；給定物體O，I為當前處理之物體O所在的鏡頭畫面（frame），R為物體O的區域遮罩，S為I的計算視覺顯著區域表示圖，r_S為視覺顯著區域表示圖所包含的像素，物體的視覺注意程度ATL（the attention level of an object）定義如下：

$$a = \text{avg}\big(S(r_s)\big), r_s \in S \cap R \tag{7}$$

$$ATL(O) = \text{avg}(S(r_s)), r_S \in S \cap R, S(r_s) > a \tag{8}$$

算式(7)為計算出視覺顯著區域上位於物體遮罩內所有像素值平均a作為基準值，物體的視覺注意程度（ATL）便是統計數值高於a的像素分布。

要定義「物體易發意識程度」前必須先定義視線熱點的計算方式，Itti等人根據傳統視覺顯著區域表示圖發展出一套視覺搜尋系統[16]，藉由判斷人眼可能會關注的地方來預測搜尋的順序。傳統視覺顯著區域表示圖是擷取出圖片中三種容易引起視覺注意力的特徵，顏色、對比、區域方向變化，分別計算出各自從局部到全域性各層級的特徵差異變化表示圖（feature contrast map），圖中具有較高數值的區域代表該特徵於該處引發視覺刺激的程度較大，較容易吸引注意力，最後做統合累加各特徵求得最終的視覺顯著區域表示圖。而要計算出何處為視線熱點，在[16]裡頭，計算過程與先前由他們所提出視覺顯著區域表示圖的唯一差別在於，對於三種特徵變化表示圖的每一層級都以一個寬度等同於影像寬度的「高斯差」（Difference of Gaussian）濾波器做捲積（convolution），此作法用意在增強全域中相對較高的視覺注視區域的頻率。由於經濾波器處理過的每一層特徵變化表示圖最後只是單純的做累加，原作法分層分項的用意在於使用者可以依照需求，對欲強調的特徵元素乘上增幅的係數或是加上其他自定義的特徵項目，而這裡我們並沒有這方面的需求，因此直接對最後的視覺顯著區域表示圖經由濾波器做捲積。

即便物體並未受到關注，仍可能被意識到其存在，會導致玩家在比對當前注視目標的同時查覺到周圍物體有所異狀，根據這假設，我們認為距離影像內的視線熱點來說較遠的物體，當作相異處的候選較佳。以下說明視野熱點表示圖計算過程；DoG為濾波器，M為S經過捲積後的視野熱點表示圖，M的計算方式如下：

$$DoG(x,y) = \frac{1}{2\pi\sigma_{ex}^2} e^{-\frac{x^2+y^2}{2\sigma_{ex}^2}} - \frac{1}{2\pi\sigma_{inh}^2} e^{-\frac{x^2+y^2}{2\sigma_{inh}^2}} \quad (9)$$

$$M_k \leftarrow |M_{k-1} + M_{k-1} * DoG|_{\geq 0}, M_0 = S, k \text{ from 0 to 5} \quad (10)$$

算式(10)經過多次的計算取得局部範圍內視覺刺激較為強烈的地方，以此得到最終的視野熱點表示圖M，$|\cdot|_{\geq 0}$用意在除去負數值並執行標準化，標準化後的數值落在0到1

之間。原作法的搜尋法為全域性，因此濾波器遮罩大小設定為影像寬度[16]，但我們著眼點在於「視野範圍」內的視覺刺激所造成的視線轉移和視覺意識的產生，因此濾波器的寬度會根據最後輸出影像做調整，而我們的實驗中設定為影像寬度的三分之一，σ_{ex}和σ_{inh}分別是2%以及25%的濾波器寬度。接著說明移動物體的易引發意識程度（AWL），r為包含於M內的元素，$ROS(O)$（relative object size）表達移動物體所占面積與所屬畫面I大小之比例，d_{aw}為我們考量物體周圍視野熱點分布情況的範圍半徑，我們設定與濾波器寬度相同，$dist(r,O)$為r與移動物體O邊緣的距離經過標準化後的數值（基值為d_{aw}）：

$$ROS(O) = \left|\frac{\#pixels(O)}{\#pixels(I)}\right|_{\leq 5\%} \tag{11}$$

$$dist(r,O) = \left|\|r-O\|\right|_{d_{aw}} \tag{12}$$

$$AWL(O) = (c\cdot ROS(O))^2 \sum_{dist(r,O)\leq 1} M(r)\cdot(1-dist(r,O))^2 \tag{13}$$

算式(11)裡$|\cdot|_{\leq 5\%}$作用將數值範圍從0%-5%依比例縮放至0-1。算式(13)為計算物體O周圍視野熱點的分布情況，得到的數值即為量化後的「物體易發意識程度」（awareness level of an object, AWL）。由於視野內的視覺刺激會引起視覺意識並也會造成視野的轉移[23, 28]，因此若是物體周圍有視野熱點的存在，勢必會發生視線被熱點吸引過去進而讓視野轉移到物體上的情況發生。視覺意識的產生取決於是否有足夠的視覺刺激進入視網膜，進而觸發視覺神經元的「前饋」現象，而出現在視野內的物體面積越大，其所發出的視覺刺激必定越多，讓「前饋」現象益發明顯，因此在物體的易引發意識程度（AWL）的計算上，會將物體與畫面間大小比例$ROS(O)$考慮進來。此外較大的物體較容易被辨識出來[45]，其中也提到當物體與畫面比例達到一個程度，即使物體不具有高視覺刺激，仍有相當高的機率被辨識出，因此我們將$ROS(O)$標準化範圍設定為0%-5%，而上限值5%是根據[45]的實驗結果所制定。視覺刺激所造成的視覺轉移會因距離當前專注點越近，其影響力越明顯[23]，因而在考量每一單點的視覺刺激，我們認為其與物體距離$dist(r,O)$呈現負相關。（計算AWL示意圖於下頁）。

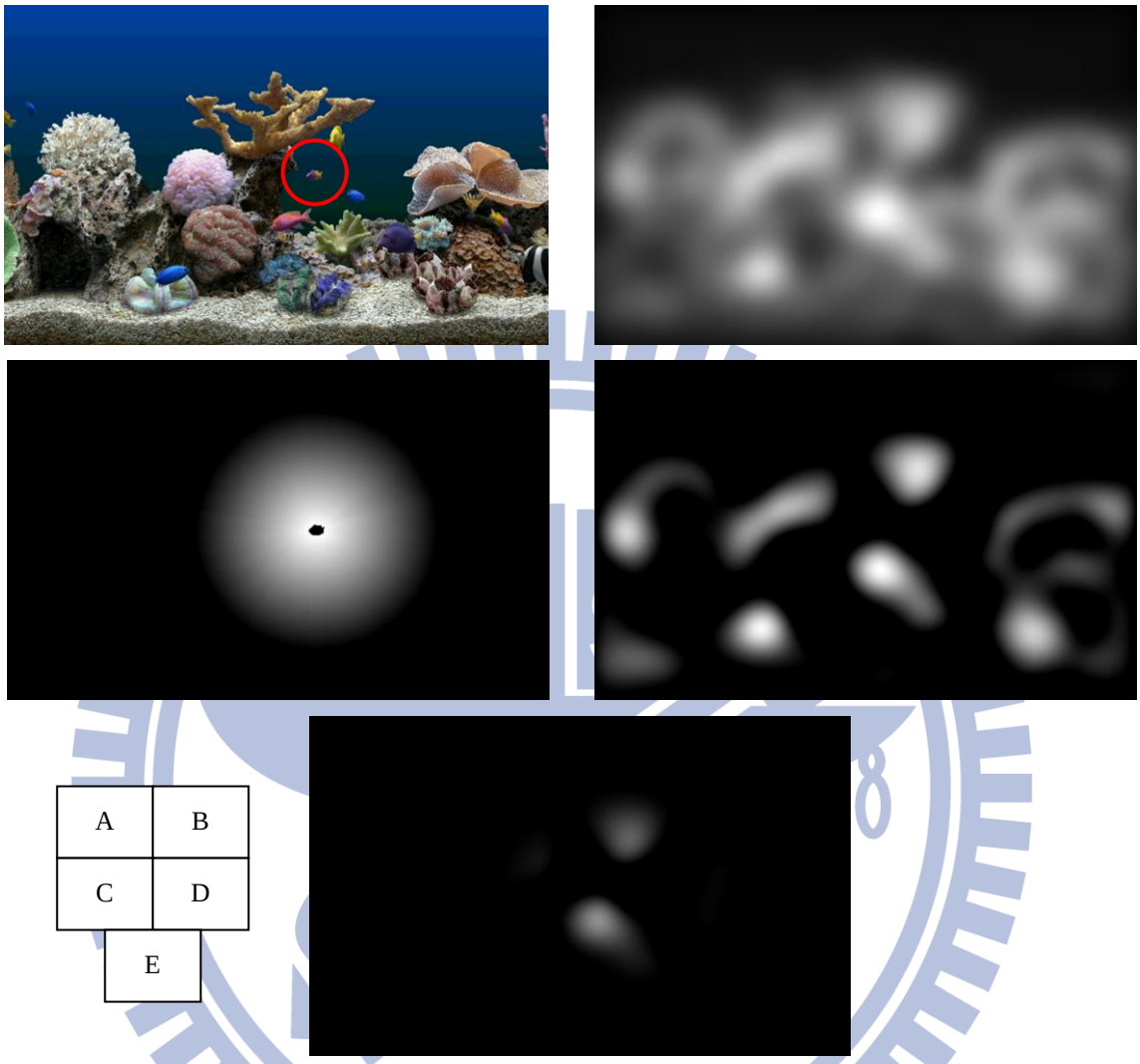

圖 9

（A）物體 O 所在影像畫面（圖中紅圈處）。（B）原圖 A 之視覺顯著區域表示圖（saliency map）。（C）圖中亮度值高低代表畫面中距離物體遠近程度，亮度越高處越靠近指定物體。（D）圖 B 經過多次的「Difference of Gaussian」濾波器 DoG 所得到的視野熱點表示圖 M。（E）圖 C 中亮度作為加權值，結合視野熱點表示圖 M 計算出物體 O 周圍可能引起視覺意識的分布圖，最後統計圖 E 中的亮度值並以物體相對比例 $ROS(O)$ 為加權值算出最終的 $AWL(O)$。

3-4 產生「主要畫面」

一般製作「Spot the Differences」是從單一張影像編輯產生，我們稱此影像為「主要畫面」。而當輸入為影片時，產生主要畫面較為直觀做法為直接挑選一張適合的鏡頭畫面，但這種方式可能會面臨兩個問題，可做為相異處的物體數不足以及高視覺刺激區域過少，前者若缺少有可能迫使我們去利用到不適合的物體做為相異處，後者的不足會導致觀察者容易聚焦在相異處。為解決這可能發生的狀況，我們使用了一種簡單並有效的方式。

我們的做法為：挑選出數張影像畫面，將各影像內的移動物體，加入至同一影像內，而我們將移動物體擺放進影像的方法是採用「Graphcut」[25]（示意圖於下頁圖9）。由於所有物體皆來自同一影片，拼貼所造成的不自然感相當微小，僅需要考量到來自不同畫面的物體是否有重疊的現象，以及重複挑選相同物體，因此挑選影像的方式在這階段扮演相當重要的角色。首先，為符合策略一，選作相異處的物體除了數量要足夠達到使用者要求，物體的視覺注意值必須小於一個門檻值。另外考量到策略三，除了加入選作相異處的物體，會盡可能的加入視覺注意值高的物體。因此挑選結合影像必須在兩種物量數量間找到一個平衡點，首先相異處數量是必定要達到，而高視覺注意值的物體雖說越多越好，但是數量過多可能造成相異處的周圍充斥過多高視覺刺激，而違反的策略二，在我們的經驗裡，總物體數量上限設定為相異處數量的兩倍，可降低此狀況發生率。

挑選的準則除了建立在上述兩個要點外，我們會希望使用內含物體較多的影像畫面，因為來自相同畫面的物體共同存在於一個畫面，會比多個來自不同畫面的物體的合理性高。而各畫面間的時間軸上的間距不能太靠近，以避免挑選到相同物體的狀況發生，在我們的情況，間距設定為一分鐘便足以應付所有輸入資料。

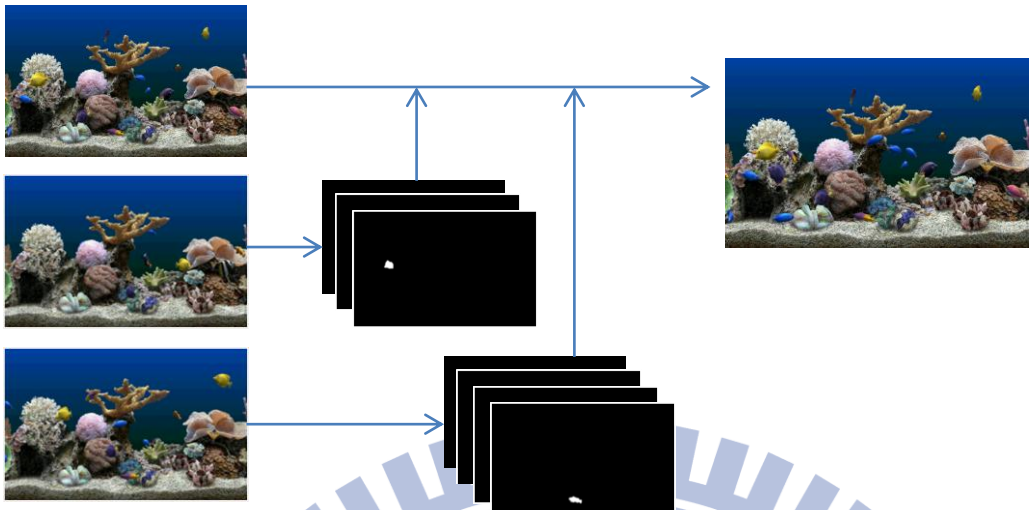

圖 10

產生「主要畫面」示意圖：左邊三張彩圖為根據章節 3-4 所挑出的影像，其中上面第一張畫面內移動物體數量最多，因此將剩餘兩張影像內的移動物體添加進第一張內，以此方法產「主要畫面」。

3-5 挑選最終主要畫面

一個不易被找出的相異處，會符合先前所訂定的策略一、二，而一組具有相當難度的「Spot the Differences」遊戲，裡面所包含的大多相異處必定不容易被發現。而不易被發現的相異處會擁有以下特性：數值低落的「物體視覺注意程度」（ATL）以及「物體易發意識程度」（AWL）。因此，前一階段所生成的數張「主要畫面」之中，何者所製作的成品難度最高，判斷方法便是比較各張所包含的所有相異處兩個數值的高低，另外加上策略三，高視覺刺激物體的數量多寡：

$$E_{O_d} = k \cdot ATL(O_d)^{n_h/n_l} + {C_{AWL}}^{-1} \cdot AWL(O_d) \quad (14)$$

$$E(I_E) = \sum_{O_d \in D} E_{O_d} \quad (15)$$

I_E為上一節所結合出的主要畫面，D為當前候選I_E中欲作為相異處的物體O_d的集合，n_h為I_E內ATL高於門檻值的移動物體數量，而n_l則是相異處數量，在我們的實驗裡ATL門檻值設定為0.5，k為控制係數，根據輸入影像的大小可做調整，例如影像較大，則全域性的ATL影響力會較小，C_{AWL}為將AWL數值標準化於1至0之間的係數，詳細算法如下：

$$C_{AWL} = 2\pi \int_0^1 {d_{aw}}^2 \cdot r^{2/c_d} = 2\pi \cdot \frac{{d_{aw}}^2 \cdot c_d}{c_d + 2} \quad (16)$$

其中d_{aw}以及c_d為先前計算AWL所設定的控制係數。

接著說明能量式的設計理念：E_{O_d}是「單一物體」選為相異處的能量表示式，根據策略三，畫面內除了選為相異處以外的高視覺刺激物體越多，相異處本身的視覺刺激越容易遭到玩家忽略。物體視覺注意值數值介於0至1之間，而相異處數量為固定值（對於各「主要畫面」而言），因此高視覺刺激物體的數量較高，會使得n_h/n_l提高，進而降低$ATL(O_d)^{n_h/n_l}$這項次的數值。而策略一、二，我們於本章第三小節制定了量化方式，便是「物體視覺注意程度」（ATL）以及「物體易發意識程度」（AWL），整體來說這兩數值越低，能量式數值也會降低，代表此物體較適合選為相異處。需注意的是

此能量式內$AWL(O_d)$計算方式裡頭的視野熱線表示圖M必須從所屬的主要畫面候選做計算，而非物體原先所屬的鏡頭畫面。

定義完單件物體，則整張影像是否為最佳的主要畫面便是分析影像內所有作為相異處的物體，如同算式(15)的計算方式，將各自能量式數值做累加作為此影像的能量值。當比較兩張主要畫面之間的能量值的時候，能量值越小的畫面裡頭作為相異處的物體整體來說會較另一張來的難被發現，而我們視其較適合製作成具有難度的「Spot the Differences」。因此，最後的步驟便是對於上階段挑選出來所有可作為「主要畫面」的影像，計算各自的能量式數值，其中數值最小的便是進入最後階段的「主要畫面」。

3-6 相異處類型

當決定好最終的主要畫面後，最後只剩一個步驟，將先前作為相異處的物體實做成相異處，並且生成與主要畫面配對的另一張影像（與主要畫面配對的影像便是將主要畫面中選為相異處的物體去除後的影像）。相異處的類型我們設定有三種：短距離位移、翻轉物體、存在與否，在此先說明三種類型的實作法；與主畫面配對的另一張影像，包含的物體並不會有選為相異處的物體，因此短距離位移以及翻轉物體是對物體作變動後添加到配對影像中，前者是在黏貼上時與原先位置有一段距離，而後者是將物體沿著某一方向做翻轉的動作後，黏貼上另一影像內。至於第三種類型則不做修改，讓物體只存在於原先的主要畫面內。

接著說明如何決定相異處的種類；由於AWL數值較高的物體容易引發視覺意識，若此類物體的相異處類型選擇為「存在與否」，玩家可能會意識到物體消失在視野內，所以這種物體無法選作為此類相異處。而餘下兩種我們並沒有限制，只會根據編輯時的方便性，以及是否會有不自然處作為考量，例如「短距離位移」會判斷移動後是否會遮蓋住其他物體。其中除了「存在與否」這種類型，其餘兩種皆有編輯的動作，這裡有一些小技巧可以使編輯後相異處更加難以被發掘；「短距離位移」移動向量的

方向可以依據兩張圖片的排列方向來決定，比如橫向並排的情況下，位移向量與水平線夾角會小於45°。「翻轉物體」則是根據畫面內水平面方向做翻轉，使得物體在編輯後並無不自然處。雖說相異處種類會影響該物體被偵測率，但為了遊戲的豐富性及策略性，我們會試著去平衡各種類的相異處數量，以避免某種類型出現過於頻繁。至此所有生成「Spot the Differences」所需的動作已全部結束。（本頁附圖11為相異處種類實例。）

圖 11

相異處範例：由左至右依序為短距離位移、翻轉物體（此範例為左右翻轉）、存在與否。由於充分的背景資訊，三種範例並未發現編輯時所造成不自然處。

四、實驗與結果討論

在本章節會說明用以佐證我們做法的實驗流程及結果討論，第一節列出為產生測試資料所輸入影片的資訊，我們主要設計了兩組實驗，第一組於第二、三節說明流程以及結果討論，接著於第四、五小節經由第二組實驗我們更深入探討「物體易發意識程度」對於相異處偵測的影響，最後小節結合實驗數據分析及經由我們方法所產生的「Spot the Differences」遊戲結果做最後的討論。

4-1 輸入影片

我們使用五段影片，其中四段為真實世界的場景，一段為模擬的電腦動畫，解析度依序為1440*1080、1440*1080、840*524、720*540，單位為像素，真實場景影片使用手持攝影機架設於三腳架上拍攝。以下為影片其餘資訊：

影片內容	影片長度（框架）	主要移動物體
交通大學小木屋	5000 (30 框架/秒)	行人
交通大學浩然圖書館前廣場	5000 (30 框架/秒)	行人
虛擬水族箱	2000 (30 框架/秒)	魚（3-D model）
馬路（臺北市南海路）	700 (30 框架/秒)	車

表格 1 實驗一、二中測試資料的來源影片。

4-2　實驗一設計

測試資料：共四組圖片，各組兩張圖片左右相接，中間以間隔十個像素寬，圖片解析度調整為各寬720像素，使用圖片組合於下頁開始之圖13至圖16。（解答於本論文頁42附圖28。）

測試平台：22吋液晶寬螢幕，解析度為1680*1050（像素）。

受測者：十五位，年齡分布為二十二至三十間，四位女性，十一位男性。

測試方式：受測者以滑鼠點擊相異處位置，若是點擊正確，會以紅色方框標示出來，實驗並無限制點擊需位於左側或右側圖片。圖片順序為亂數，每張圖限制時間五分鐘，為了避免受測者進行毫無意識的點擊舉動，我們制定了以下方法：初始可點擊次數與相異處數量相同，若點擊錯則扣除一次，點擊正確則歸還扣除的次數。最後在測驗結束後詢問受測者對於遊戲的難度，等級分為難、適中、易。（附圖12為測試畫面）

紀錄資訊：滑鼠點擊的位置及時間，途中放棄者我們視為挑戰失敗，因此答題時間計為五分鐘（300 sec.）。

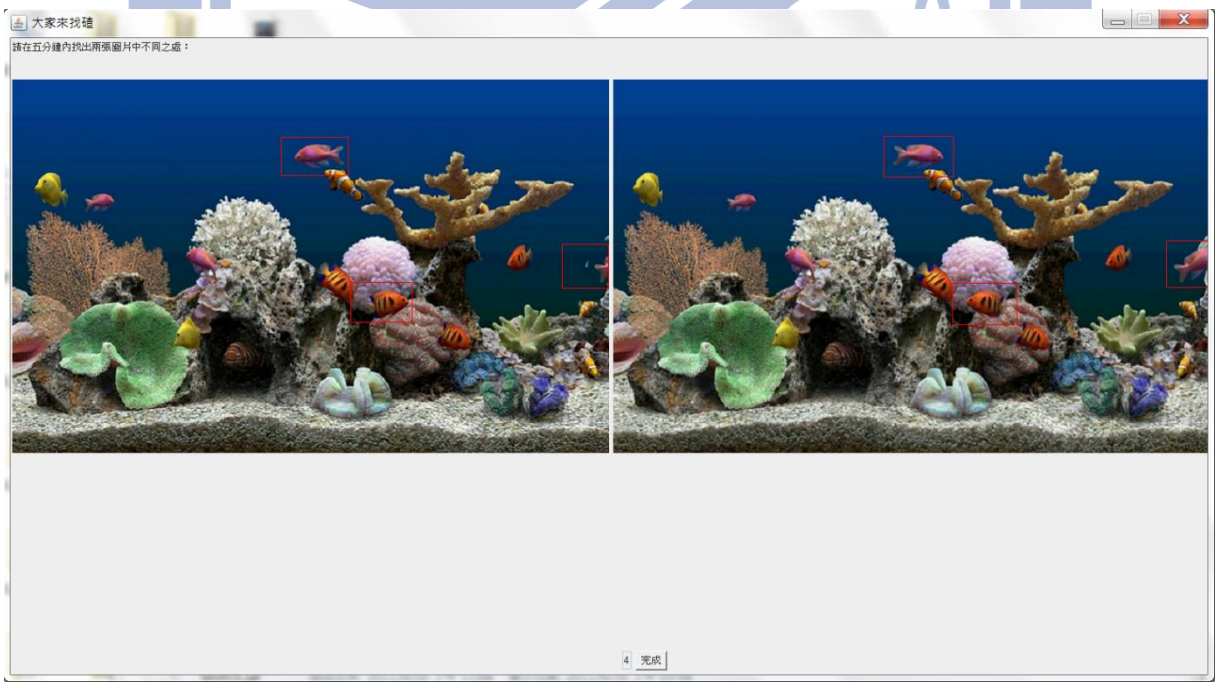

圖 12

實驗測試界面，未發現相異處數量顯示於下方，受測者可自由點選左右兩圖中所發現的相異處，當尋找出所有相異處或是選擇放棄時須點選下方「完成」按鍵結束測試。

圖 13

交通大學小木屋前路口，共有四處相異處。

圖 14

交通大學浩然圖書館前廣場，共四處相異處。

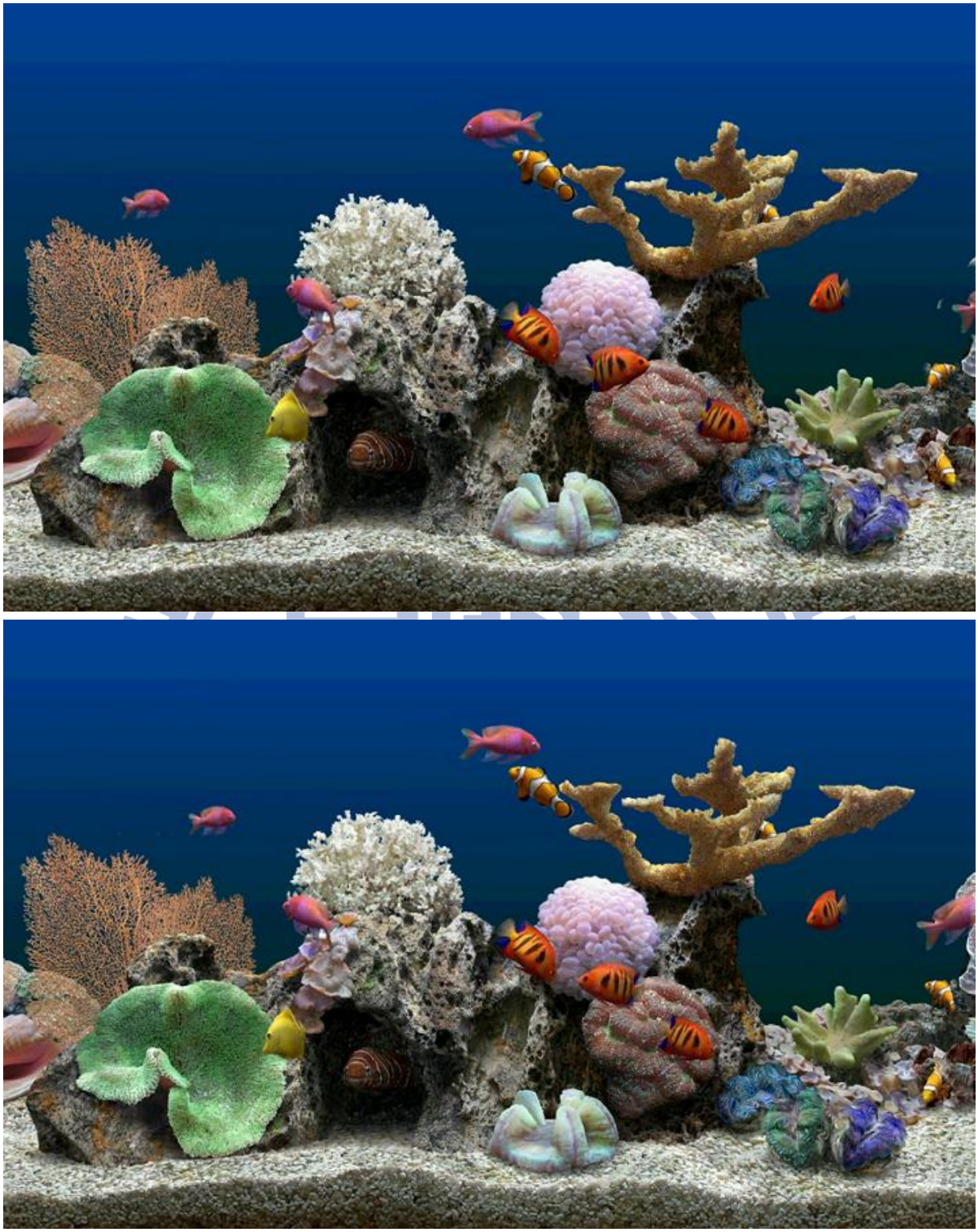

圖 15

虛擬水族箱，共七處相異處。

圖 16

臺北市南海路，共有四處相異處。

4-3 實驗一結果分析

我們以相異處的發現順序來做第一步的分析，見附圖17、18，可發現大多數第一次發現以及最後發現的相異處大多落在某些特定物體，以這兩種分布圖並結合先前對於物體的分析資料可得到下頁表格1：

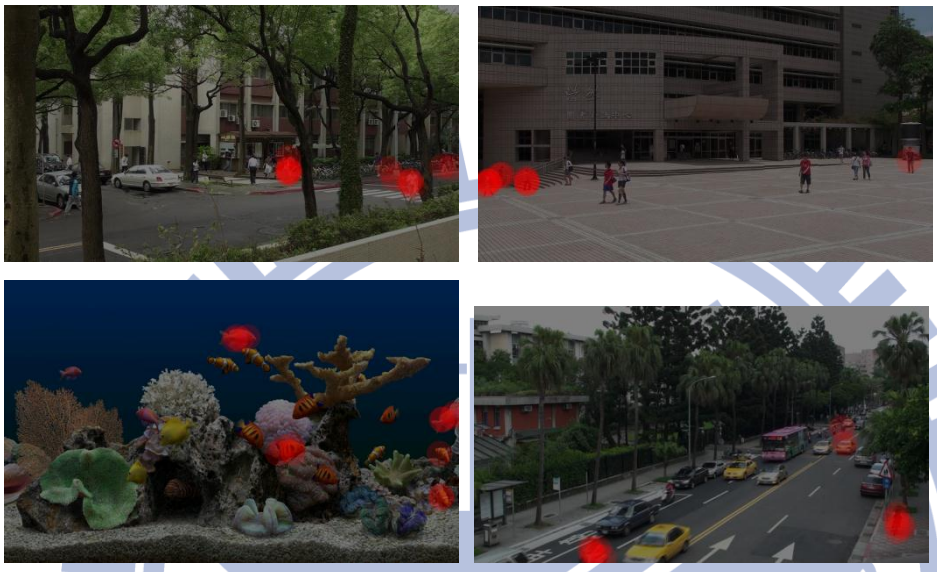

圖 17

上為第一發現群內相異處的分布圖，紅色區域為受測者點擊位置。

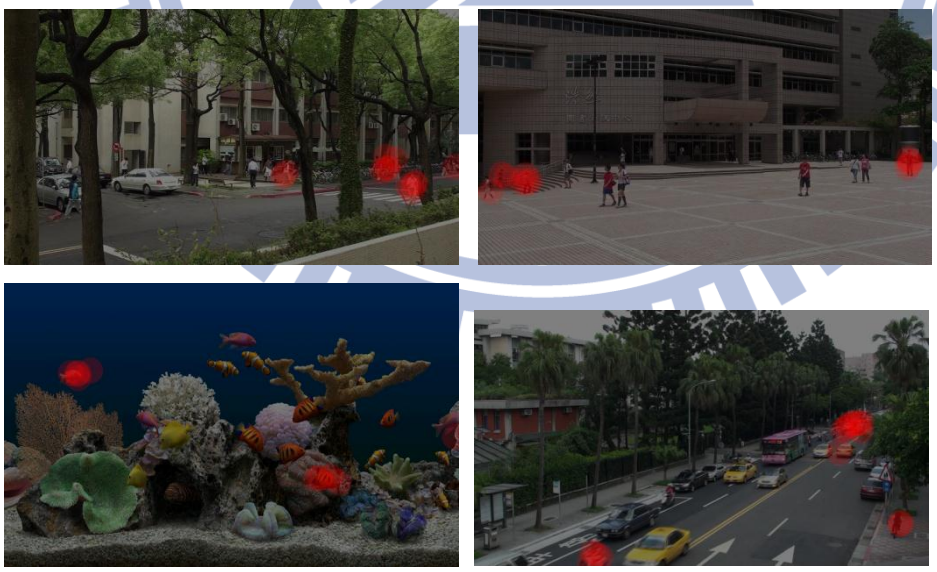

圖 18

上為最後發現群內相異處的分布圖，紅色區域為受測者點擊位置。

編號	ATL	AWL	種類	第一發現群	最後發現群	所屬影片
1	0.1367	0.0207	3			交通大學小木屋
2	0.2746	0.0392	1		○	
3	0.3693	0.0296	2		○	
4	0.4740	0.0663	3	○		
5	0.1145	0.0189	3			交通大學浩然圖書館前廣場
6	0.2576	0.0147	3			
7	0.33065	0.0139	2		○	
8	0.3395	0.0257	2	○		
9	0.365334	0.0750	1			虛擬水族箱
10	0.385872	0.0476	1			
11	0.402179	0.0294	1		○	
12	0.412253	0.0430	2			
13	0.425189	0.0765	2	○		
14	0.460589	0.0495	2			
15	0.490067	0.0411	2	○		
16	0.200082	0.0236	3			馬路（台北市南海路）
17	0.233212	0.0239	1		○	
18	0.298589	0.0330	2			
19	0.364541	0.0064	3	○		

表格 2

相異處分析表，每一橫列代表一個相異處的分析結果，欄位一：相異物編號；欄位二：相異物的「物體視覺注意程度」值；欄位三：相異物的「物體易發意識程度」；欄位四：相異物種類，1 短距離移動、2 翻轉物體、3 存在與否；欄位五：相異處是否為第一發現群；欄位六：相異處是否為最後發現群；最後第七欄位為來源影片。

可以很明顯看到每組測試資料中第一發現群裡必含有物體視覺注意程度最高的相異處，或是物體易發意識程度較高（編號13），而在最後發現群這欄位絕大多數屬於物體易發意識程度低者。至此，我們可以認為第一及第二假設是成立的，視覺注意程度與易發意識程度較高較高的相異處容易在早期被發現。

而有些同樣易發意識程度低的相異處雖覺不屬於第一發現群，但卻未進入最後發現群，原因或許是「相異處種類」，可發現皆屬於「存在與否」的類型（編號1、5、6、16）。如同我們在3-6所說，此種相異處是最能夠引起以視覺意識的類型，但有時為

了遊戲的豐富性不得不捨棄全使用同種，或少數相異處種類，在表格2裡頭也可發現，即便讓編號1、5、6並未發揮其低物體易發意識低的特性，整體遊戲難度並未下降，也可以發現在限時五分鐘的情況下，我們成功的讓這三個相異處所屬的遊戲答對率只有八成以及八成六。

詢問過實驗一的受測者對於測試資料的難易度的看法後，其中只有兩位認為難度偏易，進一步詢問原因後，一位表示自己十分擅長此類遊戲，另一位表示找到測資的破解法（破解法於下章節做討論）。四位表示難度適中，剩餘八位表示難度偏難，並且也反映在遊玩時間上，這八位受測者答題時間皆大於平均答題時間。因此，我們相信本論文的做法是足以產生對於一般人來說，較為困難的「Spot the Differences」。

另外我們能發現在第一發現群裡的相異處，物體易發意識程度（AWL）的數值分布相當廣泛（編號4為該組別中數值最高者，編號19為該組別中最低），可看出相異處偵測行為中就如同Rensink所提出的論點，視覺注意力為重要因素[44]。但同時由編號7、編號8以及編號11、13這兩對各處於相同影片的相異處可觀察出，即使ATL數值相當接近但被偵測的先後順序卻相差甚遠，就如同我們對於相異處被偵測率計算方式的看法，AWL是造成此結果的原因。為了解釋這種情況並且證明我們對於相異處被偵測率的論點，設計了另一個實驗以佐證物體易發意識程度確實會影響相異處偵測率。

影片內容	**答對率**	**平均答題時間**
交通大學小木屋	86.67%	109.66 sec.
交通大學浩然圖書館前廣場	80.00%	136.02 sec.
虛擬水族箱	60.00%	192.63 sec.
馬路（臺北市南海路）	100.0%	70.25 sec.

表格 3

未答對者答題時間設定為 300.0sec.。

4-4 實驗二設計

於上個實驗中我們證實了在計算相異處的被偵測率時，視覺注意力對於人眼的刺激強度對於真實結果有著相當大的影響力，而本實驗的目的在於驗證物體易發意識程度（AWL）的高低同時也會反映在被偵測率上。

測試資料：從實驗一中挑選出AWL數值低落的物體，即是頁33表格2中編號3、7、11，由於第四個影片「馬路」所產生的測資中AWL數值低落物體所處位置難以產生實驗二所需的測資，因此並未選作為實驗二的測試圖片。根據我們所定義的方程式(13)計算方式，在盡可能不更動ATL的情況下以下列兩種方式提高其AWL數值：一、將物體放大以提高低層級視覺刺激的發散量，此舉目的在於調高$ROS(O)$的數值。二、在物體的周圍放置具有高視覺注意力的物體，增加此物體進入視野範圍內的機率，此舉用意在提高視野範圍內的$M(r)$數值。由上述兩動作所產生的兩組圖片以及原始組合共三組作為本實驗的測試資料，測試圖以圖13、14、15共三組為主做修改並附上與原圖片組中左圖的比較圖19至24於下頁，呈現方式與實驗一相同，各組兩張圖片左右相接，中間以間隔十個像素寬，圖片解析度調整為各張寬720像素。由於圖片組16編輯上有相當程度的困難性，因此實驗二測資並無將其納入。

測試平台：22吋液晶寬螢幕，解析度為1680*1050（像素）。

受測者：與實驗一無重複的二十位受測者，年齡分布為二十二至二十五間，八位女性，十二位男性。

測試方式：受測者以滑鼠點擊相異處位置，若是點擊正確，會以紅色方框標示出來，實驗並無限制點擊需位於左側或右側圖片。圖片順序為亂數，每張圖限制時間五分鐘，並限制點擊次數不可過於頻繁，避免受測者進行毫無意識的點擊舉動（方法如實驗一所述）。（測試畫面如本論文頁26附圖12所示）

紀錄資訊：滑鼠點擊的位置及時間，途中放棄者我們視為挑戰失敗，因此答題時間計為五分鐘。

圖 19

左為原圖 13 中上方的圖片，右為將紅框內物體變大後的測試資料。

圖 20

左為原圖 13 中上方的圖片，右圖為測試資料，紅框內為添加的高視覺注意力程度物體。

圖 21

左為原圖 14 中上方的圖片，右為將紅框內物體變大後的測試資料。

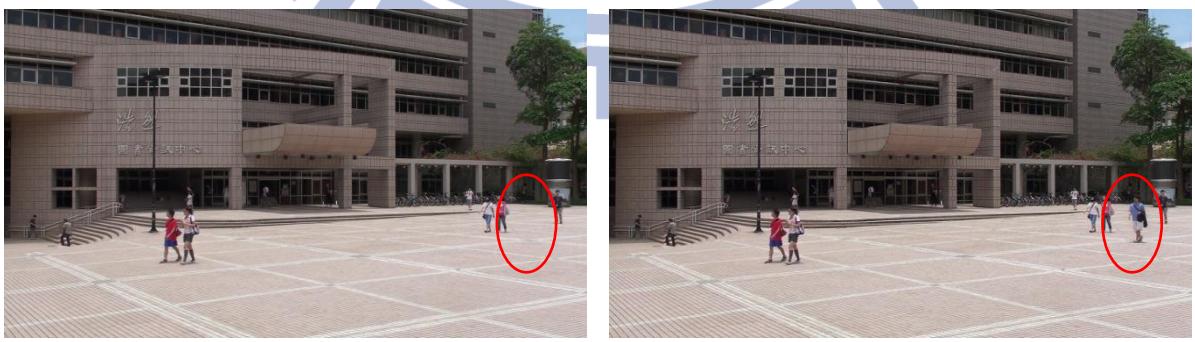

圖 22

左為原圖 14 中上方的圖片，右圖為測試資料，紅框內為添加的高視覺注意力程度物體。

圖 23

左為原圖 15 中上方的圖片，右為將紅框內物體變大後的測試資料。

圖 24

左為原圖 15 中上方的圖片，右圖為測試資料，紅框內為添加的高視覺注意力程度物體。

4-5 實驗二結果分析

下頁表格3為實驗二經過整理後的測試結果，所挑選出的物體經由調整後無法避免的*ATL*必定會有些許的變動，但可由下頁表格3欄位2以及欄位3看出變動幅度不如*AWL*大（*ATL*最大變動倍度為1.039，而*AWL*最小變動倍數為1.252），另外「第一發現群」和「最後發現群」與表格1不同在於下頁表格3只考慮實驗二中我們為驗證「物體易發意識程度」與相異處被偵測率的關係所挑出的低*AWL*數值物體，以下依序由輸入影片做分組討論。可從下頁表格3發現「虛擬水族箱」的類別中，我們所挑選做測試的相異處皆屬於原版圖片組中的最後發現群，因此若該物體的被偵測率提升則受測者會更快結束遊戲，並且也會減少因偵測不到而導致遊戲失敗的情況發生，而我們可以從該類別的「放大物體」類型發現，雖然低*AWL*數值物體仍然是最後發現群但是答對率大幅上升，平均達題時間也減少許多，並且「添加高*ATL*物體」此類別該物體也並不屬於最後發現群，因此在這影片類別中我們所挑選的物體因*AWL*數值提升而成功的提高被偵測率。「浩然圖書館前廣場」此類別所挑出的低*AWL*數值物體同樣在原版圖片組中屬於最後發現群，在「放大物體」類型中該物體脫離了最後發現群，「添加高*ATL*物體」裡答對率大幅上升且平均達題時間也減少。「交通大學小木屋」此類別的低*AWL*數值物體原並不屬與第一發現群也不屬於最後發現群，但可從下頁表格3中看到該物體經由兩種方式提升*AWL*數值後躍升為第一發現群，由於此相異處原並非最後發現群，因此影響答對率的因素會落在屬於最後發現群相異處的被偵測率，因此此類別的答對率並無顯著提升，但是平均答題時間仍因相異處變得較易被偵測而縮短。

測資類型	ATL	AWL	答對率	平均答題時間	第一發現群	最後發現群	所屬影片
放大 (圖 19 右)	0.40635	0.0662	85.71%	116.94sec. (有效樣本數:7)		○	虛擬水族箱
添加 (圖 20 右)	0.40106	0.0412	80.00%	136.48sec. (有效樣本數:4)			
原版 (圖 19 左)	0.40217	0.0294	66.67%	158.26sec. (有效樣本數:3)		○	
放大 (圖 21 右)	0.34153	0.0272	100.0%	81.60 sec. (有效樣本數:5)			交通大學浩然圖書館前廣場
添加 (圖 22 右)	0.32951	0.0174	100.0%	59.14 sec. (有效樣本數:7)		○	
原版 (圖 21 左)	0.33065	0.0139	71.43%	123.81sec. (有效樣本數:7)		○	
放大 (圖 23 右)	0.36242	0.0580	75.00%	90.32sec. (有效樣本數:4)	○		交通大學小木屋
添加 (圖 24 右)	0.36014	0.0357	85.71%	127.37sec. (有效樣本數:7)	○		
原版 (圖 23 左)	0.3693	0.0296	75.00%	140.68sec. (有效樣本數:4)			

表格 4

實驗二結果整理，由於物體經過更動後 ATL 數值與原版圖有異，但變動幅度仍遠小於 AWL 數值變化量。

4-6 結果與討論

透過實驗一證實了我們由視覺注意力所定義「物體視覺注意程度」（ATL）確實與相異處的被偵測率有絕對關係，ATL數值高的物體絕對會較早被玩家所偵測到，雖然「物體易發意識程度」（AWL）並不是決定性的因素，但在實驗二裡頭可發現，透過一些手段調整先前我們所定義的AWL能量式裡頭各元素，讓相異處的AWL數值提高後可有效的增加其受偵測率，即便該物體的ATL數值相當低。而透過這兩個實驗證實了我們方程式(14)是能夠完整表達一個相異處其被偵測率的程度，並且遊戲難度反映在答對率以及受測者的反應，大多數的受測者無法完全找出我們所產生的遊戲裡頭所有的相異處，並且實驗一的大多數受測者反映所遊玩的「Spot the Differences」遊戲是具有難度的。下頁起附圖25至27為實驗中與測試資料相同影片，一樣透過本論文的做法但難易度偏易的圖片組，圖28為本論文裡所有「Spot the Differences」圖片組的解答。

圖 25

交通大學小木屋前路口，共有三處相異處，難度較易，可看到左側挑選到的相異處，是一輛面積相當大的車子，相當容易被辨識出。

圖 26

交通大學浩然圖書館前廣場，共四處相異處，難度較易，內含非相異處的高視覺刺激物較少。

圖 27

臺北市南海路，共有四處相異處，難度較易，相異物普遍較靠近高視覺刺激的區域。

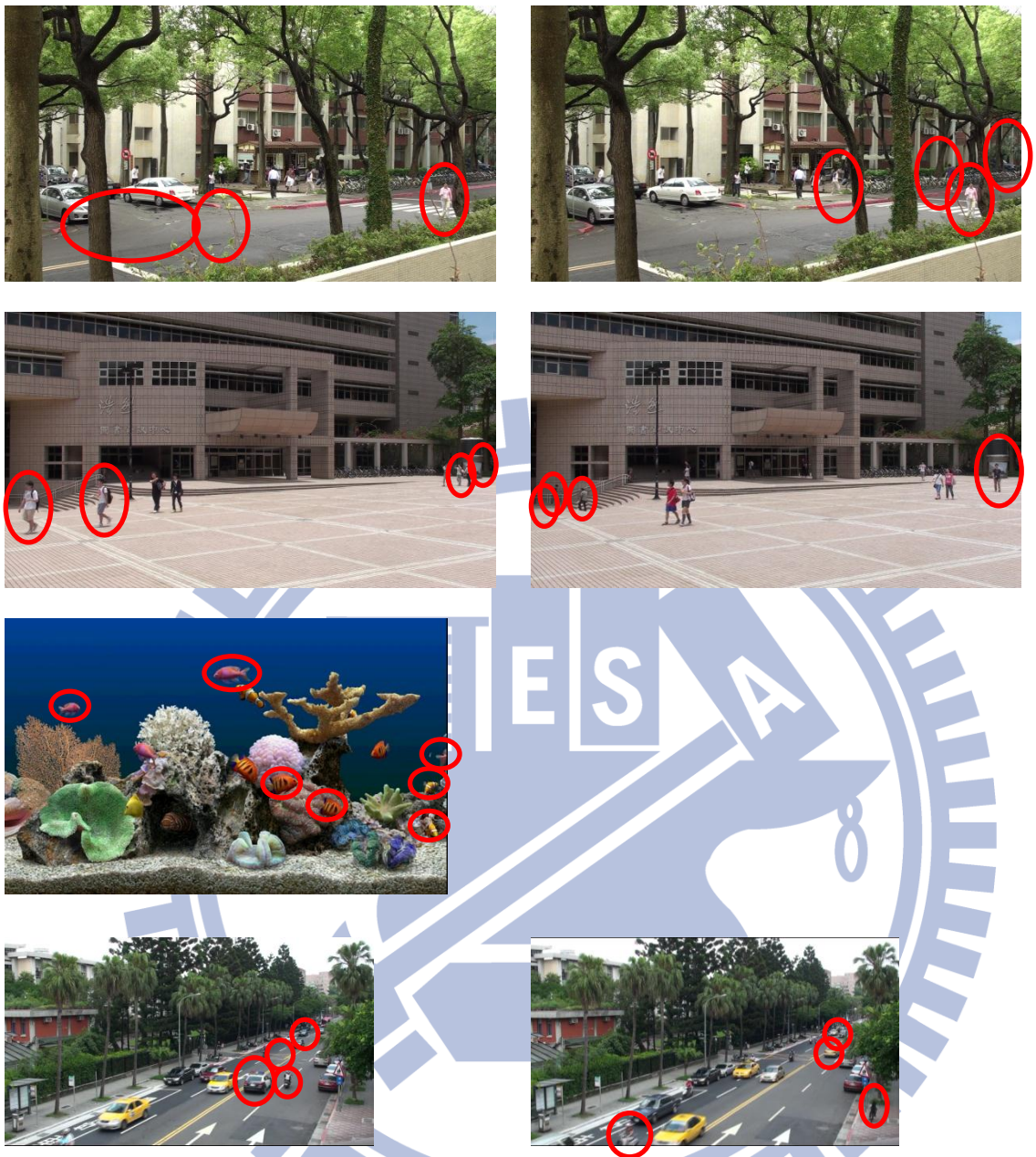

圖 28

解答，皆以左側圖片為主，第一橫列：圖 13、25；第二橫列：圖 14、26；第三橫列：圖 15；第四橫列：圖 16、27。

五、貢獻與未來展望

5-1 貢獻

在過去影像處理及電腦圖學領域尚未有人對於「Spot the Differences」此遊戲提出一個完整的分析及製作方法，即便此遊戲對於人類視知覺領域有相當的貢獻。而我們除了提出一套完整的製作流程，也成功證實了視覺刺激有助於相異處的發現，並且定義了物體與背景間視覺意識（visual awareness）的相對關係對於相異處偵測的影響。

5-2 未來展望

我們的做法僅限於將完整物體透過編輯成為相異處，與一般人工製作相比，我們無法做到許多獨具創意的更動，因此相異處受限於只能對完整物體做編輯，不能對局部區域做更動，而這也是第四章節裡，有一受測者所說的破解方式，只需關注在物體上即可。對於輸入資料的類型，目前限制只能接受固定視角的影片，並且無法有效的將疊合物體做分離的動作。上述不足部分，我們相信只需再資料分析階段加入以下兩項分析資訊即可做到：對物體本身的色塊作分割，分別對每一色塊以我們分析物體的方式算出各自的視覺顯著程度，即可做到物體局部的更動。對移動物體建立移動軌跡，建立各個物體的完整移動資訊，便可分離出疊合的物體。

我們產生「Spot the Difference」的方式相信還可應用到以下兩個層面：一、以影片的方式作呈現，並且藉此研究人眼對於真實移動物體的敏感度是否與平面影像相同會有「變盲」的現象發生。二、提供腦科學、視知覺相關領域一個工具，方便研究人員產生所需的測試資料。三、可製作使用者介面，提供一般人製作屬於自己的「Spot the Differences」遊戲，進而與人交流分享或是互相挑戰。

參考資料

[1] J. R. Bergen and B. Julesz, "Parallel versus serial processing in rapid pattern discrimination," *Nature*, vol. 303, pp. 696-698, 1983.

[2] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 26, pp. 1124-1137, 2004.

[3] J. Braun and B. Julesz, "Withdrawing attention at little or no cost: detection and discrimination tasks," *Attention, Perception, & Psychophysics*, vol. 60, pp. 1-23, 1998.

[4] B. Bridgeman, D. Hendry, and L. Stark, "Failure to detect displacement of the visual world during saccadic eye movements," *Vision Research*, vol. 15, pp. 719-722, 1975.

[5] C. M. Christoudias, B. Georgescu, and P. Meer, "Synergism in low level vision," in *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, 2002, pp. 150-155 vol.4.

[6] H.-K. Chu, W.-H. Hsu, N. J. Mitra, D. Cohen-Or, T.-T. Wong, and T.-Y. Lee, "Camouflage images," *ACM Transactions on Graphics*, vol. 29, pp. 1-8, 2010.

[7] C. Currie, G. McConkie, L. Carlson-Radvansky, and D. Irwin, "Maintaining visual stability across saccades: Role of the saccade target object," *Manuscript submitted for publication*, 1995.

[8] W. Einhauser, M. Spain, and P. Perona, "Objects predict fixations better than early saliency," *Journal of Vision*, vol. 8, 2008.

[9] E. Fukuba, H. Kitagaki, A. Wada, K. Uchida, S. Hara, T. Hayashi, K. Oda, and N. Uchida, "Brain Activation during the Spot the Differences Game," *Magnetic Resonance in Medical Sciences*, vol. 8, pp. 23-32, 2009.

[10] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, 2010, pp. 2376-2383.

[11] S. Gould, R. Fulton, and D. Koller, "Decomposing a scene into geometric and semantically consistent regions," in *Computer Vision, 2009 IEEE 12th International Conference on*, 2009, pp. 1-8.

[12] J. Grimes, "On the failure to detect changes in scenes across saccades," *Perception (Vancouver Studies in Cognitive Science)*, vol. 5, pp. 89-109, 1996.

[13] K. Hong-Wen, C. Xue-Quan, Y. Matsushita, and T. Xiaoou, "Space-Time Video Montage," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, 2006, pp. 1331-1338.

[14] M. Irani, P. Anandan, J. Bergen, R. Kumar, and S. Hsu, "Efficient representations of video sequences and their applications," *Signal Processing: Image Communication*, vol. 8, pp. 327-351, 1996.

[15] L. Itti and C. Koch, "Computational modeling of visual attention," *Nature reviews neuroscience*, vol. 2, pp. 194-203, 2001.

[16] L. Itti and C. Koch, "A saliency-based search mechanism for overt and covert shifts of visual attention," *Vision Research*, vol. 40, pp. 1489-1506, 2000.

[17] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 20, pp. 1254-1259, 1998.

[18] B. Julesz, "Experiments in the visual perception of texture," *Scientific American*, vol. 232, p. 34, 1975.

[19] B. Julesz, "Texton gradients: The texton theory revisited," *Biological Cybernetics*, vol. 54, pp. 245-251, 1986.

[20] B. Julesz, "Textons, the elements of texture perception, and their interactions," *Nature*, 1981.

[21] T. A. Kelley, M. M. Chun, and K. P. Chua, "Effects of scene inversion on change detection of targets matched for visual salience," *Journal of Vision*, vol. 3, 2003.

[22] C. Kim and J. N. Hwang, "An integrated scheme for object-based video abstraction," in *8th ACM International Conference on Multimedia*, 2000, pp. 303-311.

[23] C. Koch and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry," *Hum Neurobiol*, vol. 4, pp. 219-27, 1985.

[24] G. Kreiman, C. Koch, and I. Fried, "Category-specific visual responses of single neurons in the human medial temporal lobe," *nature neuroscience*, vol. 3, pp. 946-953, 2000.

[25] V. Kwatra, A. Schodl, I. Essa, G. Turk, and A. Bobick, "Graphcut textures: image and video synthesis using graph cuts," *ACM Transactions on Graphics*, vol. 22, pp. 277-286, 2003.

[26] V. A. F. Lamme, "Why visual attention and awareness are different," *Trends in cognitive Sciences*, vol. 7, pp. 12-18, 2003.

[27] O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau, "A coherent computational approach to model bottom-up visual attention," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 802-817, 2006.

[28] F. F. Li, R. VanRullen, C. Koch, and P. Perona, "Rapid natural scene categorization in the near absence of attention," *Proceedings of the National Academy of Sciences*, vol. 99, p. 9596, 2002.

[29] A. Mack and I. Rock, *Inattentional blindness*: The MIT Press, 1998.

[30] J. McCann and N. Pollard, "Local layering," presented at the ACM SIGGRAPH 2009 papers, New Orleans, Louisiana, 2009.

[31] G. W. McConkie and D. Zola, "Is visual information integrated across successive fixations in reading?," *Attention, Perception, & Psychophysics*, vol. 25, pp. 221-224, 1979.

[32] J. Nam and A. H. Tewfik, "Video abstract of video," in *Multimedia Signal Processing, 1999 IEEE 3rd Workshop on* 1999, pp. 117-122.

[33] J. K. O'Regan and A. Noe, "A sensorimotor account of vision and visual consciousness," *Behavioral and brain sciences,* vol. 24, pp. 939-972, 2001.

[34] J. K. O'Regan, R. A. Rensink, and J. J. Clark, "Change-blindness as a result of 'mudsplashes' [letter]," *Nature,* vol. 398, pp. 34-34, 1999.

[35] C. Pal and N. Jojic, "Interactive montages of sprites for indexing and summarizing security video," in *Computer Vision and Pattern Recognition, 2005 IEEE Computer Society Conference on*, 2005, p. 1192 vol. 2.

[36] H. Pashler, "Familiarity and visual change detection," *Attention, Perception, & Psychophysics,* vol. 44, pp. 369-378, 1988.

[37] N. Petrovic, N. Jojic, and T. S. Huang, "Adaptive video fast forward," *Multimedia Tools and Applications,* vol. 26, pp. 327-344, 2005.

[38] W. Phillips, "On the distinction between sensory storage and short-term visual memory," *Attention, Perception, & Psychophysics,* vol. 16, pp. 283-290, 1974.

[39] A. Pope, R. Kumar, H. Sawhney, and C. Wan, "Video abstraction: Summarizing video content for retrieval and visualization," in *Signals, Systems & Computers, 1998. Conference Record of the Thirty-Second Asilomar Conference on*, 1998, pp. 915-919 vol. 1.

[40] Y. Pritch, A. Rav-Acha, A. Gutman, and S. Peleg, "Webcam synopsis: Peeking around the world," *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on,* 2007.

[41] Y. Pritch, A. Rav-Acha, and S. Peleg, "Nonchronological video synopsis and indexing," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* pp. 1971-1984, 2008.

[42] A. Rav-Acha, Y. Pritch, and S. Peleg, "Making a long video short: Dynamic video synopsis," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, 2006, pp. 435-441 vol.1.

[43] R. A. Rensink, "Change detection," *Annual Review of Psychology,* vol. 53, pp. 245-277, 2002.

[44] R. A. Rensink, J. K. O'Regan, and J. J. Clark, "To See or not to See: The Need for Attention to Perceive Changes in Scenes," *Psychological Science,* vol. 8, pp. 368-373, 1997.

[45] U. Rutishauser, D. Walther, C. Koch, and P. Perona, "Is bottom-up attention useful for object recognition?," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* 2004.

[46] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation," *Computer Vision-ECCV 2006,* pp. 1-15, 2006.

[47] D. J. Simons and C. F. Chabris, "Gorillas in our midst: Sustained inattentional blindness for dynamic events," *Perception*, vol. 28, pp. 1059-1074, 1999.

[48] D. J. Simons and D. T. Levin, "Failure to detect changes to people during a real-world interaction," *Psychonomic Bulletin & Review*, vol. 5, pp. 644-649, 1998.

[49] M. A. Smith and T. Kanade, "Video skimming and characterization through the combination of image and language understanding," in *Content-Based Access of Image and Video Database, 1998. Proceedings., 1998 IEEE International Workshop on*, 1998, pp. 61-70.

[50] J. Sun, W. Zhang, X. Tang, and H. Y. Shum, "Background cut," *Computer Vision-ECCV 2006*, pp. 628-641, 2006.

[51] L. Tie, S. Jian, Z. Nan-Ning, T. Xiaoou, and S. Heung-Yeung, "Learning to Detect A Salient Object," in *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, 2007, pp. 1-8.

[52] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive psychology*, vol. 12, pp. 97-136, 1980.

[53] D. Walther and C. Koch, "Modeling attention to salient proto-objects," *Neural Networks*, vol. 19, pp. 1395-1407, 2006.

[54] H. Xuming, R. S. Zemel, and M. A. Carreira-Perpinan, "Multiscale conditional random fields for image labeling," in *Computer Vision and Pattern Recognition, 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, 2004, pp. 695-702 vol.2.

[55] X. Zhu, X. Wu, J. Fan, A. K. Elmagarmid, and W. G. Aref, "Exploring video content structure for hierarchical summarization," *Multimedia Systems*, vol. 10, pp. 98-115, 2004.