任意方向及排列的文字偵測

陳伯岳
國立彰化師範大學副教授
pychen@cc.ncue.edu.tw

高偉格
國立彰化師範大學研究生
optimistic-sai@hotmail.com

摘要

在影像裡的文字是有一定重要性的，因此，如何在龐大的資料庫中快速的尋找到特定影像的文字極其困難。然而文字偵測這方面的研究大多限制在字幕、標題、車牌、路牌、書本上，這樣的限制大大的減少了數位媒體之多元化應用。而只有少數最近提出的研究，是沒有任何限制的，因為高運算複雜的關係，這樣的系統在處理一張影像往往要花費數秒鐘。

為了達到運算複雜度低且不遺失任何文字資訊的偵測，我們提出 Sobel 邊緣偵測的特徵值來判斷邊緣較密集的地方並且用以去除誤判區塊，因為我們發現在文字密集邊緣區塊的 Sobel 特徵值變化明顯較複雜背景的 Sobel 特徵值變化來得大。根據這樣的特性來作誤判區塊的消除，在不限制文字排列及方向的實驗中，結果顯示偵測一張影像裡的所有文字資訊平均只要 0.75 秒。

關鍵詞：文字偵測、影像檢索、Sobel 運算

Abstract

Texts in an image usually contain essential information about that image. Therefore, text detection is crucial to retrieving a specific image from a huge database. However, most of the existing text detection schemes set certain constraints on the target texts, such as alignment, orientation, size, color, and language. These constraints significantly limit the application scenarios. Only a few recently proposed schemes dedicate to text detection without specific constraints. However, because of the high computational complexity, such systems tend to spend couple of seconds on just one image frame. Based on a finding that the variance of Sobel gradients within a text block is generally higher, the authors propose using the results of Sobel gradients twice, at first applying to finding the candidate blocks and secondly to eliminating some non-text candidates (which are just complicated background blocks). As a result, the computational complexity is alleviated significantly and real time applications become feasible. According to the preliminary experimental results, the proposed system outperforms many existing schemes. For most test samples which contain texts in various sizes and orientations, the proposed system successfully detect all text blocks within 1 second.

Keywords: Text Detection, Image retrieving, Sobel operator

1. 簡介與文獻回顧

近幾年來，數位媒體的發展突飛猛進，更多的數位媒體資料（包含影片、相片、電子書等數位化內容）大量的增加，這些沒有組織的資料與日俱增，因此增加了我們找尋需求資訊的困難度。而這些數位資訊的發展關鍵在於我們必須如何去管理這些龐大的資料，並快速的完成資訊之傳遞與使用。

經由數位化的資訊得以永續保存，但是如何在龐大的資料庫當中快速的尋找到我們的資料極其困難，而在數位媒體資料裡的文字，是有一定重要性的，通常可將文字訊息視為數位媒體之關鍵特徵，所以現在大部分的檢索系統(搜尋工具)都是以文字為基礎的方式來搜尋資訊。以影像資料庫的搜尋而言，資料庫中的每一張影像都必須經過人工註解或加入索引標籤，這樣的方式非常的耗費人力，若可以開發一系統而自動找出影像中的文字，必能節省龐大的人工處理成本。因此如何擷取重要的文字訊息對影像搜尋而言乃一相當關鍵之技術。

文字偵測的研究已有十多年的時間，但大多將文字限制在字幕、標題、車牌、路牌、書本上[5][6][8][9][12]。這樣的限制，大大地減少了數位媒體能夠應用的廣度。許多我們認為不重要的文字訊息，例如：商家的招牌、標誌，皆被忽略掉，所以我們思考如何拿掉這些對於文字的限制，將文字偵測技術更廣泛地應用在實務中。

而影像中的文字訊息包羅萬象，有水平的、垂直的、彎曲的、也有圓型排列的，然因拍攝的角度不同，文字訊息所呈現的幾何關係也不同。所以我們提出一個不受文字型態影響的偵測系統，第一個重點為屏除所有以往的系統對於文字上的限制，並且是不受不同語系的侷限；第二個重點則在擴大了需要擷取的文字訊息範圍後，還要能維持偵測的實用性及即時性(Real time)。這是因為在影片中的文字偵測，運算速度是非常重要的。在先前的研究當[13][17][18]中，雖然分別屏除了前述的某些文字限制，但是運算複雜度卻提高很多，使得整個運算過程中一秒還不夠處理到一幀(frame)。本論文提出一個複雜度低而全面(文字無限制)的文字偵測方法，其優點為偵測出影像中的所有文字後，還可以透過使用者的需求來增加限制條件以便分類影片中的文字，進而達到快速搜尋資訊的目的。文字區塊找到後，透過影像分割處理把背景去掉，留下完整的文字資訊，最後再透過文字辨識系統即可提升數位媒體之搜尋速度。

過去的研究中，文字偵測的方法大致可分為:以邊緣(Edge)為基礎、以顏色相似度(Color Similarity)為基礎、還有以紋理(Texture)為基礎等三類方法。

以邊緣(Edge)為基礎[7][10][11][17][18]是根據文字是由許多筆畫組成的特性，也就意味著有文字的地方有邊緣聚集的可能性；這類方法先進行適合不同邊緣偵測的前處理，以增強文字的邊緣特徵或減少雜訊，偵測出邊緣資訊以後，再利用各種合併方法如連接組件法或者影像型態學(Morphological Operations)來連接預測的文字區塊。這種方式能處理背景邊緣特徵不明顯的情形，但對於文字與背景對比低的情況下，容易造成邊緣資訊的不完整，增加後續連結文字區塊的難度；另外也容易將多邊緣及邊緣聚集的背景誤判成文字區塊。

以顏色相似度(Color Similarity)為基礎的方法是依據文字有相似的顏色為前提，因此能處理出現在簡單背景的文字資訊，而最常見的做法就是依顏色相似度偵測出位置後，再利用連接組件方法將顏色及大小相似的可能文字連接成一個字串區塊。像[14]是針對灰階影像做字幕偵測，首先會將影像中像素值相近且相鄰的像素值用同一個顏色表示，再將多個文字組合成有意義的文字資訊。[1] 則提出了處理彩色影像的方法，由於有 R、G、B 三個分量，一個像素總共需要 24 位元來表示，所以作者針對減少位元以及色彩量化兩個方向來努力。首先將 R、G、B 值皆只取其最高的兩個位元作為特徵值，每個像素只採用 6 位元，再利用統計值方圖(Histogram)來找出幾個數量較多的像素值，並將相近的數值合併，以找出可能的文字區塊。

以紋理為基礎(Texture)的方法是根據文字有相似方向和固定間距的特性，而把文字視為一種特殊材質。早期，文字與背景有強烈對比，故用紋理特徵很容易找到文字區塊，但至今還沒找到一個最佳的紋理模型來代表任意文字，尤其在特徵類似或是背景複雜的情況下容易產生誤判。另外也可先用濾波器找出文字材質的候選區塊，再利用訓練樣本的方式來正確地區分這些候選區塊。[2][3]觀察到文字通常呈現水平排列且間距相同，因此若對含有文字區塊的影像做掃描將呈現週期性的水平密度變化。[15] 則以 Sobel 運算先找出水平和垂直資訊的邊緣圖後，再利用以向量量化編碼簿為基礎的貝氏分類器來進行文字樣本的訓練，以判斷候選文字區塊是否真的存在文字。

2.提出之方法

本論文所提出的方法具備兩個優點，一是屏除以往的系統對於文字上的所有限制(排列方向、位置、大小、顏色、語系)；二是能維持偵測的實用性及即時性(Real time)。以下分別逐步說明所提出之方法並討論之。

2.1 針對影像中的文字特徵設計演算法

本論文使用 Sobel 運算做邊緣偵測，首先產生兩組 3 x 3 的矩陣，分別為橫向的 G_x 及縱向的 G_y 如下：

$$G_X = \begin{bmatrix} -1 & 0 & +1 \\ -2 & 0 & +2 \\ -1 & 0 & +1 \end{bmatrix} * A \text{ and } G_Y = \begin{bmatrix} +1 & +2 & +1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} * A$$

其中 A 為代表原始影像方塊，將之與濾波矩陣做平面卷積，即可得到橫向及縱向的梯度差分近似值 G_x 及 G_y，然後以下列公式結合，即可得知圖像中每一個像素之梯度大小：

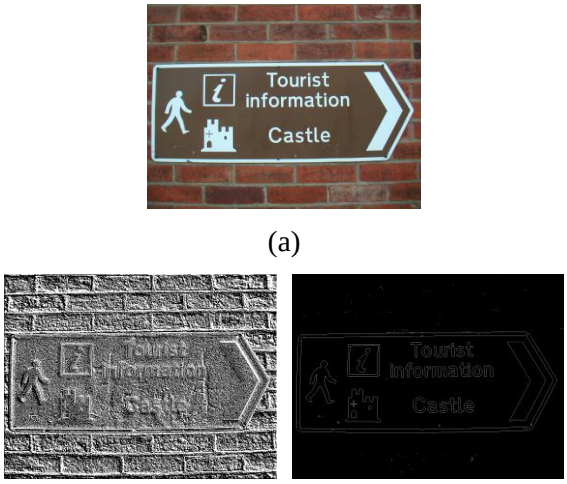
$$G = \sqrt{G_X^2 + G_Y^2}$$

為了提高效率，有時使用不開平方的近似值如下：

$$|G| = |G_X| + |G_Y|$$

Sobel 邊緣偵測的效果如圖 1 所示，圖 1(a)

為原圖，圖 1(b)為透過 Sobel 運算過後的影像，圖 1(c)為二值化後的影像。



(a)

(b) (c)

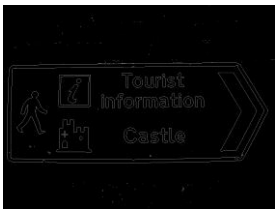
圖 1. (a).原始影像 (b).Sobel 特徵值影像 (c).二值化影像

2.2 對取得的文字特徵作處理

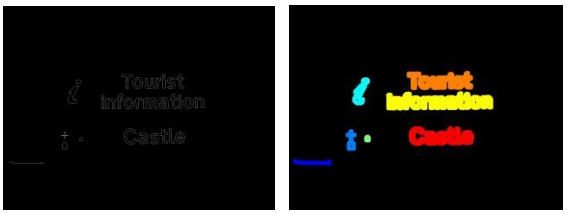
取得 Sobel 邊緣特徵值後，將分佈密度較低的邊緣去除，這是因為含有文字資訊的區域，邊緣密度較高，但這仍可能將複雜邊緣的背景誤判為文字。首先使用侵蝕做簡單的雜訊消除，並且將單一連接組件的邊緣長度大於門檻值 T_h 的消除，這是因為字母的邊緣長度通常較短，而較長的邊緣可能是招牌或路牌的邊框，或是一些較大的圖形等。然後利用膨脹的技巧將密度較高邊緣組成候選文字區塊如圖 2， 圖 2(a)為 Sobel 邊緣特徵值二值化影像，圖 2(b)為侵蝕並且消除有較長邊緣長度的連接組件影像，可以看到影像中的非文字邊緣，像是邊框與非文字圖形已消除了一部分，圖 2(c)為膨脹後的影像，不同顏色代表不同的區塊。

2.3 誤判區塊消除

圖 2(c)中的區塊除了包含有複雜邊緣的文字，也包含了複雜背景的候選文字區塊。為了消除誤判的區塊，並且以不增加太多系統運算複雜度的前提下，本研究根據多次實驗的統計，提出再次使用第一步驟取得文字特徵計算的 Sobel 運算結果。因為我們發現在候選區塊中的文字邊緣 Sobel 特徵值標準差較大，相對的，非文字而含有複雜邊緣的背景其 Sobel 特徵值標準差較小。如圖 3，縱軸為影像中文字與非文字區塊的邊緣特徵值標準差的平均值，橫軸為實驗的十張不同影像。



(a)



(b) (c)

圖 2 (a).Sobel 邊緣特徵值二值化影像 (b).雜訊去除 (c).膨脹

由圖 3 可發現，平均而言文字區塊的梯度標準差比非文字區塊的梯度標準差為高，但實際上還是有少部分非文字區塊的標準差是跟文字區塊的標準差平均是接近的。在後面的實驗結果可看到我們透過一門檻值將候選文字區塊中，邊緣特徵值標準差較小的區塊去除後，還有部分非文字區塊的殘留。另外因為誤判消除的方法是用前面步驟第一次邊緣偵測所運算過的特徵值，所以在最後誤判消除的部分，比以往研究緩慢複雜的運算快速，可以證明整個系統的運算速度是大幅提升的。

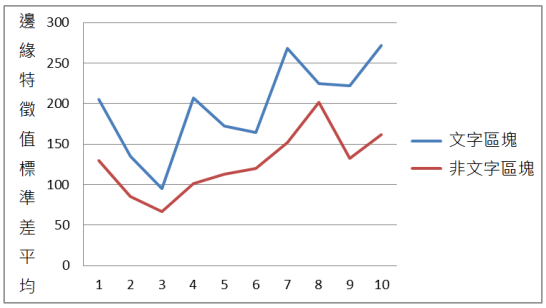


圖 3 邊緣特徵值標準差的平均比較

3. 實驗結果與分析

實驗資料針對水平排列文字、非水平排列文字、彎曲排列文字等三種排列方式之文字影像，且包含到大、中、小三種不同文字大小的影像做實驗，並且與近幾年的類似研究

[7][10][11]比較，這三篇提出的方法分別用 Sobel、Fourier-Laplacian、Sobel-Laplacian 來偵測邊緣資訊，之後在組成文字區塊的步驟也是利用膨脹概念的方法，[10][11]更運用文字區塊的幾何關係來消除誤判，例如假設正確的文字區塊的邊緣密度是很高的，並且文字資訊皆是以直線排列的型態出現，也就是不會有弧形排列或圓形排列的型態。而[7]對文字區塊沒有限制，與我們的研究最為相近，實驗結果如圖 4 到圖 9，其中(a)圖為原始影像，(b) 圖為 Sobel 邊緣偵測做二值化之結果，(c) 圖為雜訊去除結果， (d) 圖為.膨脹後的候選文字區塊，(e) 圖為透過特徵值標準差去除誤判結果，而(f) 圖則是從原始影像取出的文字資訊。

(a) (b) (c)

(d) (e) (f)

圖 4 水平文字影像一

(a) (b) (c)

(d) (e) (f)

圖 5 水平文字影像二

由圖 4-5 可看出，本論文提出之方法在水平文字的影像的效果還不錯，取出的文字資訊都很完整；而對於非水平文字影像如圖 6 及圖 7 所示，文字資訊也都完整取出，雖然複雜背景的關係，還有一些雜訊沒有消除掉，但整體取出文字資訊的結果仍佳。

(a) (b) (c)

(d) (e) (f)

圖 6 非水平文字影像一

(a) (b) (c)

(d) (e) (f)

圖 7 非水平文字影像二

最後是彎曲排列的影像圖 8 及圖 9，雖然在複雜的街景中還是能完整的取出文字資訊；由以上的實驗結果圖可以證明我們所提出的方法對於不同排列方向的文字都是有用的，並且在容易產生誤判的複雜邊緣區域，可以透過我們所提出的方法準確消除掉。

(a) (b) (c)

(d) (e) (f)

圖 8 彎曲排列文字影像一

(a) (b) (c)

(d) (e) (f)

圖 9 彎曲排列文字影像二

從實驗結果可以看到，相對於過去複雜的文字偵測方法，我們所提出的新方法重點在於二次利用已經計算好的邊緣特徵值做後續的誤判區塊消除。在多次的實驗統計中，我們發現一個現象，實際包含文字的區塊，其 sobel 邊緣特徵值變化較大(標準差較大)，而複雜背景的部分相對則較小，這是因為文字的筆劃方向較多變，所以 sobel 運算後的特徵值變化也較劇烈。本論文充分利用此一特性，大量縮短了文字偵測演算法運算時間。以類似研究[7][10][11]為例，[10]所提出的方法，平均準確度(Precision)與召回率(Recall)分別為 0.86 與 0.76，P 與 R 的 F 權重(F-measure)為 0.81，運算時間為 7.8 秒; [7]的平均準確度與召回率分別為 0.81 與 0.78，P 與 R 的 F 權重為 0.79，運算時間為 5.36 秒; 而[11] 的平均準確度與召回率分別為 0.87 與 0.72，P 與 R 的 F 權重為 0.78，運算時間為 7.9 秒; 而本論文所提出的方法，以十張各種不同類型的影像實驗結果，平均準確度與召回率分別為 0.56 與 1，P 與 R 的 F 權重(F-measure)為 0.71，運算時間為 0.75 秒，與以往的研究相比，我們的召回率是 1，這代表我們的方法並不會遺失任何文字資訊，雖然也增加了許多誤判小區塊，造成準確度降低，但在運算時間的方面大幅提升了七倍以上。

4. 結論與未來規劃

我們所提出的概念與誤判消除法，不管是在水平或非水平的任意文字皆能 100%在短時間內偵測出來(即使是最大的影像 1280x960，其偵測時間仍小於 1 秒鐘)，雖然系統偵測完的影像還有部分誤判區塊殘留，但整體取出的效果仍佳。未來則要持續改善整個系統準確度與效能，對於邊緣特徵值的運算最佳化，以減少後續的消除誤判區塊的運算，並在不增加運算複雜度的前提下提升準確度。

參考文獻

[1] A.K. Jain and B.Yu, "*Automatic Text Localization in Images and Video Frames*", Pattern Recognition, Vol.31,No 12,pp. 2055-2076, 1998

[2] A.K. Jain and Y. Zhong, "*Page Segmentation Using Texture Discrimination Masks*", Pattern Recognition & Image Process. Lab , Michigan State University,1995.

[3] A.K. Jain and Y. Zhong, "*Page Segmentation Using Texture Analysis*" ,Department of Computer Science, Michigan State University ,East Lansing ,MI 48824,U.S.A.

[4] E.K. Wong and M. Chen, "*A New Robust Algorithm for Video Text Extraction*", Pattern Recognition, vol. 36, pp. 1397-1406, 2003.

[5] J. Zhou, L. Xu , B. Xiao, R. Dai, "*A Robust System for Text Extraction in Video*", ICMV International .Conference on Machine Vision, 2007.

[6] Michael R. Lyu, J. Song and M. Cai, "*A Comprehensive Method for Multilingual Video Text Detection, Localization, and Extraction*", IEEE Transactions On Circuits And Systems For Video Technology Vol. 15, No. 2, February 2005.

[7] N. Sharma, P. Shivakumara, U. Pal, M. Blumenstein and C. L. Tan "*A New Method for Arbitrarily-Oriented Text Detection in Video*", IAPR International Workshop on Document Analysis Systems,2012.

[8] P. Shivakumara, T. Q. Phan, and C. L. Tan,"*New Fourier-Statistical Features in RGB Space for Video Text Detection*", IEEE Transactions On Circuits And Systems For Video Technology , Vol. 20, No. 11, November 2010.

[9] P. Shivakumara, W.H. and C. L. Tan ,"*An Efficient Edge based Technique for Text Detection in Video Frames*" ,The Eighth IAPR Workshop on Document Analysis Systems,2008.

[10] P. Shivakumara, T. Q. Phan, C. L. Tan, "*A Laplacian Approach to Multi-Oriented Text Detection in Video*" , IEEE Transactions On Pattern Analysis And Machine Intelligence,

Vol.33, No.2, Feb. 2011.

[11] P. Shivakumara, R. P. Sreedhar, T. Q. Phan, S. Lu, C. L. Tan, "*Multi-oriented Video Scene Text Detection Through Bayesian Classification and Boundary Growing*" , IEEE Transactions On Circuits And Systems For Video Technology, Vol.22, No.8, Aug. 2012.

[12] P.Y. Chen and C.W. Liang ,"*Text Localization Using Discrete Wavelet Transform and Neural Network*" , Department of Computer Science and Information Engineering Chaoyang University of Technology, 2004.

[13] Rafael C. Gonzalez and Richard E. Woods "*Digital Image Processing(3rd Edition)*",2007

[14] R. Lienhart and F. Stuber, "*Automatic Text Recognition in Digital Video*", in Processing of ACM Multimedia,pp11-20,1996

[15] X. Chen and H.J. Zhong," *Text Area Detection from Video Frames*", IEEE Pacific Rim Conference on Multimedia,pp.222-228,2001

[16] X. Zhao, K.H. Lin, Y. Fu, Y. Hu , Y. Liu, T. S. Huang , "*Text From Corners: A Novel Approach to Detect Text and Caption in Videos* " , IEEE Transactions On Image Processing, Vol.20, NO.3, Mar ,2011

[17] Y. Zhong, Z. Hongjiang, A.K. Jaing ,"*Automatic Caption Localization in Compressed video*", IEEE Transactions On Pattern Analysis and Machine Intelligence,Vol.22,No.4,pp.385-392,April 2000.

[18] Y. Zhang and T. S. Chua, "*Detection of Text Captions in Compressed Domain Video*", ACM Multimedia Workshops,pp.201-204,2000.