以影片彈幕為基礎的自動化關鍵片段及關鍵影格擷取

黃筠雅
資訊工程學系
國立中央大學
桃園，台灣
ilcilc1975@gmail.com

郭同益
資訊工程學系
國立中央大學
桃園，台灣
kuotony860810@gmail.com

陳弘軒
資訊工程學系
國立中央大學
桃園，台灣
hhchen@g.ncu.edu.tw

***Abstract*—尋找一部影片中的代表圖片或者關鍵片段做為預覽圖或預覽短片通常需要依賴人工的編輯與彙整。近期雖然機器學習在影片及圖片上的理解上有許多成功的技術突破及應用實例，但這種方式需要依賴大量的運算資源及訓練資料。本論文將從另一個角度切入：利用影片彈幕找出合適的代表圖片及精彩片段。我們將本文的方法與一個產業界實際使用的工具做比較。實驗結果顯示：這個簡單的方法所擷取影片的關鍵片段及關鍵截圖被大部份的受測者所認可，而且本方法輸出關鍵片段及關鍵截圖的速度遠優於被比較的方法。**

1. Introduction

讓電腦理解影片是一件非常困難的工作。近年的研究大多基於深度學習的技術對各個影格 (frame) 進行物件辨識 (object recognition)，但辨識與文意 (context) 的理解仍有相當的距離。被譽為深度學習之父的 Geoffery Hinton 亦指出：卷積式深層網路 (Convolutional Neural Network, CNN) 在適當的訓練後雖能辨識出圖像中的物件，但物件間的空間關係卻難以被 CNN 直接捕捉 [1]，理解圖像中的物件的空間關係已如此困難，自動化地理解圖片甚至是影片的文意自然更具挑戰。

本論文試圖由另一個角度切入影片理解的問題，並以自動化的影片關鍵片段擷取與關鍵圖片擷取作為技術的出口。具體而言，我們發現：現今有許多的影片網站可以讓使用者在觀看影片同時發表評論，並以彈幕的方式在影片上呈現。彈幕的資訊內容包含評論內容文字、發表時間，且最為重要的是其含有時間戳記，我們可以清楚地得知每條彈幕發布在影片當中的哪個播放時間。這與傳統的留言或評論最大的不同是：傳統留言或評論不具備時間戳記，因此無法將文字內容與影片的特定片段做連結；相反的，彈幕的文字資訊可以在一定程度上可以代表當下影片片段的詮釋資料 (meta-data)，因此，藉由彈幕具備時間戳記的特性和評論的文字，我們有機會讓電腦「理解」影片。

以此為基礎，我們提出一個非常簡單但效果很好的自動化方法擷取影片的關鍵片段與關鍵圖片，並將這個方法命名為 *Busk*。具體而言，我們發現彈幕的出現頻率高的影片片段常常是影片中最值得玩味的片段，因此引發網民的大量留言。目前，我們從網路上爬下 92 萬部影片的彈幕資訊，對其中的 26 部影片所產生的關鍵影片片段及截圖進行主觀分析，我們將 Busk 產生的影片片段及影片截圖與 KKStream 影音串流服務商[1]所研發的代表影格擷取演算法 Stiller [2] 進行比較，由招募而來的 100 名使用者對兩者進行主觀測試。結果發現：使用者對於兩種方法產生之結果皆有正面反饋 (1 至 5 的評分機制皆獲得 3 分以上)。平均而言，使用者給予 Busk 與 Stiller 的平均分數類似，但若針對對於影片內容較熟悉的人 (即：原先就已經看過影片的人) 進行調查，則 Busk 的關鍵影片片段及關鍵圖片截取結果皆優於 Stiller 的結果。除此之外，因為 Busk 的關鍵影片和關鍵圖片生成機制非常簡單，故輸出影片及圖片的的速度遠優於 Stiller。對於影片串流服務提供者而言，採用 Busk 將能在短時間內為大量的影片產生關鍵短片及關鍵截圖供預覽。

[1]https://www.kkstream.com/

2. Related work

大部份的關鍵影格擷取研究是先將影片切割為一幀一幀的影格，利用人像大小、解析度、對比、顏色等等條件挑選出較為優質的樣本，再對樣本影格做分群，在每個群集中挑選一張代表影格，再利用條件篩選出最後的關鍵影格。大部份的關鍵影格擷取演算法希望找到足以代表影片內容的關鍵截圖 [3], [4], [5], [6]，希望使用者能以這些截圖為依據比較方便地找到喜歡的影片。另一方面，某些演算法試著找尋影片中最「吸睛」的截圖，希望使用者能因為截圖而對影片產生期待，這種演算法需要考慮畫面的美感 (如：用色、構圖、對比、銳利度等) [7], [8], [9]。其中，Stiller 演算法 [2] 利用畫面的美感線索及人臉篩選的方法，找出影片中吸引人的畫面做為影片截圖，並以主觀測試驗証：在劇情類的劇集影片中其自動產生的截圖能媲美資深人類編輯所挑選的截圖。

另一方面，由於深度學習在各種非結構化資料 (如：音訊、視訊、影像、文字等) 展現極佳的效果，近期有不少作品使用深度學習技術來尋找關鍵影片截圖、關鍵影片片段及分析圖片和影片的美感 [10], [11], [12], [8]。然而，這些技術需仰賴龐大的訓練資料做為樣本，且模型的訓練及推理需要花費大量的計算資源，模型本身也較難對其產生的截圖或影片片段做出合理的解釋。

我們提出利用群眾力量所生成的彈幕來產生關鍵片段與關鍵影片截圖，這與上述的方法是截然不同的思路。同時，我們的方法能夠為產生的影片截圖或影片片段提供可能的解釋。

3. Methodology

本文的方法分為兩部分，彈幕資料的蒐集與利用彈幕資料自動產生關鍵影片片段及影片關鍵影格。以下說明這兩者的進行方式。

A. 彈幕擷取

彈幕為一種針對影片的特殊評論方式，網民能在觀看影片時進行文字評論，其最大的特徵是時間戳記。一般的評論往往是在影片下方留言並對影片進行評論，但彈幕可以在影片的任何一個時間點發出且直接顯示在影片上面，而其時間戳記就是彈幕對於影片時間軸所發出的時間點，因此，彈幕的文字資訊可被視為是該影片片段的有效詮釋。除此之外，彈幕可以使其他使用者在觀看相同影片時以類似直播的形式看到其對應時間發表的彈幕，藉此引起使用者的共鳴並且讓不同的使用者產生互動。部份影片的鐵桿粉絲甚至會以彈幕的型式為影片中的不完整處提供較為完整的解說。另一方面，評論亦可用來推測使用者對此片段的反應。

我們從 bilibili[2]網站，利用爬蟲抓取各影片的彈幕相關資料，彈幕資訊裡面包含時間戳記、彈幕長度、使用者發送時間以及彈幕內容、及留言者的資訊 (如：使用者等級)。我們一共爬取約 92 萬部影片，並從中隨機挑選一萬部影片進行分析。

B. 關鍵影格及關鍵片段生成

由於使用者可以在任意時間點發出彈幕，因此彈幕大量出現的片段通常代表著特殊事件或情節的發生，因為頻率高的地方相當可能表示在這個區段存在著某些吸引人的元素。基於上述的彈幕的特性，我們目前的實驗先以彈幕的頻率做為關鍵影格及關鍵片段的生成依據。

在生成關鍵影片片段方面，我們計算每部影片中彈幕出現頻率最高的五秒鐘。另外，由於彈幕出現時間可能會有所延遲，所以再多取前後各五秒來當作緩衝，產生總長為 15 秒的關鍵精彩片段。在生成關鍵影格方面，我們以取出的關鍵片段中最中間的影格作為關鍵影格。

4. Experiement

A. 基準方法

我們將本文中的方法與 Stiller [2] 做比較。Stiller 是用來產生代表影片的影格的一套方法。為了讓 Stiller 也能產生影片片段，我們將 Stiller 產生的截圖的畫面結果在影片中的時間戳記前後各取 7.5 秒做為影片的關鍵片段，便可與此研究產生結果進行比較。

B. 實驗設定

由於影片精彩程度為主觀感受，我們只能由主觀測試的結果做為方法好壞的評量依據。我們使用禮券為報酬招募了 100 位受試者。我們選定了 26 部影片，並使用本文的方法及 Stiller 分別為這些影片產生關鍵片段及關鍵影格，要求受試者對這兩個方法的輸出影片及輸出影格進行評分。我們設立了兩種不同的評分機制，其一為二選一，需從 Stiller 的結果及本系統的結果中選出其一覺得較為精彩的；另一種為分數制，使用者必須對 Stiller 及本系統產出的影格及影片片段進行評分，分數從非精彩片段到精彩片段依序為一分到五分。

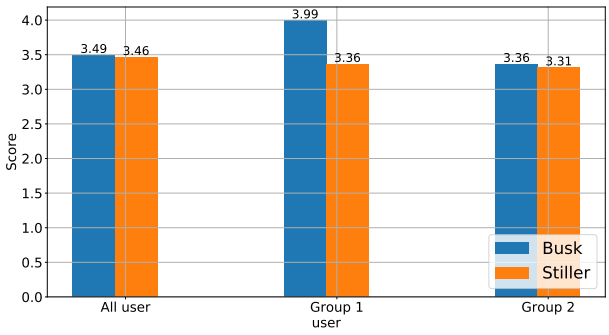
[2]https://www.bilibili.com/

C. 彈幕統計資訊

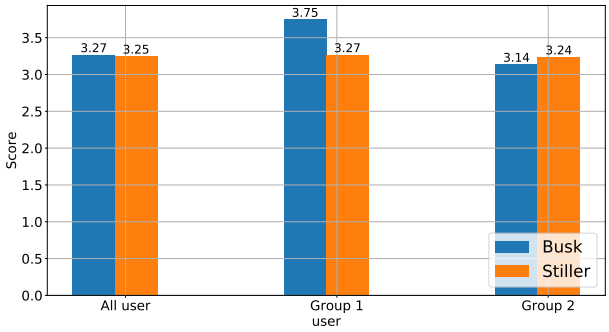
我們從下載的 92 萬部影片中隨機抽取一萬部影片進行字幕的統計分析。

這一萬部影片共包含 2,979,552 則彈幕，平均每部片約有 298 則彈幕，影片的彈幕數量約有 22% 在 10 則以下，但約有 30% 在 100 則以上。另一方面，每則彈幕留言所使用的字數呈現長尾分佈，大部份的彈幕的字數在 3 個至 10 個之間，但也有少部份的彈幕留言多達數十字，平均的字數為 9.78。

D. 整體效果比較



(a) 關鍵 15 秒影片片段擷取結果評分



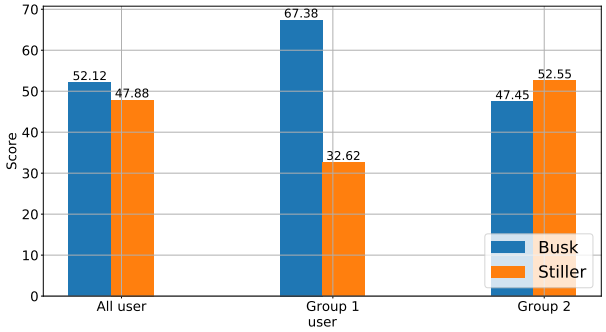
(b) 關鍵截圖結果評分

圖 1. Busk 與 Stiller 的評分結果比較

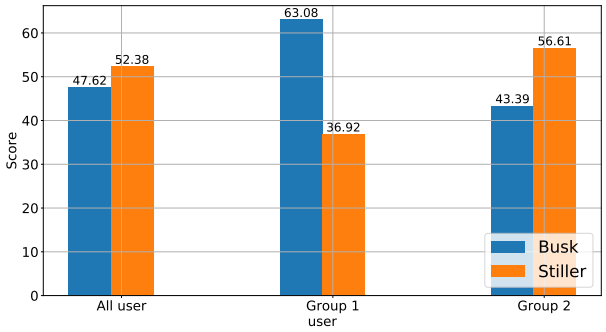
圖 1 展示使用者對本文提出的 Busk 與 KKStream 的 Stiller 所產生的影片片段 (圖 1(a)) 及圖片 (圖 1(b)) 的整體評分結果。使用者對於 Busk 的輸出結果與 Stiller 的輸出結果給予的平均分數十分接近 (參看兩張圖的最左邊的兩條長條及分數)。

另一方面，我們想對影片內容較熟悉的受測者 (如：原本即看過影片的受測者) 對於 Busk 及 Stiller 的輸出的評價做比較，因此我們要求每個使用者在評分前先回答對每部影片的內容是否熟悉或曾經看過。對每部影

片，熟悉該影片內容的人被分至 group 1，不熟悉的被分到 group 2。由圖 1 可看出：group 1 的受測者對於 Busk 的評價明顯高於 Stiller (參看兩張圖的中間的兩條長條及分數)，可見得 Busk 所產生的 15 秒關鍵短片及截圖比較能夠引發熟悉該影片的人的共鳴。



(a) 關鍵 15 秒影片片段擷取結果評分



(b) 關鍵截圖結果評分

圖 2. Busk 與 Stiller 的二選一結果比較

我們也要求使用者直接從 Busk 及 Stiller 的輸出中選擇比較喜歡的輸出，結果如圖 2 所示。平均而言，受測者比較喜歡 Busk 輸出的短片，但比較喜歡 Stiller 輸出的截圖 (參看兩張圖的最左邊的兩條長條及分數)。然而，若只針對熟悉影片內容的受測者 (group 1) 做調查，則發現受測者明顯認為 Busk 的結果較好 (參看兩張圖的中間的兩條長條及分數)。這個結果很值得玩味：雖然 Busk 可能產生了比較能夠代表影片意涵的短片及截圖，但從未看過影片的人可能比較喜歡 Stiller 產生的截圖。因此，Busk 與 Stiller 似乎應該被用在不同的情境：如果是為了要吸引新的使用者觀看影片，採用 Stiller 的截圖可能有較好的效果；但如果是要讓已經看過影片的使用者「回味」，Busk 的結果可能較符合使用者的期待。

E. 不同類型的影片效果比較

除了分析整體效果外，我們並將影片分為六種不同種類的類型，分析各種類型的效果。這六種類型分別為綜藝、遊戲實況、電影、現場音樂表演、MV、以及動畫。表 I 展示的是所有使用者對於 Busk 和 Stiller 在這六類的影片所產出的短片及截圖給的平均分數。當輸出為 15 秒的短片時，除了遊戲實況之外，Busk 的表現均優於 Stiller。當輸出為關鍵截圖時，兩者的效果差不多。

表 II 給出「對影片內容熟悉」的受測者 (即：上節提到的 group 1) 對兩個方法的產出給予的平均評分。這裡我們可以觀察到兩件事：第一，相較於一般使用者 (group 1 + group 2)，對影片的主題較熟悉的使用者通常會對 Busk 及 Stiller 給出的結果給予較高的評分，這代表無論是 Busk 或 Stiller 產生的結果可能都能抓到影片的重點；第二，當我們比較 group 1 的受測者的評分與所有受測者的平均評分，無論在短片輸出任務或截圖輸出任務，Busk 的進步幅度均高於 Stiller，顯示在不同類型的影片中，Busk 的效果都很不錯。

我們仔細看過了測驗的影片及 Busk 和 Stiller 輸出的影片片段及關鍵截圖後，觀察到以下的現象。以綜藝節目而言，各個綜藝節目有自己的流程及橋段設計，忠實觀眾較容易瞭解節目想要呈現的效果，但對於不熟悉這些節目的受測者，某些橋段會讓他們覺得莫名其妙。另外，我們也發現 Busk 很容易找到影片的經典橋段，這可能是因為影片的經典橋段常引起廣大討論「刷屏」，例如：在「葉問」片中的經典台詞「我要打十個」出現的幾秒，出現大量的彈幕留言，也被 Busk 剪輯至 15 秒精華片段中。可能是基於類似的原因，在現場音樂表演、MV、及動畫方面，Busk 的效果也都優於 Stiller，尤其是熟悉影片內容的受測者給 Busk 的評分明顯較高。

F. 結果產生速度比較

我們從 26 部影片中抽樣 5 部，Stiller 產生結果之速度約為影片長度的 16% 至 54%，若長度為一小時的影片，需耗費約 9 分鐘至 32 分鐘才能產生結果。而由於本研究方法僅處理文字檔案，因此速度有極大的差異，一部影片產生結果時間約為 24ms 至 64ms，速度遠優於 Stiller。

5. Discussion

綜合上述，彈幕的時間戳記性質使得我們可以利用使用者提供的資訊來進行影片的標記與分析，利用彈幕出現頻率找尋精彩片段為一個利用其特性最顯而易見的方法，在本研究中也取得相當不錯的成果。然而，目前僅有部份的影音網站支援彈幕這種獨特的使用者介面, 且擁有彈幕的影音網站多偏向次文化，因此，本方法最明顯的缺點就是資料的限制 — 一旦影片不存在其相對應的彈幕，本方法則無法使用。實務上，比較可行的作法可能是融合 Busk 以及其他的關鍵片段擷取及關鍵截圖程式一起使用，讓各個方法截長補短，或者根據影片類型的不同選擇合適的片段擷取與圖片擷取程式。

我們想進一步討論本實驗的部份結果。從表 I 及表 II 可以發現：熟悉影片的受測者對 Busk 及 Stiller 都給予較高的分數，這可能代表 Busk 及 Stiller 真的能夠反映影片的精彩片段，但也可能代表著受測者存在著「經驗偏誤」：對於自己已經熟悉的物品或事件給予較高的分數。倘若受測者真的存在「經驗偏誤」，現今許多依賴受測者回饋的實驗項目都需要重新考量實驗的適切性。

表 I
所有使用者對各類型影片的輸出短片及圖片的平均評分，Busk 和 Stiller 兩者分數較高的以粗體表示之。

輸出格式	方法	綜藝	遊戲實況	電影	現場音樂表演	MV	動畫
15 秒影片	Busk	**3.45**	3.07	**3.65**	**3.68**	**3.63**	**3.76**
	Stiller	3.38	**3.27**	3.48	3.03	3.54	3.28
圖片	Busk	**3.34**	2.83	3.17	**3.57**	3.34	**3.71**
	Stiller	3.27	**3.21**	**3.36**	3.08	3.34	2.97

表 II
熟悉影片內容（原本就看過影片）的使用者對各類型影片的輸出短片及圖片的平均評分，Busk 和 Stiller 兩者分數較高的以粗體表示之。括號內的分數代表與平均分數（表格 I）間的差距，差距較大者以粗體表示之。

輸出格式	方法	綜藝	遊戲實況	電影	現場音樂表演	MV	動畫
15 秒影片	Busk	**3.87** (+.**42**)	3.60 (+.**53**)	**4.01** (+.**36**)	**3.90** (+.**22**)	**4.03** (+.**40**)	**4.14** (+.**38**)
	Stiller	3.54 (+.16)	**3.67** (+.40)	3.51 (+.03)	3.02 (−.01)	3.61 (+.07)	3.26 (−.02)
圖片	Busk	**3.76**(+.**42**)	3.54(+.**71**)	3.34(+.**17**)	**3.64**(+.**07**)	**3.73**(+.**39**)	**3.97**(+.**26**)
	Stiller	3.41(+.14)	**3.7**(+.49)	**3.41**(+.05)	3.08(+.0)	3.45(+.11)	3.05(+.08)

另一方面，同樣根據上述兩個表格，對影片內容較熟悉的受測者傾向於給 Busk 比 Stiller 更高的分數，這雖然暗示著 Busk 可能給予較精確的關鍵片段，但某些項目不熟悉影片內容的受測者則比較喜歡 Stiller 的結果，這可能代表 Stiller 的結果比較吸引一般人。假如我們希望影片截圖或短片能夠吸引使用者觀看整部影片，能「吸引新使用者」的 Stiller 可能比「較精確地抓出關鍵片段」的 Busk 來得更合適。另一方面，倘若需要較正確地抓出影片中較關鍵的片段，則可能應該選擇 Busk。

最後，過去雖然有人研究彈幕這種特別的使用者介面，但大多是研究使用者體驗。就我們所知，本文是第一個針對彈幕與影片的精彩片段進行研究，雖然採用了一個十分簡單的方法，但初步成果顯示這個方向是可行的。同時，由於我們已擁有 92 萬部影片的彈幕資訊，我們將能從彈幕的文字進行自然語言的分析，因此可能的未來研究方向包括：從文字的資訊找出精彩片段、從文字內容觀察使用者對特定片段的情緒及反應等。此外，若是能夠取得使用者 IP 資訊，我們甚至可以觀察不同地區的網民對同一個片段是不是具有不同的反應。例如：我們發現中國的網民在某支足球影片中特別喜歡韓國隊發生失誤的幾個片段，但韓國的球迷可能不會喜歡相同的片段，結合彈幕資訊以及 IP 後，本方法將有可能採用不同類型的使用者的反應，針對不同類型的使用者給予不同的精彩片段輸出。此外，我們也計畫能以彈幕為基礎進行影片「片段」的推薦。目前的影片推薦只能一次推薦一部影片，但如果能利用彈幕資訊建立影片每一個片段的詮釋資料 (meta-data)，則有可能只根據使用者觀看的某一個片段去推薦相關聯的其他影片的片段。

Acknowledgment

我們感謝 KKStream 羅經凱博士對本計畫的建議，及台大資工系曹峻寧同學提供 Stiller 系統的使用說明。

References

[1] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in *Advances in Neural Information Processing Systems*, 2017, pp. 3856–3866.

[2] C. Tsao, J. Lou, and H. Chen, "Thumbnail image selection for VOD services," in *IEEE Conference on Multimedia Information Processing and Retrieval*, March 2019, pp. 54–59.

[3] X. Zeng, W. Li, X. Zhang, B. Xu *et al.*, "Key-frame extraction using dominant-set clustering," in *2008 IEEE international conference on multimedia and expo*. IEEE, 2008, pp. 1285–1288.

[4] Y. Zhuang, Y. Rui, T. S. Huang, and S. Mehrotra, "Adaptive key frame extraction using unsupervised clustering," in *Proceedings 1998 International Conference on Image Processing. ICIP98 (Cat. No. 98CB36269)*, vol. 1. IEEE, 1998, pp. 866–870.

[5] C. Sujatha and U. Mudenagudi, "A study on keyframe extraction methods for video summary," in *2011 International Conference on Computational Intelligence and Communication Networks*. IEEE, 2011, pp. 73–77.

[6] C. Gianluigi and S. Raimondo, "An innovative algorithm for key frame extraction in video summarization," *Journal of Real-Time Image Processing*, vol. 1, no. 1, pp. 69–88, 2006.

[7] Y. Deng, C. C. Loy, and X. Tang, "Image aesthetic assessment: An experimental survey," *IEEE Signal Processing Magazine*, vol. 34, no. 4, pp. 80–106, 2017.

[8] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang, "Rating image aesthetics using deep learning," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 2021–2034, 2015.

[9] L. Mai, H. Jin, and F. Liu, "Composition-preserving deep photo aesthetics assessment," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 497–506.

[10] N. Ballas, L. Yao, C. Pal, and A. Courville, "Delving deeper into convolutional networks for learning video representations," *arXiv preprint arXiv:1511.06432*, 2015.

[11] B. Mahasseni, M. Lam, and S. Todorovic, "Unsupervised video summarization with adversarial lstm networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 202–211.

[12] J. Wan, D. Wang, S. C. H. Hoi, P. Wu, J. Zhu, Y. Zhang, and J. Li, "Deep learning for content-based image retrieval: A comprehensive study," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 157–166.