International Journal of Artificial Intelligence and Neural Networks– IJAINN
Volume 7: Issue 1 [ISSN : 2250-3749]
Publication Date : 06 April, 2017

A Video Text Extraction System

Danial Md Nor, Soo-Ling Wong, Nabilah Ibrahim, Jean-Marc Ogier

***Abstract* – Digital video now plays an important role in entertainment, education, and other multimedia applications. Artificial text in video is normally generated in order to complement or encapsulate the visual content and thus is an important transporter of information that is highly significant to the content of the video. Problem arises when the content of video getting more in time and user facing the struggle to search and edit the desired information. The user cannot get or copy the artificial text directly from video but need to retype the text content in the video. This will consumes a lot of times and work. The main objective of this project is to extract the artificial text in Chinese Simplified and English appear in video and save it into word format. This paper proposed an effective and efficient caption extraction from video. Proposed methods which comprised of video segmentation, image de-nosing, image segmentation and optical character recognition (OCR) are used to extract artificial text in the video. The recognition percentage of text character in this project is up to 97.8%. The recognized text is saved in word format. The video containing different language can be extracted to obtain the text for future work. This system can be improved by extracting the scene text in the video.**

Keywords – digital video, artificial text extraction, OCR.

I. Introduction

Video is graphic multimedia source that has combines an order of images to form a moving images. Video usually has audio constituents that correspond with the pictures being shown on the display. There is no uncertainty that video has now become the most popular media type in our life with growing of digital video devices. Text detection, extraction and recognition of artificial text in video can help a lot in video content analysis and understanding, since text can provide clear and direct explanation of the information consisted in the videos.

Danial Md Nor, Soo-Ling Wong, Nabilah Ibrahim (FKEE)
University of Tun Hussein Onn Malaysia
Malaysia

Jean-Marc Ogier (L3i)
University of La Rochelle
France

The project aims to detect and recognize video images for searching and storing artificial texts through image processing. There are two types of text appeared in the video: scene text and artificial text. Artificial text in video is normally generated in order to supplement or summarize the visual content. Artificial text is artificially added in order to explain the multimedia content while scene text is textual content that was captured by a camera as part of a scene, such as text on T- shirts or road signs. The artificial text appearance is characterized by Lienhart's method [1]. The project aims to detect and recognize video images for searching and storing artificial texts through image processing.

II. Proposed Algorithm

The flow chart of the proposed text extraction process is shown in Fig.1. The colour MPEG video is the input of the algorithm. The video text as the output is the segmented text that recognized by the OCR.

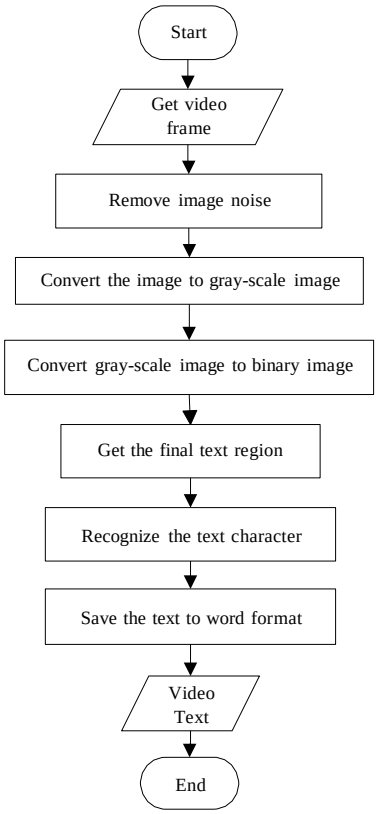

Figure 1. Flow chart for video text extraction system.

A. *Video Segmentation*

Video was extracted into the individual frame or image for video segmentation. The individual image contained in the video will be saved as image in Portable Network Graphics (PNG) file format. PNG is lossless compression and supports true colour or grayscale, and indexed. PNG file has smaller sizes than either Graphics Interchange Format (GIF) or Tag Image File Format (TIFF) while maintain image quality [2]. The image with no repeated artificial word will be selected as the input for de-noising pro-cess. The artificial text with best pixel intensity will be the priority to be chosen.

B. *Image De-noising*

Image noise may arise due to random variation of brightness or colour into image produced by digital camera. It also originate in the unavoidable shot noise of an ideal photon detector. The soft-thresholding rule is chosen over hard-thresholding in wavelet de-noising because it produces more visually pleasant images over hard thresholding [3].

Fig. 2 shows the first level decomposition step of two dimensional (2D) image using DWT which consists up-sampling and filtering. The detail coefficient sub bands is transformed from input image due to the decomposition. The filters L and H shown in Fig. 2 are one-dimensional low pass filter (LPF) and high pass filter (HPF) respectively for image decomposition and original image decimated by two after the filter operation. This improves the frequency resolution as the frequency uncertainty is reduced by half [4]. The LL sub band comes from low pass filtering in both directions and it is the most like original picture. The HL comes from high pass filtering in the horizontal direction and low pass filtering in the vertical direction, known as the horizontal fluctuation. The LH sub band comes from low pass filtering in the horizontal direction and high pass filtering in the vertical direction and, known as the vertical fluctuation. The HH sub band known as it comes from high pass filtering in both direction so it is Diagonal Fluctuation.

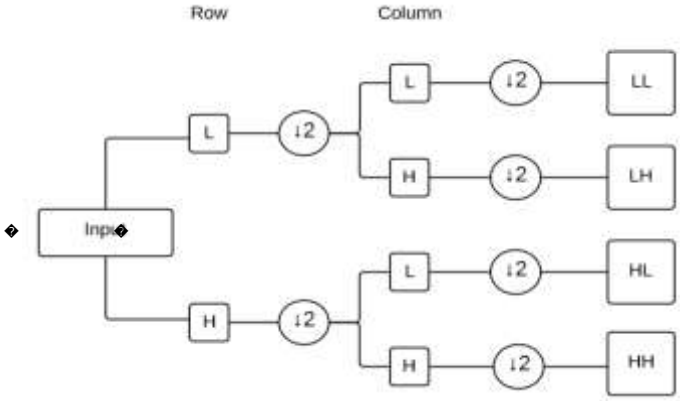

Figure 2. First level decomposition of 2D DWT.

Fig. 3 shows wavelet reconstruction process where consists of up-sampling and filtering. The detail coefficients can be assembled back into the original signal without loss of information. First level sub bands is multiply by two before the summation of filter operation. The L and H is added to get the reconstructed image from IDWT (Inverse Discrete Wavelet Transform).

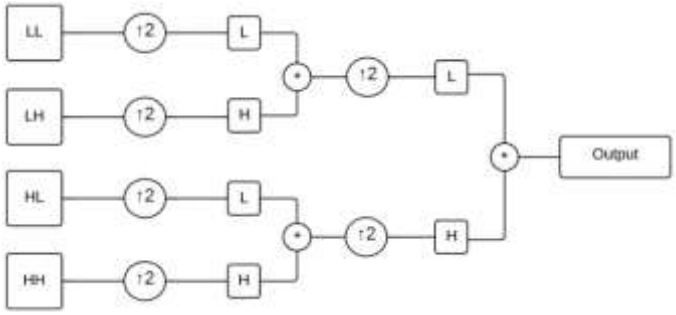

Figure 3. IDWT two 2D image with first level reconstruction step.

The sub band LL alone is further decomposed into four sub bands labelled as LL1, LH1, HL1 and HH1 to obtain the second level of decomposition, from Fig. 3 as shown in Fig. 4.

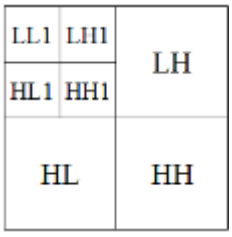

Figure 4. DWT based image decomposition second level decomposition.

Universal threshold [3] is formulated as,

$$T=\sigma\sqrt{(2\log M)} \quad (1)$$

where σ^2 is the noise variance and M is the number of pixels. As M increases, bigger threshold can be get, which tends to over smoothen the image. The threshold is subtracted from any coefficient that is greater than the threshold for soft thresholding as (2). This moves the time series toward zero.

$$coef[i] = coef[i] - thresh \quad (2)$$

The Peak Signal to Noise Ratio (PSNR) is the ratio between maximum possible power and corrupting noise that affect representation of image. PSNR is usually expressed in decibel scale, dB. The PSNR is commonly used to measure of quality of reconstruction image. The signal in this case is original data and the noise is the error introduced. It is defined via the Mean Square Error (MSE) and corresponding distortion matric, the Peak Signal to Noise Ratio [5].

MSE assesses the quality.

Second decomposition of Haar two-dimensional (2-D) DWT is used to de-noise noisy image or original image since noise can degrade image quality and increase difficulty to extract text. DWT is multiresolution analysis that can analyse image at different frequency with different resolution. It can provides sufficient image information for analysis and synthesis. Haar DWT is the simplest among wavelets thus reduce computation time [6]. Fig. 5 shows the two dimensional image de-noising. Noise image in RGB is transformed into coefficients sub-bands in Figure 4 is obtained by using Haar 2-D DWT. Noise variance is estimated from noisy image. Threshold value is calculated

using universal thresholding as (1). Soft thresholding function will be applied to the sub-band coefficients except the low pass or approximation sub-band to change the coefficients. Unscaled white-noise is used to mask unwanted sound in the noisy image using sub-bands coefficient because it contains many different frequencies of sound [7]. Then, the IDWT is used to get the reconstructed image using reconstructed filter, HPF and LPF which noise is entirely suppressed from noise image.

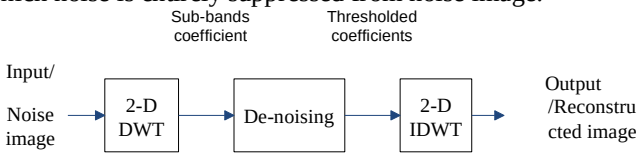

Fig. 5 Block diagram of two dimensional image de-noising.

C. *Image Segmentation*

The actual aim of the segmentation step is to divide the pixels of each frame of a video into two classes:

i. region which do not contain text and
ii. region which possibly contain text.

Region which do not contain text are discarded and regions which might contain text is kept. It is called as candidate regions since it is not exactly a superset of the character regions. Equation below is used to convert the coloured or RGB image into grayscale image, Y.

$$Y = 0.299R + 0.587G + 0.114B \quad (3)$$

The intensity of the text edges is higher than that of the non-text edges. Thus, an appropriate threshold value can be selected and non-text edges in the detail component sub-bands preliminarily removed [1].

The thresholding operation is a grayscale value remapping operation g defined by

$$g(x,y) = \begin{cases} 0, & f(x,y) < T \\ 1, & f(x,y) \geq T \end{cases} \quad (4)$$

Where (x,y) represents a grayscale value, T represent threshold value, g(x,y) represents threshold image and f(x,y) represents grayscale level image pixels. Each pixel value in input image is compared to the threshold value, T [6]. This method is fast and simple [8]. The pixel will be binarized as white or black when the input pixel exceed the T or less than T respectively Image is simplified for easier analysis. Text pixels are separated from the background [8]. The character pixels of text separated from background pixels formed image that contains only white character pixels on a black background [8]. Image segmentation with thresholding is used to obtain potential text blocks [1].It is necessary to perform pre-processing procedure to extract artificial text before a recognition algorithm is applied.

D. *OCR Recognition*

OCR is performed on the segmented text images. Optical patterns corresponding to alphanumeric or other characters is classified using Optical Character Recognition (OCR) [10]. The output is directly passed to a standard OCR software package in order to translate the segmented text into ASCII [11]. The ASCII then is able to be saved and searched using online search engine.

Experiment was carried out on video in MPEG-4 format. The video image sequences are obtained. The video image with 480x640 resolution in PNG format that consisted of different artificial text character is the input of the Haar DWT. Universal thresholding is used to threshold the detailed coefficient of second level decomposition in order to remove noise in the image. The threshold value used for different image frame calculated using (1) and threshold value is shown in Table 1.

TABLE 1
Threshold value for first and second level decomposition of Haar DWT

Input noisy image	Variance, σ^2		First level threshold value,T1	Second level threshold value,T2
	First level, L1	Second level, L2		
058.png	1.3676e+03	1.3750e+03	17.36	17.41
150.png	2.1724e+03	2.3232e+03	21.88	22.63
198.png	3.3944e+03	3.5231e+03	27.35	27.86
371.png	1.1673e+03	1.2055e+03	16.04	16.30
455.png	1.0203e+03	1.0695e+03	15.00	15.35
610.png	2.5610e+03	2.3751e+03	23.76	23.88
733.png	4.6155e+03	4.7344e+03	31.90	32.31

Table 1 above shows variance and threshold value for first and second level decomposition in each input noisy image. First level variance is get from the input noisy image. While the second level variance is get from first level Haar DWT de-noised image. Second level variance is higher than first level variance because more noise is removed from image. Higher variance meant larger pixel range in image. Thus, image has better intensity and details in image can be displayed better. The variance in Table 1 is used in (1) to calculate the universal threshold for the corresponding of frame image.

TABLE 2
Threshold value for different input noisy image.

Input image	T
058coe.png	135
150coe.png	136
198coe.png	94
371coe.png	122
455coe.png	112
610coe.png	105
733coe.png	105

Since the variance is used to calculate the universal value while the number of pixel is constant with 480x640 pixels. The increased in variance from first level to second level Haar DWT results the increased in threshold value.

Table 2 above shows the input image is the original video frame image after Haar DWT de-noising and its corresponding threshold value, T. The de-noised images are converted into the the grayscale images using (5). The grayscale images are further thresholded using (6) to binary image. Different input image have different intensity. Thus, different threshold value

needed to threshold the image in order to obtain the clear text character. Intensity of the text edges is higher the non-text edges. The grayscale image consists of 0 to 255 intensity level. The threshold value is selected from 0 to 255. If input pixel of grayscale image smaller than T, each pixel is binarized as black or '0'. If input pixel of grayscale image greater or equal T, each pixel is binarized as white or '1'. Colour of background and text is changed to black and white colour respectively. Foreground objects is separated from the background in image.

TABLE 3
MSE and PSNR for original image and reconstructed image.

Original image	MSE for first level de-noising	MSE for second level de-noising	PSNR for first level de-noising/decibel(dB)	PSNR for second level de-noising/decibel(dB)
058.png	21.1215	21.1107	34.8835	34.8858
150.png	67.8700	25.1652	29.8135	34.1228
198.png	113.4677	39.6629	27.5820	32.1470
371.png	26.2087	11.8203	33.9463	37.4045
455.png	24.4763	13.5848	34.2433	36.8120
610.png	77.7950	67.1000	29.2213	29.8636
733.png	143.7465	67.6213	26.5548	29.8300

Table 3 above shows the MSE and PSNR between original image with first and second level IDWT image. The MSE of second level de-noised image is decreased from first level de-noised image MSE indicated that the level of image distortion with relatively small image modification is decreased. The MSE obtained using (3) is used in (4) to calculate the PSNR. The PSNR of second level de-noised image is increased from first level de-noised image PSNR since the value of PSNR is inversely proportional to the MSE value. The increased in PSNR value resulted by decreased in MSE value from first level de-noised image to second level de-noised image. The PSNR values is high for first and second level de-noised image because many noise is removed from image. Higher PSNR in second de-noised meant more noise is removed from second level decomposition and better image quality is obtained.

The seven video frame text character with Chinese Simplified and English language in different letters is extracted. The text character extracted from video frame include letters, numerical digits, common punctuation marks and whitespace.

TABLE 4
Precision and recognition rate of text character.

Input Image	Number of text character in video	Number of text character extracted correctly
058.png	27	27
150.png	42	40
198.png	62	59
371.png	73	71
455.png	44	44
610.png	80	79
733.png	85	84
	413	404
	Precision= $\frac{404}{413}$ = 97.8% Recall= $\frac{9}{413}$	

Table 4 above shows the recognition percentage of video frame image with different text word. The precision is high which is $\frac{404}{413}$ and the recall is low which is $\frac{9}{413}$. The recognition percentage for the system is up to 97.8%.

TABLE 5
Recognition rate according to the type of language.

Input Image	Language			
	Chinese Simplified		English	
	Number of text character in video	Number of text character extracted correctly	Number of text character in video	Number of text character extracted correctly
058.png	3	3	24	24
150.png	9	8	33	32
198.png	13	13	49	46
371.png	17	16	56	55
455.png	8	8	36	36
610.png	17	16	63	63
733.png	18	18	67	66
Total	85	82	328	322
	82/85≈ 96.5%		322/328≈98.2%	

Table 5 above shows the recognition rate of different video frame with different language. The Chinese Simplified text character is recognized better than English text character although the recognition of English is higher than the Chinese Simplified. Because there is large number of English text character to be extracted and incorrect recognition is decreased.

III CONCLUSION

In this paper, an efficient and effective caption extraction method for video is proposed. The Haar wavelet transform is used to remove noise of PNG images of MPEG-4 video sequences. The selective threshold value, T is used to remove complexity of image by converting RGB image into binary image. The MATLAB function is developed on video acquisition, video segmentation, image de-noising and image segmentation. Artificial text in videos often carries the most important information. This information may help the video indexing and video content understanding. When the text is recognized by OCR, the accuracy of the methods can be improved. The recognition percentage of text character can up to 97.8%. The text in the video can be saved or searched for the translation, learning and documentation purpose.

Acknowledgment

The authors thanks Ministry of Higher Education Malaysia for the support through MDC Multidisciplinary Fundamental Grantt U091. We also would like to thank to the University Tun Hussein Onn Malaysia and University of La Rochelle for giving the opportunity to learn by means of gaining so much knowledge.

References

[1] Lienhart R. and Effelsberg W. (2000). Automatic Text Segmentation and Text Recognition for Video Indexing. ACM/Springer Multimedia Systems, vol.8, pp. 69-81.
[2] Trevor Morris Phototgraphics. (2014).The PNG File Format. Retrieved on May 6, 2015, from http://morris- photographics.com/photoshop/articles/png-format.html
[3] Kumar, V., & Kumar, A. (2013). Simulative analysis for Image De-noising using wavelet thresholding techniques, 2(5), 1873–1878.
[4] Lei, L., Wang, C., & Liu, X. (2013). Discrete Wavelet Transform Decomposition Level Determination Exploiting Sparseness Measurement, 7(9), 691–694.
[5] Ouni, S. (2012). A New No-reference Method for Color Image QualityAssessment, 40(17), 24–31.
[6] Chung-Wei Liang and Po-Yueh Chen, DWT Based Text Localization, International Journal of Applied Science and Engineering, 2(1), pp. 105 -116, February 2004.
[7] HowStuffWorks. (2015). What is white noise?. Retrieved on May 11, 2015, http://science.howstuffworks.com/question47.htm
[8] Jung, K. (2004). Text information extraction in images and video: a survey.Pattern Recognition, 37(5),977–997. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0031320303004175
[9] Malobabi, J., Connor, N. O., Murphy, N., & Marlow, S. (n.d.)., "Automatic Detection and Extraction of Artificial Text in Video". 5th International Workshop on Image Analysis for Multimedia Interactive Services, 2-5.
[10] Jain, a. K. (2000). Automatic caption localization in compressed video. In IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, pp. 385–392.
[11] Yusufu, T., Wang, Y., & Fang, X. (2013). A Video Text Detection and Tracking System. 2013 IEEE International Symposium on Multimedia, 522–529.

About Author (s):

Danial Md Nor received his Ph.D. in video image processing and retrieval from joint PhD programme between ULR, France and UTHM in 2016. Currently, he is the senior lecturer at Faculty of Electrical and Electronic Engineering, UTHM. Danial is a Member of IEEE Malaysia Section, EmbCOS FKEE UTHM and Associate Member of UACEE. He has published more than 20 papers in journals and conferences. He is also reviewer for CAMP, KMICE and was Head of Research Grant for Content-based Indexing and Retrieval of Low Resolution Document in 2009. He is member of several groups for ongoing research grants in image processing, content-based image and video retrieval, embedded system, computer network and Document Analysis and Content Management.

Wong Soo Ling was born in Sarawak, Malaysia. She received the B.Eng in electrical engineering with Hons. from UTHM Malaysia, in 2015. Soo Ling is a Member of IEM Student Chapter. Her current research are in image processing, content-based image and video retrieval.

Nabilah Binti Ibrahim was born in Kedah, Malaysia. She received the B.Eng degree in Communication Engineering in 2007, and M. Eng in Computer Science in 2009, both from Shibaura Institute of Technology (SIT) Japan. In 2013, she obtained her PhD in Electrical Engineering from Tohoku University Japan. From 2008, she joined the Department of Electronic Engineering, UTHM, Malaysia, as a Tutor and became a Senior Lecturer in 2013. Currently, she is an International Coordinator of Japan, China and Korea in International Office, UTHM. Her current research interests include image processing, signal processing, and diagnosis of cardiovascular disease. Nabilah is a Member of IEEE Malaysia Section, BEM, and IEM.

Jean-Marc Ogier received his PhD degree in computer science from the University of Rouen, France, in 1994. Now full professor at the University of La Rochelle, Professor Ogier is the President of University of La Rochelle. His works mainly of Document Analysis and Content Management. Author of more than 160 publications / communications, he managed several French and European projects dealing with historical document analysis, either with public institutions, or with private companies. Between 2005 and 2013, Professor Ogier was a Deputy Director of the GDR I3 of the French National Research Centre (900 researchers depending on the French National Research Center CNRS). He is also Chair of the Technical Committee 10 (Graphic Recognition) of the International Association for Pattern Recognition (IAPR), and is the representative member of France at the governing board of the IAPR. Jean-Marc Ogier has been the general chair of the program chair of several international scientific events dealing with document analysis (DAS, ICDAR, GREC, etc)