Encouraging the Usage of Neural Network in Video Text Detection Application

Suresh Kumar Balasubramaniyan, Praveen Kumar Mani,
Arun Kumar Karthikeyan, and Ganesh Shankar

Blekinge Tekniska Högskola,
School of Computing (Sektionen för datavetenskap och kommunikation),
37179-Karlskrona, Sweden
{suresh.draco,praveencse88,aruncs08,gshindi}@gmail.com

Abstract. The video text detection is an expert system in which we study how the performance can be enhanced by adding neural network. The implementation of video text detection using algorithm based approach [9] is taken and compared with the neural networks based implementation [11]. A standard protocol [10] for evaluating the video text detection approach is taken and its metrics are used for the comparative study. With this comparison, the evaluation of both the systems for better performance can be done. The conclusions necessary for enhancing the usage of neural network is drawn based on the comparison study. The paper is about encouraging the use of neural network in an expert system (Video Text Detection Application).

Keywords: Neural networks, Expert system, OCR (Optical Character Recognition) and video text detection.

1 Introduction

The expert system [5] is basically a computer program which repeatedly possesses different questions each with a yes or no answer. This process is almost like following through a maze which has many left or right turns. It is done by giving yes or no answer for turning right or left. Usually expert systems start with a small base of information. There are many thousands of expert systems commercially available which exist in applications such as medicine, chemistry and engineering field. The already known limitation of expert system is that its inability to learn new knowledge on its own. But a new knowledge can be added from the outside manually to the knowledge base [6].

The neural network system "video text detection [11]" and an expert system "overlay text detection [9]" are taken to compare the performance of the video text detection. Both systems use different architecture to detect the text from the video. The following paragraphs briefly discuss about existing video text detection methods and how the performance, efficiency etc can be improved.

Text detection is essential to video information retrieval and indexing. Current methods cannot handle diverse contrast or fixed in a complex background. To handle

N. Meghanathan et al. (Eds.): CCSIT 2012, Part II, LNICST 85, pp. 231–240, 2012.

these difficulties, this paper propose a capable text detection approach, which is based on invariant features, such as edge strength, edge density, and horizontal distribution. Text in video is an extremely compact and an accurate clue for video indexing and summarization. Most of the video text detection and extraction techniques have guess on background contrast, text color, and font style. But only few methods can handle multi-language texts well because various languages have various writing style. New results on a different number of video images and comparisons with other methods are reported in detail.

There are many researches going on text detection for document analysis and text based video indexing. Text recognition methods are typically either connected component-based or texture based. In the connected component-based process it can perceive text regions efficiently, however there are numerous difficulties for other graphical objects; this is basically happen because in digital video text is embed in complex backgrounds and the inspection that texts in video images contains distinct texture properties. Through these explanations, text detection can be posed as texture classification problem where the exact problem knowledge is available prior to classification. Video text detection presents a number of challenges since the properties of text can vary. For example, the text in video varies considerably in font size and style; the intensity of text pixels can be higher (normal text) or lower (reverse text) than that of background; and the text can be distorted or blurred due to sensor properties or due to camera or object motion. Other problems, such as lighting and background variation, make the task even more difficult. Previous work on text detection has focused on the detection of graphical text superimposed on a video frame [1, 2, 3, and 4]. Although these methods are able to detect graphical text under certain constraints, it is much harder to achieve good performance for scene text, where those constraints are usually not satisfied.

2 Background and Related Work

In video text detection, there is a problem of text location in digital video as supervised texture classification. To overcome this issue support vector machine (SVM) is used for the texture classifier. The SVM is considered as the expert system application and compared with the neural network based text detection method [12].

To detect texts in an image, SVM shifts the detection window over all locations in the image. This will only detect texts at a single scale, however. To achieve multi-scale detection, SVM incrementally resize the image and run the detection window over each of these resized images. The experiment was performed on 2500 key frames with a size of 320x240 manually selected from 200 Korean news archives and 200 commercials. SVM method detected 94.5 % of the text regions in a set of 1,500 test images with false-detection rate of only 4.2%. Errors occurred primarily because of low resolution. The same experiment is done with neural networks (NNs) [13]. The network has two hidden layers of sizes 50 and 30, respectively and was trained by back propagation algorithm minimizing mean squared error. To avoid the local minima, reported results with NN was obtained by training 10 networks with different initial weights, and selecting the minimal error over all the results.

In contrast to the previous approaches where either neural network or pure expert system is used to Video Text Detection application, we enforce the usage of both neural network and expert system. Our idea is based on the observation that text regions typically are rich of corners and edges and corners and edge points are nearly uniformly distributed in text areas. There are four features we used in this approach: corner density, edge density, the ratio of vertical edge density and horizontal edge density, and center offset ratio of edges [14].

The comparison of the algorithm (HUA) with the other three text detection schemes using our PE protocol on the above mentioned testing data is done. The following criteria's are mainly analyzed; clips, textboxes, missed textboxes, false alarms, detection rate. The first scheme (denoted by QI) for comparison is from "Integrating visual, audio and text analysis for news video" [15], the second one (XI) is from "A video text detection and recognition system" [16], and the third one (XI-2) is an improved version of the second one, in which detection results in consecutive frames are used to enhance the final performance. The evaluation results of the four algorithms are listed in the table [10], algorithm HUA produce better detection results than the other three algorithms.

3 Research Question

At the end of the systematic literature review, we formulated a research question for the comparison study to reveal the usage of neural network and its improvement.

R.Q.1: Which of the system among feed-forward neural network and expert system based on overlay procedure has better performance on video text detection?

We are making a protocol based comparison with the help of metrics proposed by Xian-Sheng Hua [10]. Here we are taking feed-forward neural network based video text detection proposed by Huiping Li and David Doermann [11], and comparing its performance with the algorithm based system proposed by Wonjun Kim and Changick Kim [9].

4 Implementation of Video Text Detection Using Neural Network

This system is proposed by Huiping Li and David Doermann [11]. Here the architecture uses feed-forward neural network for the classification process in the text detection process. The architecture needs the external training from the user by providing the ground-truth of the video text. The neural network is tuned up with the training process and further it performs the detection process without the manual help from the user side.

4.1 Training Process

The architecture uses external tool for generating ground-truth values for the sample set of video frame images. The input is stored in the text file and given to the neural

network. The main training sample is split into text and non-text samples. Text sample is given as the sample digital frames images with fixed window values showing the blocks in which the text and the background is defined. The non text samples are given with the blocks of background textures in the same window. The proposed architecture uses a four step algorithm to select the text blocks in the sample window given. After the manual training from the user side, the architecture is trained with the bootstrap method without the user intervention.

4.2 Classification Process in the Neural Network

The text block and the non text block is classified using binary notation with 1 to represent text block and 0 to represent non-text block. It results in the binary frame with 1's and 0's. The parameter that increases the complexity in the detection process is the skew angle. It is the angle between the text box axis and the text direction's vertical axis. It is 90 degree when we are dealing with the straight letters.

5 Implementation of Video Text Detection Using Expert System on Overlay Text

The architecture proposed in this system [9] is an algorithm based approach that determines overlay text from complex background. The explanation to the architecture proposed in this paper is as follows. It is almost intended to determine the subtitles of the video scenes. The following is the procedure or algorithm implemented in this implementation.

5.1 Overlay Text Region Detection

The overlay text will be the opposite texture to the background that is, if the background is dark then overlay text is bright. It is determined that the color of the text changes logarithmically with the color of the overlay. The transition map is found in this way for all frames.

$$d(T_n, T_{n-3}) = \sum_{(x,y)\in T} (T_n(x,y) \otimes T_{n-3}(x,y))$$
$$\text{if}(d(T_n, T_{n-3}) < th) TR_n = TR_{n-3}$$
$$\text{otherwise, find new } TR_n$$

Where T_n and T_{n-3} represents the transition map value of n^{th} and (n-3) rd frame respectively. Then if the values of n^{th} frame and (n-3) rd frame are same then overlay value is set to be 0 and 1 if the values are different. Similarly the value of TR in the formula is the detected overlay text regions in the n^{th} and (n-3) rd frame.

5.2 Overlay Text Region Extraction

Gray scale text region is given as input to the color polarity computation and with that the color polarity is determined. The inside region is filled with pixels and the final text is extracted out.

6 Customized Protocol for Performance Evaluation of Video Text Detection Method

In this section we have gathered data for formulating a protocol based on which both the implementation taken for comparison will be evaluated.

The performance metrics are briefly explained in this section. The parameters are taken from the protocol defined by Xian-Sheng Hua [10]. The limitation in this protocol is that they are not categorized, which when categorized can be enhancing the comparison process. In this paper we are customizing this protocol as a category based protocol as follows, (this classification is in context with the process where this parameter can be evaluated in the video text detection procedure).

- Textbox detection
  - Textbox Location.
    - In the given video frame or the image, the text location represents the co-ordinates(x,y) of the textbox
  - Textbox Height.
  - Textbox Width.
- Text property
  - Text String.
  - Text Length.
  - Character Height Variation.
- Detection complexity
  - Skew Angle.
    - It is the angle between the x-axis of the image and the horizontal axis of the text block.
  - Background Complexity.
  - Colour and Texture.
  - String Density.
    - It is width of the string or text box.
  - Contrast.
  - Recognisability Index.
    - It varies from 0 to 3 which vary according to human recognisability.
- Accuracy
  - False Alarm Rate.
  - Detection Rate.
- Processing speed.

7 Research Implementation

The neural network system "video text detection [11]" and an expert system "overlay text detection [9]" are taken to compare the performance of the -video text detection. Both systems use different architecture to detect the text from the video. The parameters used by the system are Textbox Location, Textbox Height, Textbox Width, Text String, Text Length, Character Height Variation, Skew angle, Color and Texture, Background complexity, String density, Contrast and Recognisability Index. These parameters are determined by both the neural network and an expert system, but at the different stage of video text detection architecture.

7.1 Video Text Detection System, a Neural Network System

First the architecture is analyzed to find at which stage the parameters that determine the performance of this system are calculated [11]. In this system the training process is initially conducted to analyze the background of the image with the text. Then the first phase of the architecture is feature extraction. In this phase the text length, character variation, color and texture of the text parameters are determined from this phase. Next phase is classification; here textbox location, textbox height, textbox width, background complexity and contrast color from the background are analyzed. Next phase is skew estimation, here skew angle-the angle which formed between the vertical line and inclination of the text. Finally text block is generated.

7.2 Overlay Text Detection System Based on Automated Testing, an Expert System

Overlay text detection system based on automated testing, an expert system is analyzed as follows. First the architecture is analyzed [9] to find at which stage the parameters that determine the performance of this system are calculated. The system has Overlay Text Region Detection and Overlay Text Extraction. In overlay text region detection phase, the text region that is textbox parameters which covers the text is found. They are Textbox Location, Textbox Height and Textbox Width. Transition map is generated and then candidate region is extracted from the generated map. Then the overlay text region is determined, refined and updated. Then in text extraction phase the Text String, Text Length, Character Height Variation, Skew angle, Color and Texture and Background complexity parameters are determined. Especially in color polarity computation phase the color, texture and background complexity are determined.

7.3 Performance Evaluation

7.3.1 Goal

- To have comparison study on both neural and expert oriented implementation of video text detection application.

- To find in which area which implementation shows high performance. (Efficiency, speed etc)

In the analysis procedure we analyzed the system to determine in which stage of the implementation the parameters described in the protocol can be found. The values of each parameter for both the implementations are obtained separately from [8], [9], [10] and [11]. With the idea from the analysis procedure, we compared the parameters of both implementations as follows,

System 1: Implementation using neural networks [11].
System 2: Implementation using expert system with algorithm based approach [9].

7.3.2 Experiment Explanation

The parameters like recognizability index and skew angle are the properties of the text that explain how difficult the text is to be identified. We are going to compare this parameter with each implementation using the false alarm rate and processing speed. The false alarm rate is the number of failed detection process done by the system and processing speed is the time taken for the process. Now false alarm rate of implementation of system 1 is plotted in 2D graph along with the false alarm rate of system 2. With this the analysis of the performance of each system can be made.

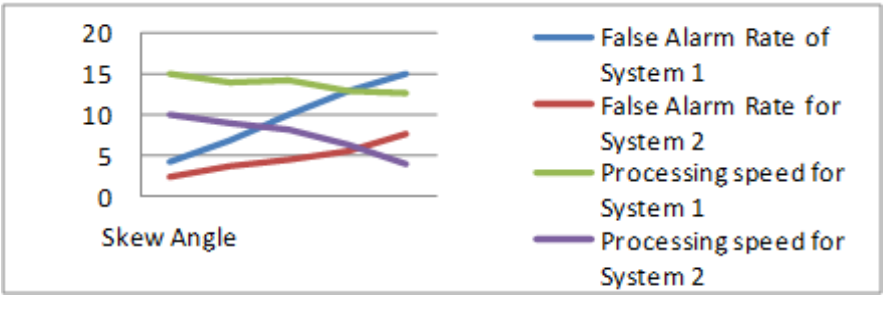

Fig. 1. Evaluation based on Skew Angle

In this analysis the parameters false alarm rate and processing speed (Frames per second) is evaluated by increasing the skew angle (X-axis). The increase in the skew angle means, according to the protocol defined it increases the complexity of the detection process.

Analysis Report. Neural network based implementation shows less in efficiency based on false alarm rate that is higher false alarm rate and expert system based on algorithm based approach shows higher efficiency that is lower false alarm rate. There is a variation at the top of the curve used to denote false alarm rate of system 1, it is because the neural network produces less false alarm rate with the high amount training.

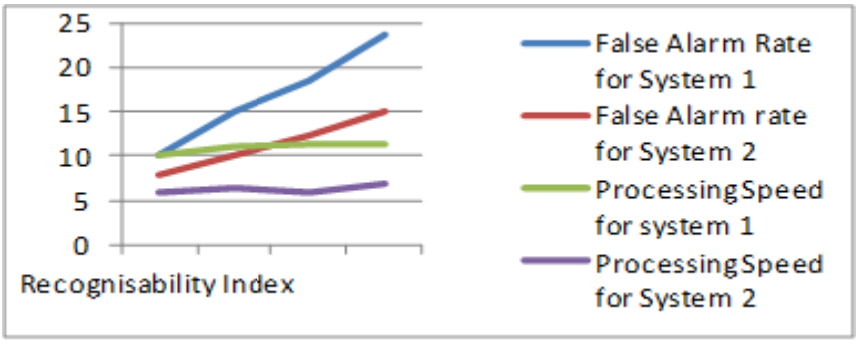

Fig. 2. Evaluation based on Recognisability Index

In this analysis report we are deducing the two performance evaluation criteria False Alarm Rate and Processing Speed with the help of Recognisability Index.

Analysis Report. It is to the report generated with the evaluation using Skew Angle. And also the other complexity detection parameters also produced the same result as the skew angle wherein the neural network based system produces better processing time and higher False Alarm Rate.

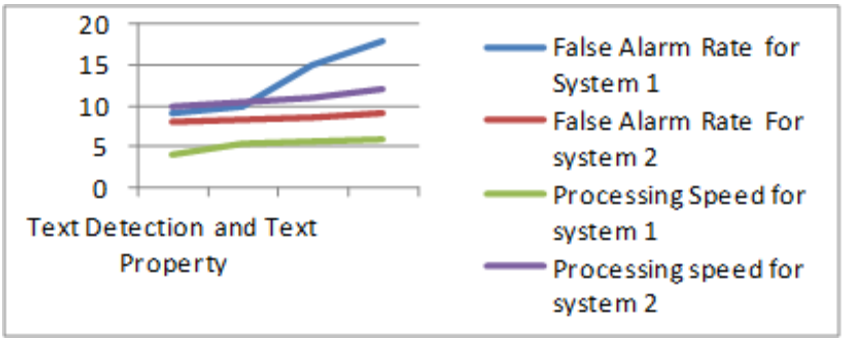

Fig. 3. Evaluation based on Text Detection and text property

Analysis Report. The performance with reference to the text detection parameters and the text property parameters of both the system shows similar variations of skew angle report.

All the above findings are obtained individually from [9] and [11] which are the different implementations of video text detection application.

8 Conclusions and Result of the Research

The performance of neural network in the expert system video based text detection is deduced with comparison to a similar expert system with the same goal but in

algorithm based approach. The evaluation resulted to the conclusion that usage of neural network enhances the performance in terms to processing speed considerably.

The neural network makes the classification faster than the algorithm oriented approach. But when taking accuracy as criteria for the performance, usage of neural network drops down. The accuracy decreases in the neural based system with the increase in the complexity involved in the text detection process. But with the proper and required training given to the neural network oriented system, efficiency can also be improved.

To enhance the usage of neural network in expert system, the training process in which the frame is given as sample is to be enhanced with appropriate learning method. In the system we took for comparison the training method involves ground-truth value which does not significantly increase the accuracy as similar to the algorithm based approach.

Finally, to enhance the usage of neural network in expert systems, we should increase the accuracy instead of processing speed. With the good training method added to the neural network based implementation, efficiency factor also can be increased. Because, accuracy is more important than processing speed in terms of efficiency.

References

1. Kim, H.: Efficient Automatic Text Location Method and Content-based Indexing and Structuring of Video Database. Journal of Visual Communication and Image Representation 7, 336–344 (1996)
2. Sato, T., Kanade, T., Hughes, E., Smith, M.: Video OCR for digital news archives. In: Proceedings of IEEE Workshop on Content-Based Access to Image and Video Databases (1998)
3. Lienhart, R., Stuber, F.: Automatic text recognition in digital videos. In: Proceedings of ACM Multimedia, pp. 11–20 (1996)
4. Shim, J., Dorai, C., Bolle, R.: Automatic text extraction from video for content-based annotation and retrieval. In: Proceedings of ICPR, pp. 618–620 (1998)
5. Carretero-Diaz, L.E., Lopez-Sanchez, J.I.: The Importance Of Artificial Intelligence-Expert Systems- In Computer Integrated Manufacturing. In: Proceedings of the International Conference on Engineering and Technology Management, IEMC 1998. Pioneering New Technologies: Management Issues and Challenges in the Third Millennium (Cat. No.98CH36266), pp. 295–301 (1998)
6. McCarthy, J.: Generality in artificial intelligence. Commun. ACM 30(12), 1030–1035 (1987)
7. Carling, A.: Introducing Neural Networks. SIGMA PRESS-Wilmslow, United Kingdom
8. Anderson, M.F., Cohen, M.E., Hudson, D.L.: Combination of a Neural Network Model and a Rule-Based Expert System to Determine Efficacy of Medical Testing Procedures. In: Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Images of the Twenty-First Century (Cat. No.89CH2770-6), vol. 6, pp. 1991–1992 (1989)
9. Changick, K., Wonjun, K.: A New Approach for Overlay Text Detection and Extraction from Complex Video Scene. IEEE Transactions on Image Processing (February 2009); Sponsored by IEEE signal processing society

10. Zhang, H.-J., Liu, W., Hua, X.-S.: An Automatic Performance Evaluation Protocol for Video Text Detection Algorithms. IEEE Transactions on Circuits and Systems for Video Technology (April 2004)
11. Doermann, D., Li, H.: A video Text Detection System Based on Automated Training. In: Proceedings 15th International Conference on Pattern Recognition (2000)
12. Shin, C.S., Kim, K.I., Park, M.H., Kim, H.J.: Support Vector Machine-Based Text Detection in Digital Video. In: Proceedings of the 2000 IEEE Signal Processing Society Workshop Neural Networks for Signal Processing X, vol. 2, pp. 634–641 (2000)
13. Jeong, K.Y., Jung, K., Kim, E.Y., Kim, H.J.: Neural Network-Based Text Location for News Video Indexing. In: Proc. ICIP 1999, Japan (1999)
14. Hua, X.-S., Chen, X.-R., Wenyin, L., Zhang, H.-J.: Automatic location of text in video frames. In: Proc. ACM Multimedia 2001 Workshops: Multimedia Information Retrieval (MIR 2001), Ottawa, ON, Canada, October 5, pp. 24–27 (2001)
15. Qi, W., et al.: Integrating Visual, Audio and Text Analysis for News Video. In: Proc. Int. Conf. Image Processing (ICIP 2000), Vancouver, BC, Canada (2000)
16. Xi, J., Hua, X.-S., Chen, X.-R., Wenyin, L., Zhang, H.-J.: A Video Text Detection and Recognition System. In: Proc. 2001 IEEE Int. Conf. Multimedia and Expo. (ICME 2001), Tokyo, Japan, August 22-25, pp. 1080–1083 (2001)
17. Ye, J., Huang, L.-L., Hao, X.L.: Neural Network Based Text Detection in Videos Using Local Binary Patterns. In: Chinese Conference on Pattern Recognition, CCPR 2009, November 4-6, pp. 1–5 (2009), doi:10.1109/CCPR.2009.5343973
18. Taylor, G.W., Wolf, C.: Reinforcement Learning for Parameter Control of Text Detection in Images from Video Sequences. In: Proceedings of the International Conference on Information and Communication Technologies: From Theory to Applications, April 19-23, pp. 517–518 (2004), doi:10.1109/ICTTA.2004.1307859
19. Li, M., Wang, C.: An Adaptive Text Detection Approach in Images and Video Frames. In: IEEE International Joint Conference on Neural Networks, IJCNN 2008. IEEE World Congress on Computational Intelligence, June 1-8, pp. 72–77 (2008)
20. Li, H., Doermann, D., Kia, O.: Automatic Text Detection and Tracking in Digital Video. IEEE Transactions on Image Processing 9(1), 147–156 (2000)
21. Li, H., Doermann, D.: A Video Text Detection System Based on Automated Training. In: Proceedings of the 15th International Conference on Pattern Recognition, vol. 2, pp. 223–226 (2000)