LOCALIZATION OF OVERLAID TEXT BASED ON NOISE INCONSISTENCIES

Too Kipyego Boaz and Prabhakar C. J.

Department of Computer Science, Kuvempu University, India

ABSTRACT

In this paper, we present a novel technique for localization of caption text in video frames based on noise inconsistencies. Text is artificially added to the video after it has been captured and as such does not form part of the original video graphics. Typically, the amount of noise level is uniform across the entire captured video frame, thus, artificially embedding or overlaying text on the video introduces yet another segment of noise level. Therefore detection of various noise levels in the video frame may signify availability of overlaid text. Hence we exploited this property by detecting regions with various noise levels to localize overlaid text in video frames. Experimental results obtained shows a great improvement in line with overlaid text localization, where we have performed metric measure based on Recall, Precision and f-measure.

KEYWORDS

Overlaid text detection, Text information extraction, Text localization.

1. INTRODUCTION

Given the fast and widespread penetration of multimedia data into all areas of life, nowadays cheaper and high performance digital media is being relied upon as the primary way to present news information, sports and entertainment regularly which captures current events as they occur. Digital videos and images form larger part of archived multimedia data files and they are rich in text information. Text has a well-defined unambiguous meaning that reflects the contents of the video and images. Therefore, it is imperative to build algorithms that provide mechanisms to ensure reliability of multimedia data and enables efficient browsing, querying, retrieval, and indexing of the desired contents available in the on-line digital libraries worldwide. Extraction of these texts comes with a number of challenges and they include, low resolutions, colour bleeding, low contrast, font sizes and orientations [1]. Text Information Extraction (TIE) in videos is a very important area in research, because text contains high-level semantic information, and therefore have many useful applications in the following areas; 1) Document analysis and retrieval, 2) Vehicle's license plate detection and extraction, 3) Keyword based image search, 4) Identification of parts in industrial automation, 5) Address block location etc.

Text that resides in the digital videos and images can be classified based on the following two criteria: 1) Naturally and 2) Overlaid. Text that appears naturally on the captured video without human input is called scene text, whereas text that is artificially overlaid to an already existing video or image is known as Caption or Graphics text. Scene text becomes hard to extract as they form part of the video or image and are characterized by low contrast, reside in complex textured background, and therefore requires advanced techniques to extract its texture features. Caption text is relatively easier to extract when compared to scene text, this is because they are overlaid during video editing, where the most important thing that is considered during video editing is readability by using high contrast colours between the video background colour and the overlaid text colour.

In general, the TIE from videos involves three major steps: 1) text localization, 2) text segmentation, and 3) text recognition. Text localization algorithm is applied to locate the text regions by drawing a rectangular bounding box around it. Text segmentation or extraction is performed to compute the foreground pixels from the localized text region, whereas, text recognition is conducted to convert the segmented text image into plain text.

With the knowledge that caption text is not part of the original captured video, but is overlaid during the editing process, we assume that this procedure causes the duplication of pixels within the edited region and brings in some detectable changes and inconsistencies into video properties such as noise variance, chromatic aberration and statistical changes [2], [3]. In the field of multimedia forensics these image property inconsistencies are analyzed to determine the authenticity of the image. In this area there have been two approaches for determining the authenticity of the image according to [2], a) Source identification and b) forgery detection. In the case of source identification approach it deals with identifying the source digital device, whereas forgery detection focuses on discovering evidence of tampering by assessing the authenticity of the digital media. Further, forgery detection methods can be classified into: 1) active and 2) passive (Blind) approaches. A digital watermarks method (passive approach) [4] has been proposed as a means for fragile, content authentication, tampering detection, localization of changes, and recovery of original content. Whereas the blind approach is regarded as the new direction with the interest in this field has over the last few years rapidly increasing. In contrast to active approaches, blind approaches do not need any explicit priori information about the image. The blind approach always works with the absence of any digital watermark or signature.

Mahdian et al., [3] proposed to employ a blind approach to detect tampered regions where they have considered noise inconsistencies as key component; they assumed that a given authentic image has uniform noise levels, while tampered regions where two or more images from a different source are spliced together, introduce various noise levels into the image. They have also considered the case where some parts of the image may be copied and pasted into a different part of the image, with the intention to hide an object or a region of the image, this process causes duplication of pixels within the pasted region, when locally random noise was added to the image by the authors they noted a mismatch resulting to noise inconsistencies in the image. Kobayashi et al., [5] proposed a method to detect superimposition generated from video not contained in the original sequence. Their method employs identification of noise inconsistencies between the original video and superimposed regions to detect forgeries.

In this study we propose an approach that performs overlaid text localization on the basis of the noise characteristics. Pixels within a non-overlapping window of the edited regions are differentiated from the rest by computing their noise variance and determining their relationship with its neighbours. A merging technique is employed that groups neighbouring blocks with similar noise variance together. The detection of various homogeneous noise levels is an evidential existence and residence of overlaid text in the original digital media. An important feature of the caption text is its linear property as text always appears in a linear form. Text in a line is expected to have similar width, height and spacing between the characters and words [6] and is rich in edge information. Therefore we have incorporated other existing cheaper methods such as edge detection to help filter out the regions that do not contain caption text.

The rest of this paper proceeds as follows; related work on caption text detection and localization is given in section 2, with section 3 giving a detailed description of the proposed approach, section 4 provides the experimental results, while section 5 concludes the paper.

2. RELATED WORK

Caption text contained in videos is a rich source of information, where caption based content retrieval has become a popular focus for researchers dealing with video content retrieval. Localization in video frames has also been a hot area in the TIE, where information media comes into focus, as users want to retrieve and get the information published in the digital archival system. It has been a challenge to browse and retrieve the intended content as videos are stored in compressed graphics media form which cannot be retrieved easily as the current tools cannot read and recognise them efficiently in the said picture format. Prior to localization of text, text detection is the initial step aimed at finding out if the video frames contains text or not, and has been considered in many recent studies [7], [8]. The authors employed a scene change algorithm to select target frames from the video sequence at affixed time interval and a colour histogram is used to segment into various colour planes where text lines are detected and stored.

After text detection algorithm has confirmed the existence of text within the image frame, it is followed by text localization which is the most important part of text Information Extraction and it involves finding the position of text region in the image frame [9], [10], [11], a good localization result is a set of tight bounding box around each text region [1].

Current video text localization approaches can be classified into two categories based on the number of frames involved; 1) Single frame and 2) Multiple frames. Multi-frame integration (MFI) utilizes the temporality of video frame sequences [12][13],[14],[15], where frame contents that appear to be true in all the selected sequence of frames are considered as text feature, because the same text existing in the multiple video frames generally tends to occur at the same location for at least 2-seconds [13]. While these methods can register good results in determining text regions, its accuracy is very low due to high false positive when the method contains other characteristics that resemble text features. The other category detects text regions in individual frames independently [16]. A great deal of work has been done to detect and localize text in video frames and these methods can be categorized again into two classes namely, 1) Texture-based and 2) Region-based text extraction [17]. This is independent of the first classification that was based on the number of frames involved.

The texture based methods [18], [19], [20] extracts textural properties by scanning the video frames at a number of scales and analyze neighbouring pixels for classification based on text properties which include edge densities, gradient magnitudes and intensity variances, in general, texture features are used to train classifiers that will eventually distinguish between pixels that belong to the text and those belonging to non-text. Techniques such as Wavelet transform, Gabor filters, Fourier transform and machine learning techniques belongs to these category. Texture-based feature extraction methods [18] and [19] presumes that a video have a discriminate textural properties, between the text and non-text objects; that is the area presumed to form the image background and may contain scene text. However, texture based methods are faced with some limitations such as computational complexity, difficulty in integrating information from different scales and in ability to sufficiently detect slant text [6].

To separate caption text from scene text, Shivakumara et al., [21], [20] proposed a technique that separates the two types of text based on the nature of the backgrounds the reside on. It is noted that caption text are overlaid on a relatively less complex background as the image editors takes

into consideration text readability, unlike the scene text that are always part of the complex scene. Whereas in [20] they again proposed another method that combines wavelet-laplacian and colour features to detect text. They perform wavelet decomposition separately on each of the three colour channels of the RGB and selected the a set of the first three high frequency sub-bands on each channel the average of their results are used to enhance the text pixels.
The region based methods [22], [23], [24] works through a bottom up approach where they divide the input video frame into smaller regions. Pixels within each region are processed and analyzed for certain properties, neighbouring pixels exhibiting similar properties are then grouped together forming a connected components. Unlike texture based methods, this approach has some advantages, such as ability to detect text of various scales and multi-oriented text. Individual regions containing various connected components merging back again to form large homogeneous regions with favourable and desired properties. Methods that works with connected components, colour and edge features comprises region based techniques.

In [22] the authors proposed an approach that exploits the approximate constant colour exhibited by foreground pixels containing certain properties where they are then grouped together to form regions. Liu et al., [23] have proposed a three step tracking method to locate caption text, perform region binarization and detect the changes in caption text. The approach implements the spatial edge information within an individual frame and the binarization technique is used to classify text and non-text pixels. The result of this initial procedure will be used to detect caption changes across the successive frames.

A Temporal feature method is proposed in [25], the spatial, temporal locations, caption segmentation and post processing to identify stroke direction changes and to segment caption pixels based on consistency and dominancy of caption colour distribution which is later refined by applying post processing techniques. Akhtar et al., [26] proposed an edge based segmentation method that finds vertical gradients and its average gradient magnitude in a pixel neighbourhood for horizontally aligned artificial Urdu text detection from videos, resulting to image binarization which is smoothed using horizontal run length smoothing algorithm to merge text regions and an edge filter to remove noisy non-text regions.

In [24] they have proposed an approach for overlaid text localization using Colour Filter Array (CFA) to detect various regions as they assume that text portion has distinct intensity values with respect to the non-text region or background. This difference in the intensity values was captured in the gradient image map. While this method produces good results their main drawbacks are 1) it requires explicit prior information and 2) It performs poorly on commercially compressed JPEG images, which forms a large part of online data.

Caption text localization methods provided in this section follows all the steps of TIE in order to achieve their objectives, the ultimate goal of these methods are to provide the user with the tools capable to robustly use text image browser for direct access into to the temporal position of an interesting text-image and correct text-lines in a video document. Though these methods have achieved high accuracy rates, but they fall short of being reliable to all forms of videos.

3. PROPOSED APPROACH

In our proposed method we have developed a robust method that provides a reliable solution to localization of overlaid text on the basis of the noise characteristics. The process of video editing introduces a different noise level to already existing noise properties in the original video. Hence our approach is presented as a problem of identifying various regions with homogeneous noise levels through the computation and comparison of the relationship between pixel values in the image in order to detect various noise inconsistencies.

Various noise levels do not in itself provide a proof of the existence of overlaid text, but we have incorporated edge map to truly identify text regions. The diagrammatic flow of our approach is shown figure 1. It is discussed in details through the following steps; 1) Estimation of noise variance, 2) Block merging, 3) Edge estimation and 4) Morphological operation and Mapping.

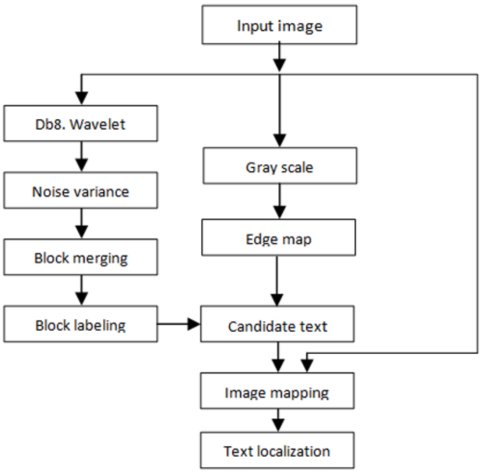

Figure 1: Flow-graph of our proposed approach

3.1. Estimation of Noise Variance

Given a noisy video frame or image that is assumed to contain overlaid text, we can formulate the image signal as shown in the equation below;

$$f_n(x, y) = f(x, y) + n(x, y). \qquad (1)$$

Where $f(x, y)$ is the uncorrupted signal, while $n(x, y)$ is a Gaussian white noise.

We have taken the original image and perform a single level pyramid decomposition using the wavelet transform to obtain the sub-bands, shown in figure 2(b) (LL_1 LH_1 HL_1 *and* HH_1). Wavelets have been widely used in noise estimation as it cuts up image data into different frequency components. The HH_1 sub-band (figure 2(c)) contains the diagonal details of the image and it represents image's highest resolution. For the sake of better visualization HH_1 is highlighted as shown in figure 2(d). Wavelet coefficients at HH_1 not only corresponds to noise, but it can be affected with image structures. If there is a textured object in a scene, it is not possible to obtain the noise component independently this is because the spatial variation is mixed in the signal. Therefore, this sub-band is further processed, with the assumption that, the wavelet coefficients with absolute value smaller than the threshold value which is computed iteratively belongs to noise [27].

The next step is to apply a non-overlapping tiling window T_i of size $N \times N$ on the refined HH_1 sub-band in order to obtain definite blocks that we assume to have relative amount of noise levels. To determine N we perform preliminary work to find the width and height of every connected component through morphological operation, where we estimated the maximum and minimum width and height of various characters and the minimum of their respective averages are considered as the value of N as shown in equation (2) and equation (3) below;

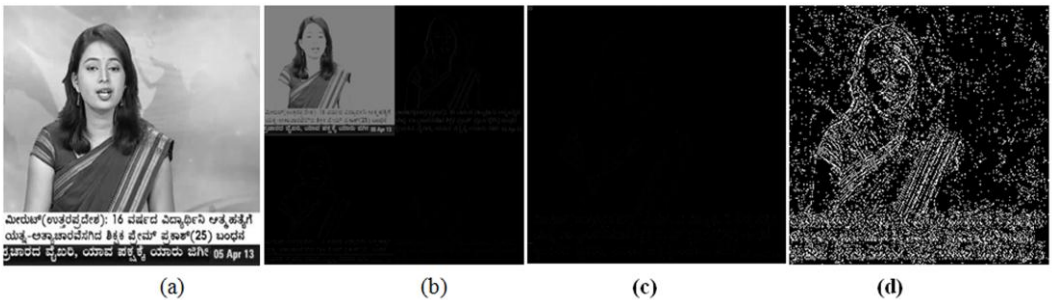

Figure 2: (a) original gray scale image, (b) The wavelet decomposition sub-bands, (c) HH_1 sub-band and (d) HH_1 sub-band (highlighted).

$$(h_{(\max,\min)}, w_{(\max,\min)}) = \max \ \& \ \min(h_i^{char}, w_i^{char}), \tag{2}$$

$$N = \min(avg(h_{(\max,\min)}), (w_{(\max,\min)})). \tag{3}$$

where i is the number of characters in the image.

After obtaining definite blocks we employed a gradient based noise variance estimator on each T_i, $i = 1, 2 \ldots, (\frac{imagesize}{N} * \frac{imagesize}{N})$, in order to obtain gradient amplitudes. Gradient estimator is a widely used technique for estimating the standard deviation $\hat{\sigma}$ of the noise high amplitudes [28]. The standard deviation can be robustly estimated using the following median measurement.

$$\hat{\sigma} = \frac{median(|HH_1|)}{0.6745} \tag{4}$$

3.2 Merging of Blocks

Once the noise standard deviation of each block is estimated, we are going to have several blocks having various noise variances figure 3(a), where every block is assigned a median standard deviation.

We have used this new information as the homogeneity condition to segment the investigated image into several homogenous sub-regions [3]. The procedure is carried out by applying a simple regions-merging technique, where generally all the neighbouring blocks that have similar or almost similar standard deviation are merged together based on the selected similarity threshold value, as computed below.

Let Th be the set threshold value, then $|\hat{\sigma}_k - \hat{\sigma}_m| < Th$ where $k = 1, 2, 3, \ldots, (\frac{imagesize}{N} * \frac{imagesize}{N})$ and $m = k + 1$.

The neighbouring blocks can only be merged if their absolute differences in their standard deviations are within the threshold value, figure 3(b) shows merged blocks, whereas figure 3(c) is a labelled image assigned with unique labels.

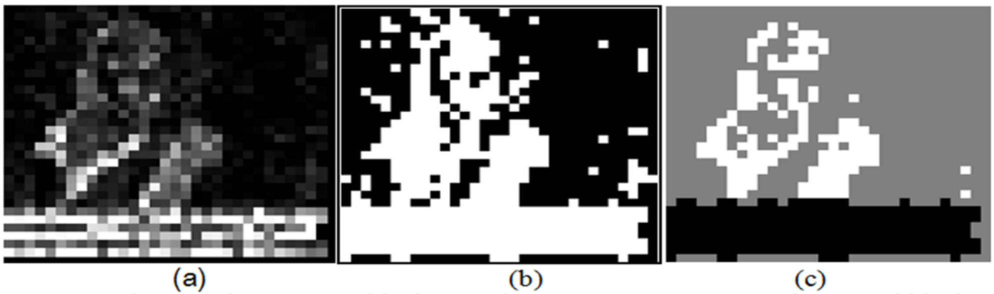

Figure 3: Intermediate results (a) Image blocking with various noise variances, (b) merged blocks and (c) Blocks assigned with labels.

The threshold value is determined in order to group neighbouring blocks together, where a stepwise threshold values Th_i is set, this should be done basically by initially analyzing the noise variances in order to determine the number of classes to allocate. The whole procedure is integrated into simple merging algorithm that is capable of grouping neighbouring blocks of an investigated image into various partitions with homogenous noise levels. The labelling is given based on homogeneity of noise levels.

3.3 Localization of Text Region

Following the labelling of homogeneous noise regions, our next step is to identify which among the labelled regions represents the area containing caption text. We have therefore integrated the result of block labelling with a low cost edge detection method. It is widely believed that edges are rich and reliable feature of text regardless of colour intensity or layout [29]. Edge strength and density are two distinguishing characteristics of text overlaid in images.

We have extracted edges from the gray scale image of the original using canny edge (figure 4 (a)), as it gives strong edges. Based on our earlier computed width and height of the image characters (equation 2 and 3), the resultant edge map is further processed to eliminate both vertical and horizontal edges with extremely longer lengths than $h_{\max}$ and $w_{\max}$ this are edges characterized by long lengths in the form of straight lines that does not belong to text,

To get the text regions we have integrated both the merged labelled blocks (figure 3 (c)) with the refined edge map (figure 4 (b)). An image analysis is then done to eliminate labelled blocks that do not have enough edge information through the determination of the ratio between edge information and the area of the block

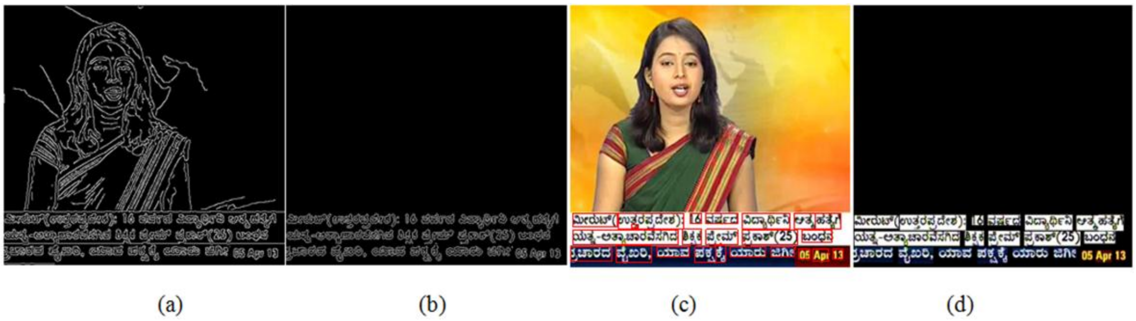

Figure 4: Intermediate results (a) Edge map, (b) refined edge map. Text localization results by (c) Bounding box and (d) extracted text regions.

Normally, text overlaid in an image appears in clusters, therefore to eliminate short and isolated edges, a morphological operation is then used to connected edges together. The size of the structuring element is taken as N (equation 3).

We can only retain those regions that contain text by eliminating small connected components with fewer pixels. This procedure will classify regions into text and non-text labelled blocks. The final stage for text localization is to label all connected components within text rich block by drawing bounding box around them (figure 4(c)). This is followed by mapping and extracting from the original RGB frame only regions within the bounding boxes as shown in figure 4 (d), which represents our text localization result.

4. Experimental Results

To the best of our knowledge, there is no known benchmark dataset available in the literature for caption text detection on the basis of the noise characteristics. We have therefore considered and created two sets of datasets obtained from the following sources; 1) Internet downloads and 2) public databases. These datasets comprise of both videos and images containing graphics text, where most of these datasets are compressed. Since our approach requires the information about noise, the compressed videos and images limit its performance, due to the evolution of sophisticated codec which reduces noise.

The compressed videos and images poses even more challenges to process than un-compressed, since the compression codec can have retouch tools in a bid to cover the traces of editing or else when saved in JPEG, as it cannot produce the exact edit but approximates. The selection of text frames for the created dataset attempts to label a frames according to the noise levels they contain, ideally the selection process should be simple, fast and cost effective in order to reduce the computational time required to process them.

Noise levels don't give text information directly but only acts as a tool to identify various homogeneous regions. The overall experimental implementation of our approach using MATLAB, with a frame size of 512x512 resolutions and block size parameter of N=10, on a PC with a 2.93 GHz core 2duo processor and 256MB memory, has an average run time of 17 seconds. All experimental results were obtained on gray scale video frames or images, using the Daubechies wavelet db8., with the additive Gaussian mean of zero.

4.1 Frame Selection

Noise level is a very important parameter to many images processing applications such as image de-noising, and accurately estimating the noise level improves their performance. The main importance of this section to this paper is to analyze the noise characteristics in each input image independently for the purpose of building a database having the desired noise characteristics. This is done by estimating the noise level in the input video frame based on the method proposed by

Liu et al., [30]. The graph (figure 5) shows estimated noise level for a gray scale image. The input to this process is a single video frame positively identified to contain overlaid text, in which we are seeking to estimate its noise information. The estimation process does require neither a reference frame nor prior information. Because our approach uses noise information it is imperative that the video frames selected should contain some noise. If the frame doesn't contain any noise, but contains overlaid text, therefore some local noise is added to the video frame to test if there will provide the desired characteristics.

Our created dataset must be either in these two conditions,

1) When an video frame or image has noise level above the set threshold, and
2) When the video frame or image does not have noise or have noise level less than the threshold value, then additive noise is added to the video frame or image.

In this section we have build a pre-processing step guided by our approach requirements and it provides a cheaper way to create a dataset consisting of video frames and images with the desired noise characteristics and is accurately selected from different internet sources. We have made this section not part of our approach computational time.

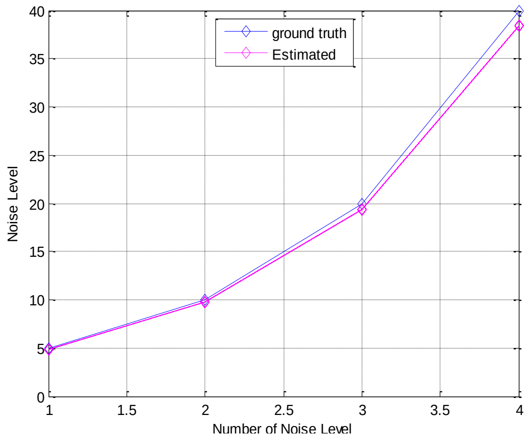

Figure 5: Noise level estimation for the gray scale image shown in figure 2(a)

4.2 Evaluation Metrics

Evaluating text localization has generated a number of test metrics where each is adapted to simulated localization results. In our approach we have adopted a widely used and more accurate metric as it produces a good averaging result on our dataset. The adopted metric test for Recall (R), Precision (P) and f-measure and uses a principle of supervised evaluation [31]. Two synthetic image maps are maintained for each evaluation process, (1) Ground truth image map and (2) Simulated localization result. The ground truth map is generated by manually marking the bounding box which surrounds the entire text block. The most important factor to consider while marking the ground truth is to make sure that some specific properties such as foreground pixels are not left out as they are our main interest. The metric provides a score to measure the performance of our technique by determining the correspondence between these two image maps. The computation of precision (equation 5) and recall (equation 6) based on pixel accuracy is given in the following definitions.

1. A pixel is classified as true positive (TP) if it is ON in both ground truth image map and output of simulated text localized frame image
2. A pixel is classified as false positive (FP) if it is ON only in the output of simulated text localized image.
3. A pixel is classified as false negative (FN) if it is ON only in the ground truth image map.

$$precision = \frac{Number\ of\ TP}{(Number\ of\ FP\ +\ Number\ of\ TP)}, \quad (5)$$

$$\mathrm{Re}\,call(R) = \frac{Number\ of\ TP}{(Number\ of\ FN\ +\ Number\ of\ TP)}. \quad (6)$$

We use pixel accurate ground truth in total of 339 video frames and images containing caption text of various languages including a combination of Kannada (regional language) and English script. The video frames and images contain caption of various text layout and font sizes.

4.2.1. Experiments on own Datasets

We have collected a number of video clips for our dataset from the commercial TV channels. In order to evaluate the performance of our proposed approach to localize caption text, we conducted experiments on individual frames extracted on video clips downloaded from TV9 Kannada channel in various situations, such as TV9 (news), TV9 (hosting), TV9 (Celebrity talk), TV9 (advertisement) and English CNN channel, as these video frames are generally structured with complex background, low contrast and blurred text. Since our proposed approach does not require any prior information about noise, it was important that only randomly and non-sequential video frames are selected with the assumption that they contain overlaid text. All of these video frames have a resolution of 512×512 and a frame rate 29.97 per second. Sample result from our dataset is shown in figure 6, where multi-size fonts do exist within the video frame, this condition posses a great challenge in text localization as each font size will require specific value of structuring element. In order to address this challenge we have perform experiments repetitively, where in every trial we have noted the localized results along with its corresponding structuring element. We will only select the results that portray a fair localization to all font sizes. The evaluation performance of our approach on our dataset is shown in Table 1.

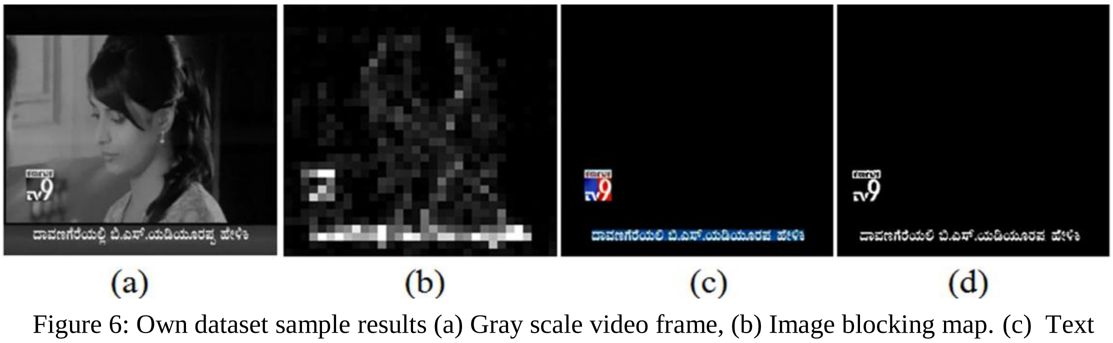

Figure 6: Own dataset sample results (a) Gray scale video frame, (b) Image blocking map. (c) Text localization results and (d) Extracted text

Table 1. Results on Recall, Precision and F-measure

Video source	Number of pixels %		
	True positive	False positive	False Negative

TV9 (news)	0.89	0.12	0.07
TV9 (hosting)	0.99	0.03	0.00
TV9 (Celebrity talk)	0.99	0.09	0.04
TV9 (advertisement)	0.98	0.04	0.03
Average	0.96	0.07	0.04

4.2.2. Experiments on Benchmark Datasets

To test the performance of our technique, we created a dataset comprising of videos and images obtained from various publicly available benchmark databases such as Image Processing Centre (IPC)-Artificial text and ICDAR-13 dataset. These are standard developed dataset containing labelled ground truth for accurate comparison and evaluation. Figure 7 shows the accurate results of localization and extraction based on IPC dataset.

The experimental results of the proposed algorithm show an improved performance even in compressed digital video frames and images and are shown in Table 2.

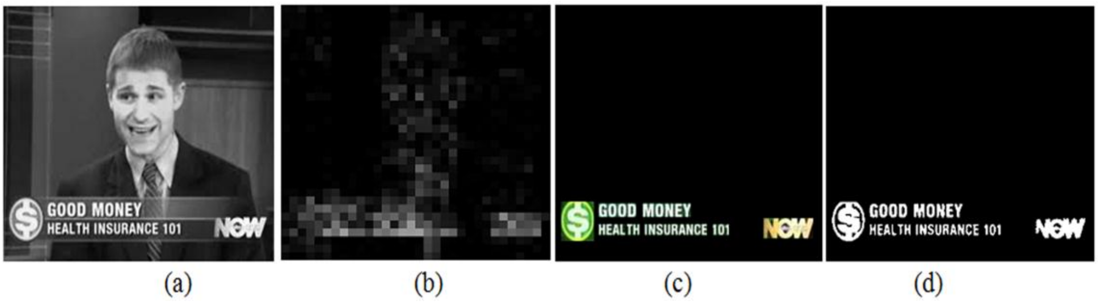

Figure 7: IPC dataset sample results (a) Gray scale video frame, (b) Image blocking map. Text localization results by (c) Extracted text regions and (d) Extracted text.

Table 2. Results on Recall, Precision and F-measure

Video source	**Number of pixels %**		
	True positive	**False positive**	**False Negative**
ICDAR-2013 [32]	0.97	0.03	0.10
IPC dataset [33]	0.98	0..02	0.04
Still images [34]	0.99	0.01	0.07
Average	0.98	0.02	0.07

4.2.3. Comparison with other Techniques

The performance of our proposed technique has been evaluated and a comparison with other recent text localization methods has been done. The recent methods which include edge based method [25] and Feature based method [18] has been implemented and experiments done using both our dataset and publicly available datasets [33], [34]. From the experimental results obtained, our method performs better in both our dataset and the benchmark dataset than the existing methods in areas with complex background and compressed video frames and images.

The visual comparisons can be clearly detected from the output results between the proposed and existing methods as shown in Figure 6, with the comparison results shown in Table 3. The proposed result (figure 6 (a)) is visually better than the existing methods even before we incorporate the edge algorithm. The existing methods [25] and [18] are capable of localizing text area, but fail to provide a tight bounding box around text area hence they enclose more non-text pixels than text pixels.

Table 3. Results on Recall, Precision and F-measure

Method	**Precision**	**Recall**	**F-measure**
Edge based [25]	0.90	0.93	0.92
Feature based [18]	0.87	0.85	0.86
Proposed	0.95	0.94	0.95

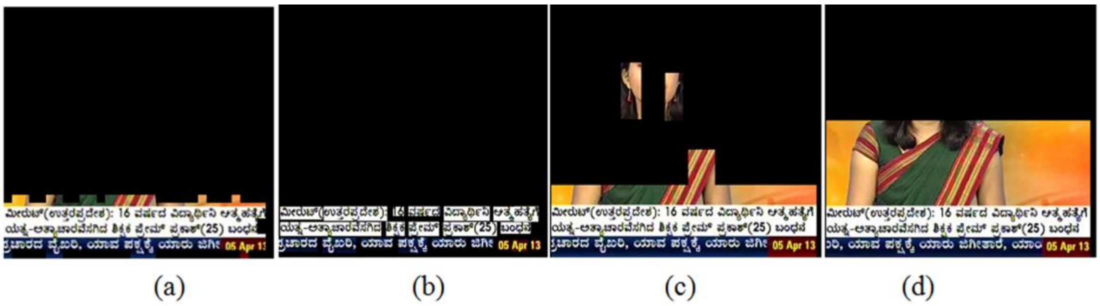

Figure 6: Comparison results. (a) Candidate text region based on block label (proposed), Final text localization (proposed), (c) Region based method [25] and (d) Feature based [18].

5. Conclusion

We have performed caption text localization based on image noise inconsistencies. Typically, captured video frames or images contain uniform noise levels. In TIE, it is assumed that there is a change in noise characteristics when text is overlaid into a video or image during the editing process. This is because within the overlaid region there will be duplication of pixels, and when additive Gaussian noise is added to it, the duplicated regions will not match exactly. The accurate identification and labelling of various regions with homogeneous noise variances provides an insight of the location of overlaid text. Text localization results of our approach show that it is possible in a simple and blind way to divide an investigated image into various segments with homogenous noise level without depending on any prior information, and this is the main strength of our approach. The primary contribution of this work is the use of noise characteristics to discover image inconsistencies in edited video frames and images. But the main drawback of the method is that, some images may also contain various isolated regions with totally different variances caused by structured image. The method can denote these regions as inconsistent with the rest of the image, and therefore it is necessary that there shall be a human interpretation of the output results.

References

[1] Crandall, David, Sameer Antani, and Rangachar Kasturi. "Extraction of special effects caption text events from digital video." International journal on document analysis and recognition 5.2-3 (2003): 138-157.

[2]G R Talmale and R W Jasutkar. Article: Analysis of Different Techniques of Image Forgery Detection. IJCA Proceedings on National Conference on Recent Trends in Computing NCRTC(3):13-18, May 2012. Full text available.

[3] Mahdian, Babak, and Stanislav Saic. "Using noise inconsistencies for blind image forensics." Image and Vision Computing 27.10 (2009): 1497-1503.

[4] J. Fridrich, "Methods for Tamper Detection in Digital Images", Proc. ACM Workshop on Multimedia and Security, Orlando, FL, October 30−31, 1999, pp. 19−23.

[5] Kobayashi, Michihiro, Takahiro Okabe, and Yoichi Sato. "Detecting forgery from static-scene video based on inconsistency in noise level functions." Information Forensics and Security, IEEE Transactions on 5.4 (2010): 883-892.

[6] Epshtein, Boris, Eyal Ofek, and Yonatan Wexler. "Detecting text in natural scenes with stroke width transform." Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. IEEE, 2010.

[7] Kim, Hae-Kwang. "Efficient automatic text location method and content-based indexing and structuring of video database." Journal of Visual Communication and Image Representation 7.4 (1996): 336-344.

[8] M.A. Smith, T. Kanade, Video skimming for quick browsing based on audio and image characterization, Technical Report CMU-CS-95-186, Carnegie Mellon University, July 1995.

[9] X. Chen, A. Yuille, "Detecting and Reading Text in Natural Scenes", Computer Vision and Pattern Recognition (CVPR), pp. 366-373, 2004

[10] Wang, R., Jin, W., Wu, L., A novel video caption detection approach using multi-frame integration, Pattern Recognition, 17th International Conference on ICPR2004, Vol. 1, 2004, pp. 449-452

[11] Jung, K., Kim, K.I., Jain, A.K., Text information extraction in images and video: A survey, pattern Recognition, Vol. 37, No. 5, 2004, pp. 977-997.

[12] X. Hua, P. Yin and H.J. Zhang, “Efficient video text recognition using Multiple Frame Integration”, IEEE Int. Conf. on Image Processing (ICIP), Sept 2002.

[13] Yi, Jian, Yuxin Peng, and Jianguo Xiao. "Using multiple frame integration for the text recognition of video." Document Analysis and Recognition, 2009. ICDAR'09. 10th International Conference on. IEEE, 2009.

[14] Li, Huiping, David Doermann, and Omid Kia. "Automatic text detection and tracking in digital video." Image Processing, IEEE Transactions on 9.1 (2000): 147-156.

[15] Li, Huiping, and David Doermann. "Text enhancement in digital video using multiple frame integration." Proceedings of the seventh ACM international conference on Multimedia (Part 1). ACM, 1999.

[16] J. Hu, J. Xi, L. Wu, “Automatic Detection and Verification of Text Regions in News Video Frames”, Int. Journal of Pattern Recognition and Artificial Intelligence, Vol.16, No.2, pp.257-271, 2002.

[17] Sharma, Nabin, Umapada Pal, and Michael Blumenstein. "Recent advances in video based document processing: a review." Document Analysis Systems (DAS), 2012 10th IAPR International Workshop on. IEEE, 2012.

[18] Sun, H., Zhao, N., Xu, X., Extraction of text under complex background using wavelet transform and support vector machine, IEEE International Conference on Mechatronics and Automation(ICMNA 2006), Vol. 2006, No. 4026310, 2006, pp. 1493-1497.

[19] H-K Kim, "Efficient automatic text location method and content-based indexing and structuring of video database". J Vis Commun Image Represent 7(4):336–344 (1996).

[20] Shivakumara, Palaiahnakote, Trung Quy Phan, and Chew Lim Tan. "New wavelet and color features for text detection in video." Pattern Recognition (ICPR), 2010 20th International Conference on. IEEE, 2010.

[21] P. Shivakumara, N. Kumar, D. Guru, and C. Tan. Separation of graphics (superimposed) and scene text in video frames. In Document Analysis Systems (DAS), 2014 11th IAPR International Workshop on, pages 344-348, April 2014.

[22] A. Jain, B. Yu, "Automatic Text Location in Images and Video Frames", Pattern Recognition 31(12): 2055-2076 (1998).

[23] Liu, Xiaoqian, and Weiqiang Wang. "Extracting captions from videos using temporal feature." In Proceedings of the international conference on Multimedia, pp. 843-846. ACM, 2010.

[24] Amit Hooda, Madhur Kathuria & Vinod Pankajakshan "Application of Forgery Localization in Overlay Text Detection" ICVGIP '14, December 14 - 18 2014, Bangalore, India (to appear).

[25] Jamil, A., Siddiqi, I., Arif, F., & Raza, A. (2011, September). Edge-based Features for Localization of Artificial Urdu Text in Video Images. In Document Analysis and Recognition (ICDAR), 2011 International Conference on (pp. 1120-1124). IEEE.

[26] Wang, R., Jin, W., Wu, L., A novel video caption detection approach using multi-frame integration, Pattern Recognition, 17th International Conference on ICPR2004, Vol. 1, 2004, pp. 449-452.

[27] D. Pastor, "A theoretical result for processing signals that have unknown distributions and priors in white Gaussian noise," Comput. Stat. Data Anal., vol. 52, no. 6, pp. 3167–3186, 2008.

[28] D. Donoho, I. Johnstone, "Ideal spatial adaption by wavelet" shrinkage,Biometrika 8 (1994) 425–455.

[29] 30_Samarabandu, Jagath, and Xiaoqing Liu. "An edge-based text region extraction algorithm for indoor mobile robot navigation." International Journal of Signal Processing 3.4 (2006): 273-280.

[30] Liu, Xinhao, Masayuki Tanaka, and Masatoshi Okutomi. "Single-Image Noise Level Estimation for Blind Denoising." (2013): 1-1.

[31] Hemery, Baptiste, et al. "Evaluation protocol for localization metrics." Image and Signal Processing. Springer Berlin Heidelberg, 2008. 273-280.

[32] ICDAR-2013dataset: http://dag.cvc.uab.es/icdar2013competition/?com=news&action=data&id=15.

[33] IPC-Artificial text dataset: https://sites.google.com/site/artificialtextdataset/.

[34] Still images: http://www-i6.informatik.rwth-aachen.de/imageclef/resources/iaprtc12.tgz.