INPAINTING SCHEME FOR TEXT IN VIDEO : A SURVEY

Mukesh Sakle[1], Rahul P. Chauhan[2]

[1]*Information technology Department, Parul Institute of Engineering and Technology, Limda, Vadodara, India*
[2]*Computer Engineering department, Parul Institute of Engineering and Technology, Limda, Vadodara, India*

Abstract

Text data present in video sequences provide useful information of paramount requirement. Although text present in video provide useful information not all are of them are necessary because it may hide the important portion of the video. So there must way to erase this type of unwanted text. This can be done in two phase first text components are detected from each frame of the video. Detected text component are then removed from the video sequences. And restore the occluded part of the video using inpaint method. This text detection and removal scheme is in two phases. Each phase is broad topic of image processing. Video text detection and Inpainting are two most important phase in this scheme. Text detection phase consist of text localization, text segmentation and recognition phase. Inpainting method is used for restoring occluded part produce due to removal of text.

Keywords—*Optical Character Recognition(OCR), Stroke width transform, Text Detection,Connected Component*

----------***----------

1. INTRODUCTION

Inpainting is like reconstructing occluded parts of images and videos. It is a synonym for image interpolation. Text information removal from image and video frames consists of two phases. First is text information extraction. Second is reconstructing the parts occluded by removing text objects in image or video frames. This is done using inpainting techniques. Text information extraction phase consist of two phases. One is text information detection in image or video frames and second is text information recognition. Text information extraction has several other phases included like text detection, text localization, enhancement and recognition. Video completion consists of filling in the occluded parts in video sequences caused by unnecessary objects or scratches on frames. In video there are different types of text included or appeared. It may be scene text, caption text or document text. Text data present in image or video frames contain useful information. Text data appears in image or video frames are used for annotation, indexing and structuring of an image. Reconstruction of the occluded parts in the image is done by inpainting techniques. Digital inpainting models, techniques have applications in image interpolation, photo restoration, super resolution and zooming, perceptual image compression and coding, and the error hiding in image transmission. Here we use two broad category of image processing that is text detection and inpainting to remove unnecessary text information form video frame sequences.

2. RELATED WORK

Text detection and removal is a scheme which attempts to develop a computer system with the ability to automatically read from the video the text content visually embedded in complex background and after removing text string, restore that part using inpainting. Optical character recognition (OCR) read characters that are optically processed. The first OCR was employed to convert typewritten reports into computer readable punched cards. OCR systems were able to process large amount of printed and handwritten characters into computer documents [1]. Here different image processing algorithms primary to image analysis in general and to image text detection in particular is introduced as well as mathematical background is also described in brief. First visual feature detection techniques is described than text detection using machine learning principal is described and than two stochastic modeling technique gaussian mixture models and Markov random field are introduced.

2.1 Visual Feature Extraction

Edge detection: Edges are the visual features used to detect text content in images. Edges have the sharp discontinuity in a gray level. Edge detection is a difficult task if amount of noise is more. Edges can be defined by its magnitude and direction. Spatial derivation and thresholding is used to find edges.

The first order derivatives of a given image function defined as vector it is also called gradient

$$\nabla I = \left(\frac{\partial I}{\partial x}, \frac{\partial I}{\partial y} \right)$$

For digital image gradient can be calculated using difference operation. Edges can be calculated by finding local maxima of the gradient magnitude. Image contain great amount of noise there for selection of local maxima is done by thresholding.

Second order derivatives contain noise more than first order derivative. Second order derivative contain more noise than first order derivative. Advantage of second order derivative is it can easily find close contour than to track. Canny [2] use both first and second order derivative to find edges.

Texture feature extraction: Texture is a visual feature that can be helpful for the detection of text string in image or video frames. The term texture represents some characteristics of a region of an image. Although images contain various objects that may have its own texture characteristics, researchers have found some texture characteristics of text string. Texture features can be characterized by histogram. We have three types of texture detection algorithm to find text strings in image. They are gaussian space features, discrete cosine transform and wavelet features. Wu [3] have proposed algorithm to detect text using gaussian functions. Wu have proposed that text feature can be extracted by calculating second order derivative of a given image processed by gaussian function. Discrete cosine transform coefficients are another texture feature for text detection. It is based on discrete cosine transform coefficients. This transformation has capability to convert spatial amplitude representation of image into spatial frequency representation. Advantage of DCT is it can distinguish some of content region by concentrating signal frequency and neglect other regions. The advantage of using DCT in text detection is it can effectively represent visual information of an image. The algorithm is fast enough to present visual information. The main issue in the text detection phase is localization of the text in complex background because unconstrained property of text. Representations of image in the form that can OCR system perform well and for that texture representation is helpful for better binarization. Wavelet feature: Li have presented algorithm for texture space representation. This algorithm takes benefits from the Haar wavelet decomposition. It is useful because it can easily find out line segments which may be text regions. Particular image input are segmented into four images that are low frequency image (L), vertical high frequency image (V), horizontal high frequency image (H), and diagonal high frequency image (D). The original image is decompose until there is only one pixel contain in sub image. Features are extracted. Mean value, second order central moment and third order central moment is calculated as a feature vector. The moment values are calculated for each sub image and in every frequency domain. Some features are than selected as text regions and other non text region. This transformation is also a fast representation algorithm.

2.1 Machine Learning Approaches

This technique is useful in text verification task. Machine learning approach is useful when we cannot generate mathematical model to distinguish text string from its background. Machine learning approach is useful when mathematical model fails. Here we have to only detect text regions there for this approach is useful rather than for general cases because here we have to only find out whether given region is text or non text. Two mostly used machine learning approaches are multilayer perceptrons and support vector machine.

Multilayer perceptrons: Perceptron is a simplest form of a network [4]. A perceptron has number of layers including input layer, output layer and number of hidden layers. Output node computes sum of the weighted sum of the input nodes. MLP basically has multilayer extension of a perceptron network which consists of one or more number of hidden layers of neurons. Here main work of MLP is to classification based on training. Classification done by computing weighted sum by neurons of the hidden layer. Each layer computes weighted sum of the previous layer. One output associated with only one neuron. We can write output layer as a vector. We want optimal weight on output layers by training. Back Propagation is one algorithm that we can use for training process.

Support Vector Machine: Support vector machine is used as a classification tool. It is also based on training. Statistical learning is the base of support vector machine. It is applied in the various classification tasks in the various fields. MLP is empirical task where as SVM is structured algorithm. SVM mostly used for generalize objects into its classical form. There are different cases in which support vector machine works these are soft margin also called linear non separable case. It is look like linear separating surface but such separating hyper plane does not exists. Second is Non linear decision surface. This is the generalized method for non linear case.

2.2 Statistical Methods

Gaussian mixture models: Gaussian use multiple gaussian to approximate distribution. We can model distribution of grayscale image using three or more gaussian. Expectation maximization algorithm is used for selecting parameters. EM algorithm makes clear the estimation process by maximum likelihood estimation with incomplete data. EM algorithm consists of different steps. E-step is used to estimate missing region from the complete dataset. It first checks conditional expectation. M-step is used to maximize complete data log.
Markov Random Field: Basically it is a branch of probability theory. Markov random field is used to model representative spatial feature also called spatial context.MRF is equivalent to Gibbs distribution.

2.3 Video in Painting Related Work

After removing text string from the image we have to restore the occluded part of video frames. But here main focus is to maintain the spatial and temporal consistency of a video. There are various method for video inpainting. But here we have to select efficient inpainting method. Because here we have to complete the procedure in less amount of time. Missing parts in the video sequence are caused by undesired object removal or damages. Restoration problem of such missing regions is modeled by video inpainting.

3. TEXT DETECTION

Text detection in complex background can be group together into bottom up, heuristic top-down methods and machine learning based top down methods.

Bottom up methods: This method scans the image into number of scales. Basically it treats an image into number of segments and distinguish particular text pattern from the

background. Geometric property like size of the region, height and the width ratio, variance of stroke, width of stoke, illumination of region, gradient value are used to classify different object and we select "character" region. There regions are than grouped together to form text component. Text components are grouped together according to their alignments. These methods are very sensitive for the character size, font size, alignment, gradient value.

Heuristic top down methods: In this method Heuristic filters are used to detect text block in image. This text block then segmented from its background. The primary operation is detect text in image than secondary part is separate it from its background. This algorithms are useful for detect text in complex image but still difficulties are there in text segmentation. There are various properties by which text are detected. These are vertical edge, texture and edge orientation. One system detects text string using horizontal variance of text component. Spatial properties are used in connected component analysis. One system uses vertical edge to find text component in image. These vertical edges are than grouped together to find text regions in the image by smoothing process. Although these methods to detect text in image are fast but generate false alarm that need to consider.

Machine learning based top down method: When background texture of image is complex than heuristic top down method are of no use. Because it is not a robust method to find text in all type of image. Some system exploited machine learning tools to detect text in image or video sequences [4]. Wavelet features and derivative features are extracted to locate text regions in image or video frames. Fixes size of block of pixel is classified for text and non text region using artificial neural network. The main drawback of this system is it is not efficient in terms of computation cost. This may produce false alarm.

Leon [5] uses geometric property like color, size, alignment, motion of text to detect caption text in image. To detect various objects in image we have to use descriptor which distinguish particular object from input. Here hierarchical based image model is used as a descriptor to classify text object. Property of caption text is exploited to detect it from the complex image. Caption text can be found from image by using contrast value because pixels of caption text have similar contrast value. Area is analyzed using discrete wavelet transform. Wavelet transform have efficient texture representation. Feature vectors are generated for the classification of text. Different geometric property like occupancy, aspect ratio, height of text string, area occupied are calculated and fed to the classification algorithm. Classification is done based on tree analysis.

Ezaki [6] use character features to detect text in image. Consistency of a gradient of pixel is used to recognize text component region from the background region. Character feature like edge strength, edge density and horizontal distribution of intensity level of text pixels is used to localize text. First use of edge density is used by applying local thresholding after that edge strength feature is used for better binarization. Local thresholding technique is used to keep low contrast text. Local profiles are used as classifier for identification of text object. Contrast levels of connected components are analyzed.

Here main focus is to locate both low contrast and high contrast text component in Cai [7]. There for Thresholding is employed to locate both low and high contrast text. Edge detection is first applied to image. Thresholding is used for refinement of image. Text enhancement is done for better binarization of image. Two operators are used edge strength smoothing and edge clustering power operator. Classification is done based on project profile of horizontal and vertical distribution of text.

Liu [8] method is based on the reliable feature of edges like edge strength, density and the orientation of text component. These are the properties of text embedded in image. Intensity value of a text pixel is same. Global thresholding is used to identify text component. Here magnitude of second derivative of intensity is computed for the measurement of edge strength. Edge density is calculated using average edge strength. Text component has considerably high edge strength than that of background. Feature map are generated exploiting this type of property. These feature maps are fed to the classification engine for the recognition of text. OCR system does the recognition task.

Molesh [9] uses a simple and effective operator called stroke width transform. Performance of stoke width transform is proportional to textural representation of image. There for first edge detection is applied on image and then stroke width transformation is applied. To detect edges of an image one another transformation is applied. That is bandlet transform. Bandlet based effectively represent textural structure of an image. Image is segmented in sub images called square image. This image is further segmented in sub image until it is found geometrical regularity in that image. Now for each square bandlet coefficients are calculated. Three level of thresholding is employed to extract text regions form the background. Feature vectors are then generated respect to text property. These feature vectors are fed into the classification algorithm. K-means algorithm is use as a classification tool. K-means algorithm detects text strings. Advantage of bandlet based over wavelet bases is wavelet transform cannot identify object in seismic like structure. There for here use of bandlet base is employed. This algorithm is fast and detects accurate text string located in image.

4. VIDEO TEXT REMOVAL

After removing text from the video sequences we must restore the occluded part of that region so survey of different inpainting algorithm is provided in this topic. There is various methods by which we can do inpainting to restore image.

Bui [10] take benefits from inter and intra scale dependency. Wavelet transformation has ability to represent textural feature effectively. This will be use to approximate boundary

data to be fill in the missing region. This method assumes the global structure of image and use property like shape and structure to interpolate data into missing regions. Wavelet transformation is suitable for image containing text in complex background. The algorithm decomposes image into four frequency composition. For each segment wavelet coefficient is found. Inpainting is done considering scaling function.

The method in [11] uses ophotelines to inpaint missing parts. Propagation of values are done with the use of values reside on the boundary of occluded regions into the regions itself. Its direction is in the minimal change of the values. This algorithm is best for the images that have fewer regions to be inpaint. This algorithm is fast.

5. DISCUSSION

From the survey it is observed that some of the text detection techniques follow general object identification approach. Some of the techniques perform well in their defined domain. Here Comparison of existing text detection and video inpaint method are given.

Table 1: Comparison of Text Detection Methods

Technique	Advantage	Disadvantage
Region Based	Fast for locating caption text	Difficult to locate scene text
Character Feature Based	Thresholding is simple	Segmentation is slow
Texture Feature based	Keep low contrast text	Only use local thresholding
Edge Based	Execution time is slow	Can only deal with printed characters against clean background
Connected Component based	Accurate text localization	Contain multiple phases

Table 2: Comparison of Inpainting Method

Method	Advantage	Disadvantage
Wavelet Based	Utilizes inter and intra scale dependency to maintain image structure and texture quality	Masking for regions is defined manually.
Partial differential equation based	Produce good results if missed regions are small one.	When the missed regions are large this algorithm will take so long time.
Texture synthesis based	Useful for small regions	Cannot handle natural images
Exemplar based	Provide good result if structure is simple	Use of greedy algorithm

6. DISCUSSION AND CONCLUSION

From the survey it is observed that to remove text from the video sequences, first employ technique of text detection from the image first and then after removing text, to restore the occluded part using video inpainting technique. Performance of the whole scheme is dependent on accurate text detection method as well as robust video inpainting technique. Further existing text detection technique is highly dependent on edge detection. This scheme may not identify that the identified text component is either caption text or scene text to remove. It may remove all the text components appear in the video sequences. Vast amount of work is needed in this scheme.

REFERENCES

[1] S. Mori, C. Y. Suen, and K. Yamamoto. Historical review of ocr research and developement. Proceedings of the IEEE, pages 1029–1058, 1992.

[2] J. F. Canny. A computational approach to edge detection. IEEE Trans. on Pattern Analysis and Machine Intelligence, 8(1):679–698, 1986.

[3] V. Wu, R. Manmatha, and E. M. Riseman. Textfinder: An automatic system to detect and recognize text in images. IEEE Trans. on Pattern Analysis and Machine Intelligence, 11(20):1224–1229, 1999.

[4] H. Li and D. Doermann. Text enhancement in digital video using multiple frame integration. In Proc. ACM Multimedia, volume 1, pages 385–395, Orlando, Florida, USA, 1999.

[5] Leon, M., Vilaplana, V., Gasull, A. and Marques, F., "Caption text extraction for indexing purposes using a hierarchical region-based image model," IEEE ICIP 2009, El Cairo, Egypt, 2009.

[6] N. Ezaki, M. Bulacu, L. Schomaker, "Text Detection from Natural Scene Images: Towards a System for Visually Impaired Persons", Int. Conf. on Pattern Recognition (ICPR 2004), vol. II, pp. 683-686.

[7] M. Cai, J. Song, and M. Lyu, "A new approach for video text detection," in Proc. IEEE ICIP, Feb. 2002, pp. 117–120.

[8] Xiaoqing Liu and Jagath Samarabandu, " Multiscale Edge-Based Text Extraction From Complex Images" , Multimedia and Expo, 2006 IEEE International Conference,2006.

[9] A. Mosleh, N. Bouguila, and A. Ben Hamza, "Image text detection using a bandlet-based edge detector and stroke width transform," in Proc. Brit. Mach. Vis. Conf., Sep. 2012, pp. 63.1–63.12.

[10] Dong wookcho and Tien D. Bui "Image In painting Using Wavelet-Based Inter and Intra-Scale Depedency" IEEE Transactions on Image Processing, 2008.

[11] Bhimaraju Swati, Naveen Malviya, Shrikant Lade "Analysis of Exemplar Base In painting for Adaptive Patch Propagation using Wavelet Transform". IJETAE Volume 3,May 2013.

[12] Marcelo Bertalmio, Luminita Vese, Guillermo Sapiro (2003), "Simultaneous Structure and Texture Image In painting", IEEE transactions on image processing, vol.12.