I.J. Image, Graphics and Signal Processing, **2014, 12, 53-64**
Published Online November 2014 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijigsp.2014.12.08

Survey of Region-Based Text Extraction Techniques for Efficient Indexing of Image/Video Retrieval

Samabia Tehsin, Asif Masood and Sumaira Kausar
National University of Science and Technology (NUST), Islamabad, 46000, Pakistan
Email: samsatti@yahoo.com, amasood@mcs.edu.pk, sum_satti@yahoo.com

Abstract—With the dramatic increase in multimedia data, escalating trend of internet, and amplifying use of image/video capturing devices; content based indexing and text extraction is gaining more and more importance in research community. In the last decade, many techniques for text extraction are reported in the literature. Methodologies of text extraction from images/videos is generally comprises of text detection and localization, text tracking, text segmentation and optical character recognition (OCR). This paper intends to highlight the contributions and limitations of text detection, localization and tracking phases. The problem is exigent due to variations in the font styles, size and color, text orientations, animations and backgrounds. The paper can serve as the beacon-house for the novice researchers of the text extraction community.

Index Terms—Text extraction, Document analysis, Survey, Text localization, Text tracking.

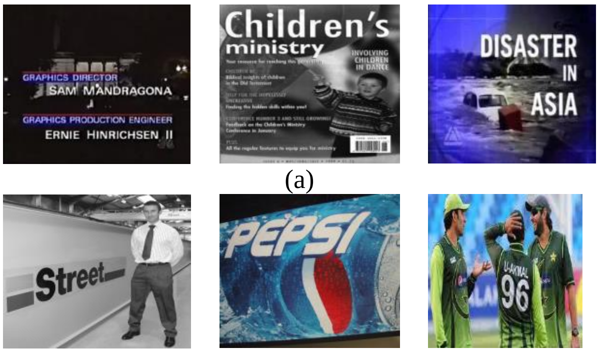
I. Introduction

In recent years there is a rapid increase in multimedia libraries. The amount of digital multimedia data is growing exponentially with time. Thousands of television stations are broadcasting every day. With the vast spread of affordable digital cameras and inexpensive memory devices, multimedia data is increasing every second. Ranging from cameras embedded in mobile phones to professional ones, Surveillance cameras to broadcast videos, every day images to satellite images, all these increasing multimedia data. According to Flickr statistics; just in 2013, 43 million images per month are uploaded that is 1.42 million per day in average [1]. And according to youtube official announcement, 72 hours of videos are uploaded to the site every minute and watched over 3 billion hours a month [2].

With this dramatic increase in multimedia data, escalating trend of internet, and amplifying use of image/video capturing devices; content based indexing and text extraction is gaining more and more importance in research community.

In literature text embedded in images and videos is classified in two groups, caption text and scene text. Caption text is laid over the image/video during editing stage e.g. score of match and name of the speaker. It is also known as artificial text or superimposed text. Caption text usually highlights or recapitulates the multimedia's contents. This formulates caption text principally positive for construction of keyword index. Fig. 1(a) presents some examples of caption text.

Scene text is actual part of the scene e.g. brands of the products, street signs, name plates and text appearing on t-shirts etc. Scene text physically present in the scope of camera view during image/video capture. Fig. 1(b) presents examples of scene text in images and video frames.



(a)

(b)

Fig 1. Images extracted from different domains (a) Caption text (b) Scene text

Section 2 explains the architecture of text extraction process, section 3 highlights the applications of text extraction process, Section 4 explains the problems faced by text extraction researchers, state of the art techniques and their limitations for text localization are presented in section 5 and 6 respectively. Section 7 gives the concluding remarks.

II. Architecture Of Text Extraction Process

Text extraction and recognition process comprises of five steps namely text detection, text localization, text tracking, segmentation or binarization, and character recognition. Architecture of text extraction process can be visualized in Fig. 2.

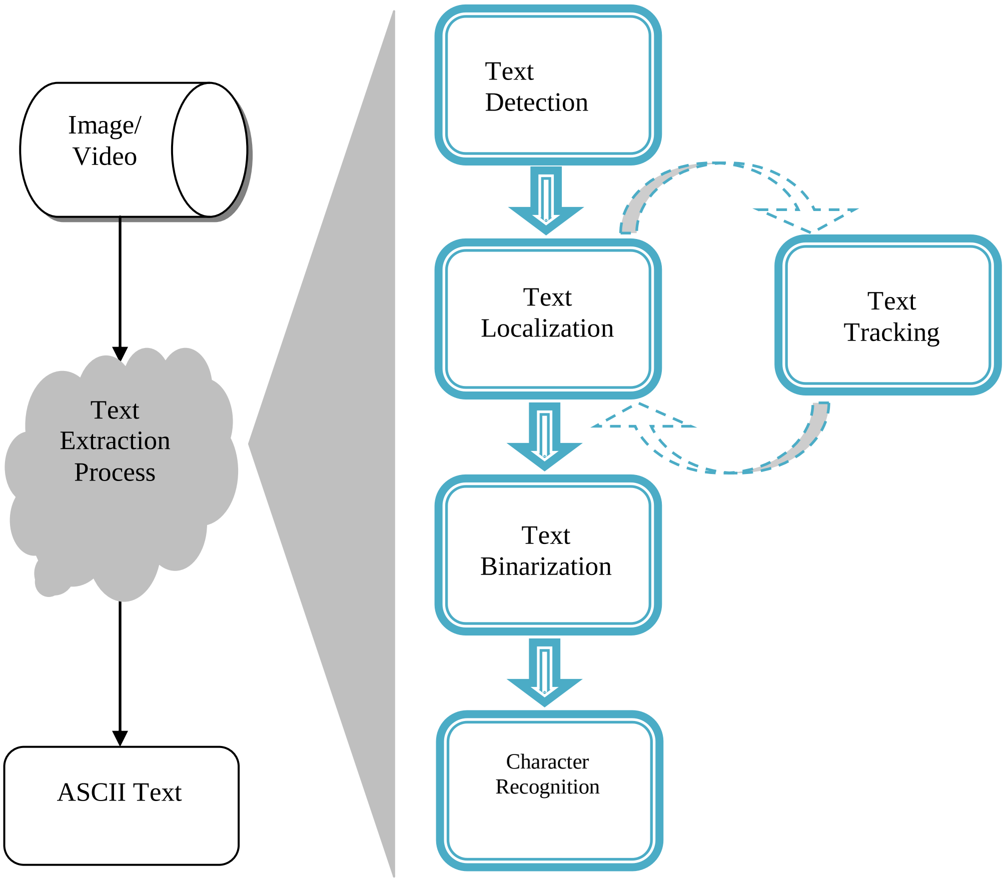


Fig 2. Architecture of text extraction process

Text Detection: This phase takes image or video frame as input and decides it contains text or not. It also identifies the text regions in image.

Text Localization: Text localization merges the text regions to formulate the text objects and define the tight bounds around the text objects.

Text Tracking: This phase is applied to video data only. For the readability purpose, text embedded in the video appears in more than thirty consecutive frames. Text tracking phase exploits this temporal occurrences of the same text object in multiple consecutive frames. It can be used to rectify the results of text detection and localization stage. It is also used to speed up the text extraction process by not applying the binarization and recognition step to every detected object.

Text Binarization: This step is used to segment the text object from the background in the bounded text objects. The output of text binarization is the binary image, where text pixels and background pixels appear in two different binary levels.

Character Recognition: The last module of text extraction process is the character recognition. This module converts the binary text object into the ASCII text.

Text detection, localization and tracking modules are closely related to each other and constitute the most challenging and difficult part of extraction process. Fig. 3 presents the output of different modules of the text extraction process.

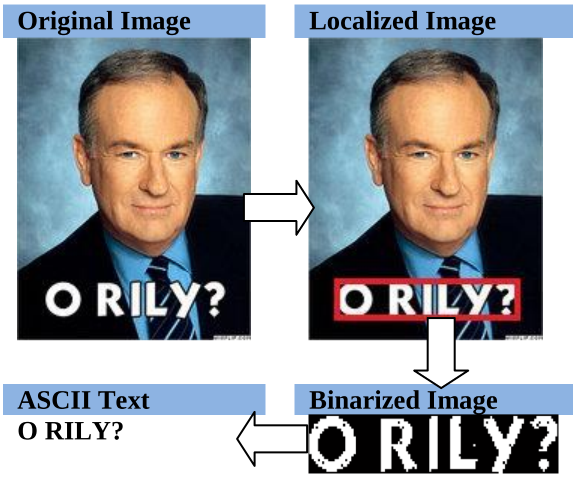


Fig 3. Modular results of text extraction process

III. Applications Of Text Extraction

Text extraction from images has ample of applications. With the rapid increase of multimedia data, need of understanding its content is also amplifying. Some of the applications of the text extraction are mentioned below.

A. Video and Image Retrieval

Content based image and video retrieval is the focus of many researchers for the last many years. Text appearing in the images gives the essence of the actual content of the image and displays the human perception about the content. This makes it a vital tool for indexing and retrieval of multimedia contents [3], [4]. This tool can give much better results than the other shape, texture or color based retrieval techniques [5]. Embedded text in the videos and images communicate human discernment about the content, hence it is most suitable for indexing and retrieval of multimedia data.

B. *Multimedia Summarization*

With the vast increase in the multimedia data, huge amount of information is available. Because of this overwhelming information, problem of overloaded information arise. Text summarization can provide the solution for the problem. Superimposed text in video sequences offer helpful information concerning their contents. Text data appear in video hold valuable knowledge for automatic annotation and generation of content summary. A variety of methods have been presented to deal with this issue. Sports video summarization [6] and News digest [7] are the well known applications of summarization of visual information.

C. *Indexing and Retrieval of Web Pages*

Text Extraction method from web images can truly improve the indexing and retrieval of web pages. Main indexing terms are embedded in the title image or banners. Instead of text, most of the sites use image to present the title of the web page. So to precisely index and retrieve web pages, text within images must be understood. This would result into enhanced indexing and more proficient and accurate searching [8].

Text extraction from web images can also help in filtering of images with offensive language. It is also helpful in conversion of web page to voice. It can also be utilized for web page summary generation [9].

Above listed applications are not the only examples of text extraction methods. There are plenty of other applications such as voice coding for blinds, intelligent transport system, Image tagging, robot vision and scene analysis etc.

IV. CHALLENGES

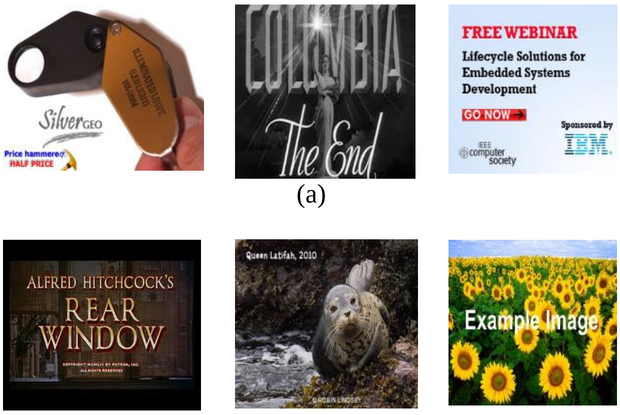
There are many challenges and difficulties for designer and developer of text extraction process. A lot of work has been done in the field of text extraction from multimedia data. But most of the work is application specific and there is still need of work in designing domain independent systems. This is because there are so many challenges when extracting text with variation in fonts, size, color, alignment, orientation, illumination and background. Problem of text extraction get very difficult because of these deviations. Some of the problems of text extraction process are listed below:

Resolution: Size of image and video frames vary from few hundreds to tens of MBs. This puts serious concern for text extraction process to deal with such variation.

Font, Size and Color: The size, font and color of the text can vary in different ranges. This aspect limits application of many text extraction methods. Most of the existing system can only deal with limited range of font styles and sizes.

Complex backgrounds: The complexity of the text background may vary from simple to much complex ones. The background can be comprised of varying colors and textures. In videos, background of the same text object may vary drastically in different frames.

Fig. 4 shows examples of font, size and color variations. It also presents images with complex backgrounds.



(a)

(b)

Fig 4. Challenges in text extraction (a) Variation in font, sizes and colors (b) Text with complex backgrounds.

Computational efficiency: In order to efficiently index and retrieve image and video data from huge multimedia reservoir, retrieval method should be very proficient.

Dynamic text: Text appears in the videos can move in arbitrary directions. It can also change the size of the text in case of zoom in/out.

Noise, Blur and compression: Low resolution images mostly suffer from blur and loose the sharp transitions at the text boundaries(See Fig. 5). GIF that is limited to 8-color palette introduces considerable quantization artifacts and dithered color. Compression artifacts also degrade the quality of edges. For example JPEG compression can produce significant distortion to characters and their boundaries.

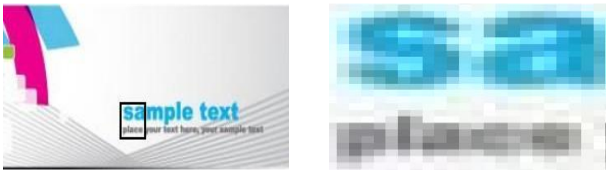


Fig 5. Effect of blurriness (a) Low-resolution image (b) Enlarged highlighted portion with antialiased edges

V. TEXT DETECTION AND LOCALIZATION TECHNIQUES

A variety of techniques for text extraction are appeared in recent past [10]-[15]. Comprehensive surveys can be traced explicitly in [16]-[18]. These techniques can be categorized into two types mainly with reference to the utilized text features i.e. region based and texture based [19].

Texture based methods pertain to textural properties of the text, distinguishing it from the background. The techniques mostly use Gabor filters, Wavelet, FFT, spatial variance, etc. These methods further use machine learning techniques such as SVM, MLP and adaBoost [20]-[23]. These techniques work in the top down fashion by first extracting the texture features and then finding the text regions.

Region based approach exploits different region properties to extract text objects. This approach makes use of the fact that there is sufficient difference between the text color and its immediate background. Color features, edge features, and connected component methods are often used in this approach [24]-[26]. These techniques typically work in the bottom up fashion by first segmenting the small regions and then grouping the potential text regions.

Texture based techniques usually give better results in complex backgrounds than region based techniques but have computationally very heavy hence not suitable for retrieval systems for hefty databases. Therefore, there is a need to improve the detection results of region-based techniques to be used for retrieval and indexing of large multimedia data. Rest of this paper will focus on the region-based text localization techniques.

Region based techniques typically work in the bottom up fashion by initially segmenting the small regions and lately grouping the potential text regions. Region based methods are generally composed of three modules. (1) Segmenting the image into small regions which aims at segregating the character regions from its background, (2) Merging and grouping of small regions to form words and sentences (3) Differentiating between text and non text objects.

Different techniques and strategies have been proposed in the literature for each stage. Each stage is highlighted separately with the beam of existing techniques.

5.1 Segmentation

Image segmentation is very helpful in text extraction applications. It identifies the regions of interest in an image. Artificial text occurrence is commonly characterized in the research community as regions of high contrast and high frequencies. Several approaches are suggested in the literature to segment the image into small regions e.g. edge, gradient, corner, color and intensity based methods. Majority of state of the art methods use one of the following techniques or used the hybrid of these techniques.

5.1.1 Edge and Gradient Based Segmentation

It is pragmatic that text object naturally has elevated edge densities and large edge gradient disparities. It is due to the fact that text consists of group of characters and has large contrasts with the background and edge-based methodologies are employed using these characteristics.

Smith and Kanade [27],[28] work on vertical edges by applying 3 x 3 horizontal differential kernel. Neighboring edges are joined after removing the small edges. To minimize the false alarms, heuristics including the size of text object, fill factor and aspect ratio are used.

Cai et al. [29] uses the color edge detection YUV color space. It uses edge features like edge density, edge strength and horizontal alignment to detect text in an image. It uses two kernels for image enhancement and projection profile is used to localize the text precisely.

Lyu et al. [30] proposed the sequential multiresolution paradigm and background complexity based adaptive thresholding edge detection method for multilingual text extraction. The method defined four language dependant features; stroke density, font size, aspect ratio, and stroke statistics; and four language independent features; contrast, color, orientation, and stationary location. It used the signature-based multiframe verification for minimizing false alarms and the dam-point-based inward filling for text extraction.

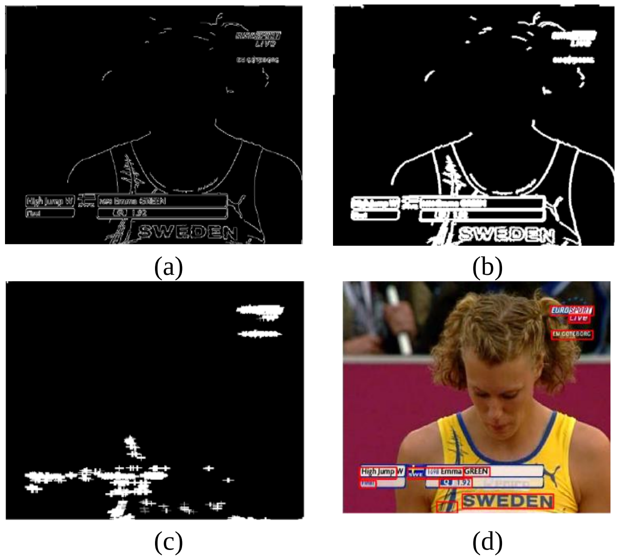
Liu et al. [31] propose a multiscale edge-based text extraction algorithm. The method consists of three stages, namely, candidate text region detection, text region localization and character extraction. Edge strength, density and variance of orientations features are used to detect the text regions. Text region localization uses the dilation operator to group the characters in a text object and finally the binary image is generated to be fed into the OCR.

Anthimopoulos et al. [32] proposed a two-stage methodology for text detection in video images. In the first stage, text lines are detected based on the Canny edge map of the image. In the second stage, the result is refined using a sliding window and an SVM classifier trained on features obtained by a new Local Binary Pattern-based operator (eLBP) that describes the local edge distribution. The whole algorithm is used in a multiresolution fashion enabling detection of characters for a broad size range. Fig. 6 shows the step wise results of the method.

Yao et al. [33] proposed an edge based algorithm for text detection. First the canny edge map is generated from the input image then stroke width transform operator is used to group the neighboring pixels to form text objects. Greedy hierarchical agglomerative clustering method is applied to aggregate the pairs into candidate chains. This method links the characters in arbitrary directions, and text may not necessarily be in the horizontal direction. Random Forest is used as the chain level classifier to get the final results.

Wei et al. [34] proposed a pyramidal scheme to detect text in images. First input image is resized into grayscale images of three different sizes. Then, the horizontal gradients, vertical gradients and the maximum gradient difference maps of the image pyramid are calculated. k-means clustering is applied on energy uniformity maps of MGD map to segregate text and non text pixels. Geometrical constraint along with the SVM is used to produce the final results.

Shivakumara et al. [35] proposed a method which used edge maps and quad tree to extract text in images. The pixels are grouped together based on their R, G and B values to enhance text information. K-means with k=2 is used to differentiate potential text candidate pixels from non text pixels. Stroke width based symmetry property is used for further authentication of potential text pixels. These authenticated text objects are then utilized as seed points to reinstate the text information with reference to the Sobel edge map of the original input image. Quad tree is employed to conserve the spatial locations of text pixels. Region growing is applied on Sobel edge map to formulate the text lines.



(a) (b)

(c) (d)

Fig 6. Lyu’s method (a) Edge map (b) Dilated edge map (c) Opening on the edge map (d) Final result after machine learning refinement

Different edge detection techniques have been used in the literature for segmentation. Canny [35], [36] and Sobel [37],[38] are the most commonly used edge detection techniques for the text extraction processes. This class of methodologies is effectual in detecting text regions when background is comparatively smooth and has low edge density. This type of methods fails when background has greater intensity transitions or it has complex patterns. These techniques work well for small and medium font sizes, but usually not suitable for large fonts.

5.1.2 Corner Based Segmentation

Text objects are affluent of corners, which are characteristically spread over the text regions uniformly. Exploiting this feature, many text detection techniques are based on density and uniformity of corners in an image.

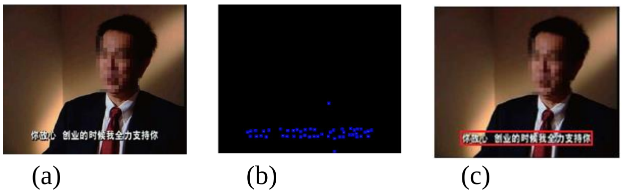
A Susan corner map is used by Hua et. al. [39] to detect the text objects in an image. Non text corners are eliminated using the corner density values. Remaining corners are merged using spatial constraints to form text objects. Corner density, edge density, the ratio of vertical edge density and horizontal edge density, and center offset ratio of edges extorted from vertical, horizontal, and overall edge maps are calculated to segregate text regions into text lines and minimize the false alarms.

Harris corners are used by Bai *et al*. [40] to detect text objects. Corners are merged and grouped on the basis of color and spatial similarities. Color and spatial similarities are measured by color histogram and Euclidean distances respectively.

Sun *et al*. [41] used corner response instead of corner counts to detect the text objects. Block-based corner density is exploited to localize text objects. The high corner density blocks are merged to generate text regions. Text verification is achieved by color deviation. Finally text localization is refined by horizontal and vertical profiles.

Zhao *et al*. [42] also used Harris corners to detect and localized caption text in images and videos. Morphological dilation is used to group the neighboring corners to form text objects. Detection results are further refined by eliminating false positives. False positive elimination is executed using area, foreground and background ratio, aspect ratio, orientation and position. This method also proposed the text tracking mechanism for moving text using text detection results and motion vectors. Results of Zhao’s method can be visualized in Fig. 7.

Corner metric and Laplacian filtering is used by Kaushik and Suresha [43] to detect text in images. Kanade-Tomasi corner detection is used to detect the corner metric. Combination of corner metric and laplacian image is achieved by image multiplication. Image multiplication also removes the noise induced by filtering and corner detection. Finally the detection results are localized and binarized to get fed into the OCR.



(a) (b) (c)

Fig 7. Zhao’s method (a) Input image (b) Corner detection result (c) Text detection result

Harris corner detector [44] is used in majority of the corner based text detection methods. Corner based methods give good results in high resolution images, with comparatively smoother backgrounds. These techniques don’t perform very well in low contrast images with low spatial and intensity resolutions. The response of corner based techniques is not appreciable in the presence of artefacts introduced by quanitization and compression.

5.1.3 Color and Intensity Based Segmentation

Color-based techniques work on the study that the color of text entities is uniform and very much contrasting from background. It means color and intensity difference is very low within text object and has high difference with the background. This observation guides many researchers to exploit the color based techniques for text detection. Commonly used color based techniques are clustering methods, intensity histograms, and binarization methods.

Ezaki *et al*. [45] proposed text extraction method that combines global and local image binarizations. This technique is fundamentally based on Otsu's binarization method and it is applied in all three-color channels. The input image is divided in non-overlapping tiles of pixels. For every such tile, Fisher's Discriminant Rate (FDR) is computed from the histogram. The FDR can be used for detecting the image tiles with a bimodal gray-level histogram. For image tiles with high FDR values, the local Otsu threshold is used for binarizing the image. For tiles with low FDR values, the global Otsu threshold is used instead.

Fu *et al*. 0 and Liu *et al*. [47] used Gaussian Mixture Modeling (GMM) of three neighbor characters to discriminate between characters and non-characters. Based on this modeling, the text in an input image is extracted in three steps. Firstly, the image is binarized and the morphological closing operation is used for merging and grouping of regions on the binary image. Then the neighborhood of all the connected components is established by partitioning the image into Voronoi regions of centroids of connected components. Finally, each connected component is labeled as character or not according to all its neighborhood relationships.

Kim [48] has proposed a technique using Local Color Quantization. First the input image is converted to 256-colored image and then Local Color Quantization is executed for every color of the image. Based on the text region features, regions are merged to form the text candidates. The weakness of this approach is the highl computational time as Local Color Quantization is performed for each and every color.

RGB color space is used by Lienhart and Effelsberg [49] for text detection in videos. Homogeneity of text color within the text object and visible contrast with its background is use to connect the neighboring pixels to form the connected components. These regions are merged depending upon the color similarity. False alarms are discarded by using basic geometrical features.

Pei and Chuang [50] proposed a text detection methodology based on 3D color histogram. Input image is first quantized to several quantized images with different number of quantized color. Each quantized image was put to 3D histogram analysis to find the text candidates. After applying some spatial constraints and relationship rules, text candidates would be identified. Finally, all the quantized images are combined to locate the text precisely.

Kim *et al*. [51] detect and localized text using the focus of mobile camera. The candidate text color in HCL space is found using mean shift algorithm based on the assumption that text region occupies most portions of the focus. This text color is considered as the seed color to binarize the focus and generate text component. Text regions are localized by expanding text component regions iteratively. Five geometric heuristic conditions are used to stop the component expansion.

Fu *et al*. [52] propose a text detection method in complex background based on multiple constraints. Preliminary segmentation is implemented by K-means clustering based on YCbCr color vectors. K=3 or 4 depending on the number of humps that appear in the histogram of an image. After obtaining CCs using clustering results, four constraints are applied to perform post-processing to eliminate background residues. (i) Color constraint, all CCs corresponding to text objects should have homogeneous colors. (ii) Edge magnitude constraint, the boundaries of text CCs should go with strong edges. (iii) Thickness constraint, character strokes should have proper size and CCs whose height or width exceeds the thresholds are removed. (iv) Components relation constraint, such as dimensional range, combination of two components, and compactness of two components.

Yi [53] proposed K-means clustering based text detection method. First, the edged image is calculated and then corresponding edges are repainted on the original image. Pixels are then sampled with locally extreme values and initialize several color clusters in RGB color space by calculating the expected color values. Then K-means cluster algorithm is executed to group the pixels with similar colors together. Color reduction segments the input image into several color layers. Each color layer consists of only foreground color on white background.

Data clustering for image data is primarily means dividing a set of pixels into natural partitions. This partitioning has very important role in image processing, analysis and understanding. K-means clustering is a technique that partitions the objects into K mutually exclusive clusters, by maximizing the inter cluster distances and minimizing the intra cluster distances. Each cluster is described by its centroid.

K-means clustering is a natural tool for clustering and segmenting colored images and because of the reason; it is widely used in text detection applications [54]-[57]. However, the k-means algorithm is highly dependent upon the value of ‘k’ i.e. number of clusters in the partition. So the number of k has to be defined for color segmentation before clustering. The optimum number of clusters may vary from image to image. There must be some methodology for finding the adaptive ‘k’ that can give optimum results for all the images. Some methods are reported in the literature that claims to calculate the k that is dependent upon the nature of the data. Sugar [58] and Lleti *et al*. [59] defined the mechanism for finding the optimum number for k, but these require a-priori clustering before the actual one, hence it decrease the computational efficiency. So it is required to find the simple yet optimum solution for this problem specific to text detection methodologies.

5.2 Merging and Grouping

Segmentation identifies the occurrence of different regions in the image but does not recognize the relation between these regions. It is substantial to merge the characters of a word to form a text object, because most of the text detection techniques work on group of characters and it is very difficult to detect the isolated character [14],[60]. This grouping can utilize the pixel level features or can exploit the high-level features.

5.2.1 Pixel Level Merging

Presently few pixel level merging methods are introduced in the literature pertaining to text detection. Dilation is most commonly used as merging technique. Dilation is the basic operation in the area of mathematical morphology. It is classically operated on binary images, but there are adaptations in grayscale images too. The primary outcome of the operator on a binary image is to steadily expand the boundaries of foreground regions. Consequently, it increases the size of the foreground objects. The dilation operator takes two pieces of data as inputs. The first is the image which is to be dilated. The second is a set of coordinate points known as a structuring element or kernel. The amount and direction of increase is dependent upon the structuring element of the dilation operator.

Many text detection techniques have used this simple operator for the merger of neighboring text characters. It is computationally very simple and hence has been adopted by many text detection techniques.

Farhoodi and Kasaei [61] use the morphological dilation operation on the processed edge map. A horizontally longer structuring element of fixed size is used for dilation. Das *et al*. [62] also used the morphological dilation operator for merger and enhancement of text regions.

Deepa and Victor [60] tried different structuring elements and found that the most suitable structuring element is a disk shaped structuring element of 6 pixels radius.

Shutao and Kwok [63] morphologically dilate the edge map with a square structure to produce a segmentation map. Square-shaped kernel of 8x8 size is used in experiments. This choice is based on the observation that most of the text objects are in square shape. Poignant *et al*. [64] used horizontal dilation and erosion to connect the characters of the same string.

Indeed, the size of the morphological operator intrinsically characterizes the size of the homogeneous segmented regions. Thus, large text areas are prone to over-segmentation, while small text regions might be skipped. Fixed size of the structuring element can only materialize for limited spatial resolutions and small range of font sizes. More so, size of the structuring element should be dependent upon the size of the text, but usually has the fixed value which cannot deal with the variation in resolution of image and size of text.

Some methodologies in literature use the pyramid approach to solve this problem and extend the range of text sizes for detection [62], [65], [66]. This highly increases the computational requirements or demands for parallel processing mechanisms.

5.2.2 Object Level Merging

Object level merging is more close to human vision and deals with the objects and regions instead of pixels. It connects the potential character objects to form the text strings. Hence the grouping and merging is dependent upon some high level features which gives better performance.

Wolf and Jolion [5] used the conditional dilation and erosion to merge the neighboring character candidates. Dilation is performed, if the relative difference of the heights of the connected component including the given pixel and the neighboring component to the right does not exceed a threshold and the horizontal distance of these two regions should not surpass a different threshold. If the assessment is convincing and the dilations count does not reach the ceiling value, then the pseudo color is placed as pixel value. The conditional erosion step performs an operation based on the additional conditions on the gray levels of the pixels instead of the shape. The image is eroded with the condition that only pixels marked with the pseudo color are eroded. The effect of this step is the connection of the all connected components which are horizontally aligned and whose heights are similar. Results of this method are shown in Fig. 8.

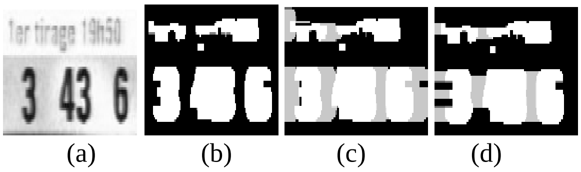


Fig 8. Wolf and Jolian method (a) Original image (b) Binarized accumulated gradients (c) After conditional dilation (d) After conditional erosion.

Minetto *et al*. [65] developed a grouping step where all recognized characters are grouped all together with their neighbors to recover the text regions. The conditions to link two characters are based on the distance between the two regions relative to their height.

Pan *et al* [67] built component tree using minimum spanning tree. This text detection method merges the text characters into words and text objects using shape difference and spatial difference. These features are chosen based on the observation that components belonging to the same text region are spatially close and have similar shapes.

González and Bergasa [68] suggested that characters should have some similar attributes, such as stroke width, height, alignment, adjacency and constant inter-letter and inter-word spacing. The process of merging chains of candidates was repeated until no more groups can be merged. This approach assumed that letters do not appear alone in an image, so those groups are rejected which have less than 3 elements.

Shi *et al*. [24] used the graph model to merge the neighboring regions to form text strings. Each node in the undirected graph is supposed as the extracted region and the neighboring nodes for each node are those ones that satisfy the criterion defined by difference in color and position, width ratio and height ratio.

Character candidates are linked into pairs Yao *et al.* [33] method. If two candidates have similar stroke widths (ratio between the mean stroke widths is less than 2.0), similar sizes (ratio between their characteristic scales does not exceed 2.5), similar colors and are close enough (distance between them is less than two times the sum of their characteristic scales), they are labeled as a pair. Next, a greedy hierarchical agglomerative clustering method is applied to aggregate the pairs into candidate chains.

These features are though defined by strict boundaries in existing techniques; the relation between the neighboring characters is not crisp.

5.3 Feature Vector

Very less work has been done in defining the novel feature vectors for text detection. Most of the existing systems use few conventional features to classify text and non text objects. These features are generally defining few geometrical features of the text objects.

Zhong *et al.* [69] used a CC-based method using color reduction. They quantize the color space using the peaks in a color histogram in the RGB color space. Each text component goes through a filtering stage using heuristics, such as area, diameter, and spatial alignment.

Two geometrical constraints are applied by Wolf and Jolion [5] to eliminate the non text and detect the text objects from videos. One is the width to height ratio and the second one is number of text pixels of the component to area of the bounding box.

Simple rules are used by Ezaki [45] to filter out the false detections. They imposed constraints on the aspect ratio and area to decrease the number of non-character candidates. Isolated characters are also eliminated from the text candidate list.

Hua *et al.* [39] used the constraints on height and width of the text candidates to reduce the false alarms. They also defined fill factor constraint to further reduce the non text objects. They defined the upper and lower limits for ratio of horizontal edge points to vertical edge points. They have also defined the upper limit for the ratio of edge points to total number of pixels in the area. Here the edge points represent horizontal edge, vertical edge and overall edge.

Epshtein *et al.* [26] present a novel image operator that seeks to find the value of stroke width for each image pixel, and demonstrate its use on the task of text detection in natural images. Many of the recent techniques are using this operator as part of text detection feature vector.

Local binary pattern is being used by Wei and Lin [34] for texture analysis. They first extracted the statistic feature of each text candidate by resizing each text candidate to 128x128 size. They then used Haar wavelet transform to decompose the text candidate to the four sub-band images including: low frequency (LL) band, vertical high frequency (LH) band, horizontal high frequency (HL) band and high frequency (HH) band. Next, they calculated the features in four sub-bands including mean, standard deviation and entropy of each sub-band. In addition to these statistic features, five features of the gray-level co-occurrence matrix (GLCM); energy, entropy, contrast, homogeneity and correlation, are calculated for each four direction in four wavelet sub-bands. 92-dimensional feature vector for each text candidate was generated, which was reduced to 36-dimensions using the principal component analysis (PCA).

After applying the morphological dilation on detected corner points in the image, [42] used five region properties as the features to describe text regions. These features are area, saturation, orientation, aspect ratio and position. The area is the foreground pixels in the bounding box. Saturation specifies the proportion of the foreground pixels in the bounding box that also belong to the region. Orientation is defined as the angle between x-axis and the major axis of the ellipse that has the same second-moments as the region. Aspect ratio of the bounding box is defined as the ratio of its width to its height. Position is defined by the region’s centroid.

Shivakumara *et al.* [15] used two features to eliminate the false positives. One is the straightness and the other one is The first feature, straightness, comes from the observation that text strings appear on a straight line (their assumption), while false positives can have irregular shapes. The second feature, edge density, is defined as the ratio of edge length to the connected component area. Ranjiniand and Sundaresan [70] used the area to find the text area blob.

There is a need of in-depth study of text structures. Anatomical study of human text detection can be useful for identification of such features. And there is also a need to mathematically model those bio inspired features to make it workable for machines. Detailed geometrical and statistical study of text objects is also required.

VI. Text Tracking

Tracking superimposed text moving across several frames of a video is relevant for exploiting its temporal occurrence for effective video content indexing and retrieval. Text appear in video data is having an important property that is missing in images .i.e. Temporal redundancy: Text in videos lasts for some time, in order to make it read by the viewer. This property can be exploited to detect the false alarms in the detection and localization phase. Most of the presented methodologies work only for images only and cannot be applied for video data [71], [26],[14]. Very few works have been done on text tracking in the videos. Conventional methods [72],[73] for the text recognition of video mainly focus on recognizing the text in each single frame independently.

Lienhart and Wernicke [8] defined a text tracker that took the text line in a video frame, calculates a characteristic signature which allows discrimination of this text line from text lines with other contents, and searches in the next video frame for the image region of the same dimension which best matches the reference signature. Moreover, the system is not only able to locate and segment text occurrences into large binary images, but is also able to track each text line with sub-pixel accuracy over the entire occurrence in a video, so that one text bitmap is created for all instances of that text line.

Wolf and Joilion [5] used the overlap information between the list of rectangles detected in the current frame and the list of currently active rectangles (i.e. the text detected in the previous frames which is still visible in the last frame). Difference in size, position and size of the overlap area are used to find the similarity between the text objects in the adjacent frames.

Li *et al.* [74] presented a text tracking approach that used the SSD(Sum of Squared Difference) for a pure translational motion model, but it will fail when the background change greatly.

Xu *et al.* [75] proposed text extraction in DCT compressed domain. Potential text blocks are located in terms of DCT texture energy. An adaptive temporal constraint method is proposed to exploit the temporal occurrence of text in a sequence of frames. Results are verified on MPEG video sequences.

Gllavata *et al.* [76] present a text tracking method based on motion vectors for MPEG videos. They tracked the text within a group of pictures (GOP) using MPEG motion vector information extracted directly from the compressed video stream.

Qian *et al.* [77] propose a methodology of text tracking for compressed video. Horizontal and vertical texture intensity projection profiles and Mean Absolute Difference (MAD) is used to track the rolling and static text respectively.

Huang *et al.* [78] used temporal information obtained by dividing a video frame into sub-blocks and calculating inter-frame motion vector for each sub-block. This deals only scrolling text with same trajectory and velocity. It also does not allow the text to scale or deform during its appearance in the video frames. Tanaka *et al.* [79] use the cumulative histogram to compute the similarity of detected text block with the blocks in the consecutive frames.

Optical flows are used by Zhao *et al.* [42] as the motion feature for moving caption text tracking. Optical flow estimation is used to compute an approximation to the motion field from intensity difference of two consecutive frames. Multiresolution Lucas-Kanade algorithm is used for optical flow estimation.

Zhen *et al.* [80] and Li *et al.* [81] dealt with multi frame integration but dealt only with stationary text in videos. Particle filter is used for text tracking in some wearable camera applications [82], [83].

Text tracking algorithms presented in the literature are mainly dependent upon few global or local features which fail for videos with complex backgrounds. Moreover, these methods can mainly apply on stationary text or text with simple movements such as simple scrolling credits. So there is a need a methodology, which can deal with complex text movements such as zoom in and out, having complex backgrounds.

VII. CONCLUSION

Text data appear in the multimedia documents can be a vital tool for indexing and retrieving the multimedia contents. Many approaches are presented in the literature to extract text from the images and videos. A detailed survey on state of the art techniques is presented in the paper. The paper covers in detail analysis of the text detection, localization and tracking techniques. The presented research also highlights the limitations and constraints of the existing methodologies. Although many approaches are presented for text detection, localization and tracking, but few aspects are fully or partially ignored in the literature. These aspects involve the variable font sized text detection and localization and text tracking for animated texts.

REFERENCES

[1] Michel, F. “How many photos are uploaded to Flicker every day, month, year?”, http://www.flickr.com/photos/franckmichel/6855169886/, Dec 5, 2013.

[2] “Official Blog: It’s YouTube’s 7th Birthday,” http://youtube-global.blogspot.com/2012/05/its-youtubes-7th-birthday-and-youve.html, Oct 2013.

[3] Misra, C., & Sural, S. (2006). Content based image and video retrieval using embedded text. In Computer Vision–ACCV 2006 (pp. 111-120). Springer Berlin Heidelberg.

[4] Antani, S., Crandall, D., & Kasturi, R. (2000). Robust extraction of text in video. In Pattern Recognition, 2000. Proceedings. 15th International Conference on(Vol. 1, pp. 831-834). IEEE.

[5] Wolf, C., & Jolion, J. M. (2004). Extraction and recognition of artificial text in multimedia documents. Formal Pattern Analysis & Applications, 6(4), 309-326.

[6] Vijayakumar, V., & Nedunchezhian, R. (2011). A Novel Method for Super Imposed Text Extraction in a Sports Video. International Journal of Computer Applications (0975–8887) Volume.

[7] Pickering, M. J., Wong, L., & Rüger, S. M. (2003). ANSES: Summarisation of news video. In Image and Video Retrieval (pp. 425-434). Springer Berlin Heidelberg.

[8] Lienhart, R., & Wernicke, A. (2002). Localizing and segmenting text in images and videos. Circuits and Systems for Video Technology, IEEE Transactions on, 12(4), 256-268.

[9] Antonacopoulos, A., Karatzas, D., & Ortiz-Lopez, J. (2000, December). Accessing textual information embedded in internet images. In Photonics West 2001-Electronic Imaging (pp. 198-205). International Society for Optics and Photonics.

[10] Minetto, R., Thome, N., Cord, M., Leite, N. J., & Stolfi, J. (2012). T-HOG: An effective gradient-based descriptor for single line text regions. Pattern Recognition.

[11] Neumann, L., & Matas, J. (2013) On Combining Multiple Segmentations in Scene Text Recognition. 12th International Conference of Document Analysis and Recognition (ICDAR)

[12] Zhao, M., Li, S., & Kwok, J. (2010). Text detection in images using sparse representation with discriminative dictionaries. Image and Vision Computing, 28(12), 1590-1599.

[13] Wang, K., & Belongie, S. (2010). Word spotting in the

wild. In Computer Vision–ECCV 2010 (pp. 591-604). Springer Berlin Heidelberg.

[14] Neumann, L., & Matas, J. (2011). A method for text localization and recognition in real-world images. In Computer Vision–ACCV 2010 (pp. 770-783). Springer Berlin Heidelberg.

[15] Shivakumara, P., Phan, T. Q., & Tan, C. L. (2011). A laplacian approach to multi-oriented text detection in video. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 33(2), 412-419.

[16] Jung, K., In Kim, K., & K Jain, A. (2004). Text information extraction in images and video: a survey. Pattern recognition, 37(5), 977-997.

[17] Liang, J., Doermann, D., & Li, H. (2005). Camera-based analysis of text and documents: a survey. International Journal of Document Analysis and Recognition (IJDAR), 7(2-3), 84-104.

[18] Sumathi, C.P., Santhanam, T., & Gayathri G. (2012). A Survey on various approaches of text extraction in images. International Journal of Computer Science & Engineering Survey (IJCSES), 3(4).

[19] Lienhart, R. (2003). Video OCR: a survey and practitioner's guide. In Video mining (pp. 155-183). Springer US.

[20] Li, C., Ding, X. G., & Wu, Y. S. (2006). An Algorithm for Text Location in Images Based on Histogram Features and Ada-boost. Journal of Image and Graphics, 3, 003, 325-331

[21] Kim, K. I., Jung, K., & Kim, J. H. (2003). Texture-based approach for text detection in images using support vector machines and continuously adaptive mean shift algorithm. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 25(12), 1631-1639.

[22] Gllavata, J., Qeli, E., & Freisleben, B. (2006, December). Detecting text in videos using fuzzy clustering ensembles. In Multimedia, 2006. ISM'06. Eighth IEEE International Symposium on (pp. 283-290). IEEE.

[23] Chen, D., Odobez, J. M., & Bourlard, H. (2004). Text detection and recognition in images and video frames. Pattern Recognition, 37(3), 595-608.

[24] Shi, C., Wang, C., Xiao, B., Zhang, Y., & Gao, S. (2012). Scene text detection using graph model built upon maximally stable extremal regions. Pattern Recognition Letters.

[25] León Cristóbal, M., Vilaplana Besler, V., Gasull Llampallas, A., & Marqués Acosta, F. (2012). Region-based caption text extraction. 11th International Workshop On Image Analysis For Multimedia Interactive Services (Wiamis).

[26] Epshtein, B., Ofek, E., & Wexler, Y. (2010, June). Detecting text in natural scenes with stroke width transform. In Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on (pp. 2963-2970). IEEE.

[27] Sato, T., Kanade, T., Hughes, E. K., & Smith, M. A. (1998, January). Video OCR for digital news archive. In Content-Based Access of Image and Video Database, 1998. Proceedings., 1998 IEEE International Workshop on (pp. 52-60). IEEE.

[28] Sato, T., Kanade, T., Hughes, E. K., Smith, M. A., & Satoh, S. I. (1999). Video OCR: indexing digital news libraries by recognition of superimposed captions. Multimedia Systems, 7(5), 385-395.

[29] Cai, M., Song, J., & Lyu, M. R. (2002). A new approach for video text detection. In Image Processing. 2002. Proceedings. 2002 International Conference on (Vol. 1, pp. I-117). IEEE.

[30] Lyu, M. R., Song, J., & Cai, M. (2005). A comprehensive method for multilingual video text detection, localization, and extraction. Circuits and Systems for Video Technology, IEEE Transactions on, 15(2), 243-255.

[31] Liu, X., & Samarabandu, J. (2006, July). Multiscale edge-based text extraction from complex images. In Multimedia and Expo, 2006 IEEE International Conference on (pp. 1721-1724). IEEE.

[32] Anthimopoulos, M., Gatos, B., & Pratikakis, I. (2010). A two-stage scheme for text detection in video images. Image and Vision Computing, 28(9), 1413-1426.

[33] Yao, C., Bai, X., Liu, W., Ma, Y., & Tu, Z. (2012, June). Detecting texts of arbitrary orientations in natural images. In Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on (pp. 1083-1090). IEEE.

[34] Wei, Y. C., & Lin, C. H. (2012). A robust video text detection approach using SVM. Expert Systems with Applications, 39(12), 10832-10840.

[35] Shivakumara, P., Basavaraju, H. T., Guru, D. S., & Tan, C. L. (2013, August). Detection of Curved Text in Video: Quad Tree Based Method. In Document Analysis and Recognition (ICDAR), 2013 12th International Conference on (pp. 594-598). IEEE.

[36] Dutta, A., Pal, U., Bandyopadhya, A., & Tan, C. L. (2009). Gradient based Approach for Text Detection in Video Frames 1.

[37] Phan, T. Q., Shivakumara, P., & Tan, C. L. (2009, July). A Laplacian method for video text detection. In Document Analysis and Recognition, 2009. ICDAR'09. 10th International Conference on (pp. 66-70). IEEE.

[38] Sharma, N., Shivakumara, P., Pal, U., Blumenstein, M., & Tan, C. L. (2012, March). A New Method for Arbitrarily-Oriented Text Detection in Video. InDocument Analysis Systems (DAS), 2012 10th IAPR International Workshop on (pp. 74-78). IEEE.

[39] Hua, X. S., Chen, X. R., Wenyin, L., & Zhang, H. J. (2001, September). Automatic location of text in video frames. In Proceedings of the 2001 ACM workshops on Multimedia: multimedia information retrieval (pp. 24-27). ACM.

[40] Bai, H., Sun, J., Naoi, S., Katsuyama, Y., Hotta, Y., & Fujimoto, K. (2008, December). Video caption duration extraction. In Pattern Recognition, 2008. ICPR 2008. 19th International Conference on (pp. 1-4). IEEE.

[41] Sun, L., Liu, G., Qian, X., & Guo, D. (2009, June). A novel text detection and localization method based on corner response. In Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on (pp. 390-393). IEEE.

[42] Zhao, X., Lin, K. H., Fu, Y., Hu, Y., Liu, Y., & Huang, T. S. (2011). Text from corners: a novel approach to detect text and caption in videos. Image Processing, IEEE Transactions on, 20(3), 790-799.

[43] Kaushik K.S., Suresha D. (2013). Automatic Text Extraction in Video Based on the Combined Corner Metric and Laplacian Filtering Technique. International Journal of Advanced Research in Computer Engineering & Technology, 2(6).

[44] Harris, C., & Stephens, M. (1988, August). A combined corner and edge detector. In Alvey vision conference (Vol. 15, p. 50), 147–152.

[45] Ezaki, N., Kiyota, K., Minh, B. T., Bulacu, M., & Schomaker, L. (2005, August). Improved text-detection methods for a camera-based text reading system for blind persons. In Document Analysis and Recognition, 2005. Proceedings. Eighth International Conference on (pp. 257-261). IEEE.

[46] Fu, H., Liu, X., Jia, Y., & Deng, H. (2006, October).

Gaussian mixture modeling of neighbor characters for multilingual text extraction in images. In Image Processing, 2006 IEEE International Conference on (pp. 3321-3324). IEEE.

[47] Liu, X., Fu, H., & Jia, Y. (2008). Gaussian mixture modeling and learning of neighboring characters for multilingual text extraction in images. Pattern Recognition, 41(2), 484-493.

[48] Kim, P. K. (1999). Automatic text location in complex color images using local color quantization. In TENCON 99. Proceedings of the IEEE Region 10 Conference (Vol. 1, pp. 629-632). IEEE.

[49] Lienhart, R., & Effelsberg, W. (2000). Automatic text segmentation and text recognition for video indexing. Multimedia systems, 8(1), 69-81.

[50] Pei, S. C., & Chuang, Y. T. (2004, June). Automatic text detection using multi-layer color quantization in complex color images. In Multimedia and Expo, 2004. ICME'04. 2004 IEEE International Conference on (Vol. 1, pp. 619-622). IEEE.

[51] Kim, E., Lee, S., & Kim, J. (2009, July). Scene text extraction using focus of mobile camera. In Document Analysis and Recognition, 2009. ICDAR'09. 10th International Conference on (pp. 166-170). IEEE.

[52] Fu, L., Wang, W., & Zhan, Y. (2005). A robust text segmentation approach in complex background based on multiple constraints. In Advances in Multimedia Information Processing-PCM 2005 (pp. 594-605). Springer Berlin Heidelberg.

[53] Yi, C. (2010, October). Text locating in scene images for reading and navigation aids for visually impaired persons. In Proceedings of the 12th international ACM SIGACCESS conference on Computers and accessibility (pp. 325-326). ACM.

[54] Shivakumara, P., Phan, T. Q., & Tan, C. L. (2009, July). A robust wavelet transform based technique for video text detection. In Document Analysis and Recognition, 2009. ICDAR'09. 10th International Conference on (pp. 1285-1289). IEEE.

[55] Lee, S., Cho, M. S., Jung, K., & Kim, J. H. (2010, August). Scene text extraction with edge constraint and text collinearity. In Pattern Recognition (ICPR), 2010 20th International Conference on (pp. 3983-3986). IEEE.

[56] Aradhya, V. M., & Pavithra, M. S. (2013). An Application of K-Means Clustering for Improving Video Text Detection. In Intelligent Informatics (pp. 41-47). Springer Berlin Heidelberg.

[57] Šarić, M., Dujmić, H., & Russo, M. (2013). Scene Text Extraction in HSI Color Space using K-means Algorithm and Modified Cylindrical Distance.

[58] Sugar, C. A., & James, G. M. (2003). Finding the number of clusters in a dataset. Journal of the American Statistical Association, 98(463).

[59] Lletı, R., Ortiz, M. C., Sarabia, L. A., & Sánchez, M. S. (2004). Selecting variables for< i> k</i>-means cluster analysis by using a genetic algorithm that optimises the silhouettes. Analytica Chimica Acta, 515(1), 87-100.

[60] Deepa, S. T., & Victor, S. P. (2013). A novel method for text extraction. International Journal of Engineering Science & Advanced Technology, 2(4), 961 – 964.

[61] Farhoodi, R., & Kasaei, S. (2005, May). Text segmentation from images with textured and colored background. In Proceedings of 13th Iranian Conference on Electrical Engineering. Zanjan, Iran.

[62] Das, M. S., Bindhu, B. H., & Govardhan, A. (2012). Evaluation of Text Detection and Localization Methods in Natural Images. International Journal of Emerging Technology and Advanced Engineering, 2(6), 277-282.

[63] Li, S., & Kwok, J. T. (2004, October). Text extraction using edge detection and morphological dilation. In Intelligent Multimedia, Video and Speech Processing, 2004. Proceedings of 2004 International Symposium on (pp. 330-333). IEEE.

[64] Poignant, J., Besacier, L., Quenot, G., & Thollard, F. (2012, July). From text detection in videos to person identification. In Multimedia and Expo (ICME), 2012 IEEE International Conference on (pp. 854-859). IEEE.

[65] Minetto, R., Thome, N., Cord, M., Fabrizio, J., & Marcotegui, B. (2010, September). Snoopertext: A multiresolution system for text detection in complex visual scenes. In Image Processing (ICIP), 2010 17th IEEE International Conference on (pp. 3861-3864). IEEE.

[66] Anthimopoulos, M., Gatos, B., & Pratikakis, I. (2007, March). Multiresolution text detection in video frames. In VISAPP (2) (pp. 161-166).

[67] Pan, Y. F., Hou, X., & Liu, C. L. (2011). A hybrid approach to detect and localize texts in natural scene images. Image Processing, IEEE Transactions on, 20(3), 800-813.

[68] Gonzalez, A., & Bergasa, L. M. (2013). A text reading algorithm for natural images. Image and Vision Computing.

[69] Zhong, Y., Karu, K., & Jain, A. K. (1995). Locating text in complex color images. Pattern recognition, 28(10), 1523-1535.

[70] Ranjini, S., & Sundaresan, M. (2013). Extraction and Recognition of Text From Digital English Comic Image Using Median Filter. International Journal.

[71] León Cristóbal, M., Vilaplana Besler, V., Gasull Llampallas, A., & Marqués Acosta, F. (2013). Region-based caption text extraction. Analysis, Retrieval and Delivery of Multimedia Content 2013, Springer New York, 21-36.

[72] Zhang, X., Sun, F., & Gu, L. (2010, August). A combined algorithm for video text extraction. In Fuzzy Systems and Knowledge Discovery (FSKD), 2010 Seventh International Conference on (Vol. 5, pp. 2294-2298). IEEE.

[73] Zhiming, W., & Yu, X. (2010, April). An Approach for Video-Text Extraction Based on Text Traversing Line and Stroke Connectivity. In Biomedical Engineering and Computer Science (ICBECS), 2010 International Conference on (pp. 1-3). IEEE.

[74] Li, H., Doermann, D., & Kia, O. (2000). Automatic text detection and tracking in digital video. Image Processing, IEEE Transactions on, 9(1), 147-156.

[75] Xu, J., Jiang, X., & Wang, Y. (2009, March). Caption Text Extraction Using DCT Feature in MPEG Compressed Video. In Computer Science and Information Engineering, 2009 WRI World Congress on (Vol. 6, pp. 431-434). IEEE.

[76] Gllavata, J., Ewerth, R., & Freisleben, B. (2004, October). Tracking text in MPEG videos. In Proceedings of the 12th annual ACM international conference on Multimedia (pp. 240-243). ACM.

[77] Qian, X., Liu, G., Wang, H., & Su, R. (2007). Text detection, localization, and tracking in compressed video. Signal Processing: Image Communication, 22(9), 752-768.

[78] Huang, W., Shivakumara, P., & Tan, C. L. (2008, December). Detecting moving text in video using temporal information. In Pattern Recognition, 2008. ICPR 2008. 19th International Conference on (pp. 1-4). IEEE.

[79] Tanaka, M., & Goto, H. (2007, September). Autonomous text capturing robot using improved dct feature and text racking. In Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on (Vol. 2, pp. 1178-1182). IEEE.

[80] Zhen, W., & Zhiqiang, W. (2010, August). An Efficient Video Text Recognition System. In Intelligent Human-Machine Systems and Cybernetics (IHMSC), 2010 2nd International Conference on (Vol. 1, pp. 174-177). IEEE.

[81] Li, L. J., Li, J., & Wang, L. (2010, July). An integration text extraction approach in video frame. In Machine Learning and Cybernetics (ICMLC), 2010 International Conference on (Vol. 4, pp. 2115-2120). IEEE.

[82] Tanaka, M., & Goto, H. (2008, December). Text-tracking wearable camera system for visually-impaired people. In Pattern Recognition, 2008. ICPR 2008. 19th International Conference on (pp. 1-4). IEEE.

[83] Goto, H., & Tanaka, M. (2009, July). Text-tracking wearable camera system for the blind. I Document Analysis and Recognition, 2009. ICDAR'09. 10th International Conference on (pp. 141-145). IEEE.

Authors' Profiles

Samabia Tehsin is a PhD Scholar at MCS, NUST. She did his MS Software Engineering from NUST in 2007. Her areas of research are Digital Image processing, computer Vision and Document Analysis.

Asif Masood-Dr Asif Masood did his BE in software Engineering from Military College of Signals (MCS), NUST in 1999. He completed his MS and PhD in Computer Science from University of Engineering and Technology Lahore in 2007. Currently, he is working in MCS, National University of Science and Technology.

Sumaira Kausar is a PhD Scholar at CEME NUST. Her research interests are Digital image Processing, Computer Vision and machine learning.

How to cite this paper: Samabia Tehsin, Asif Masood, Sumaira Kausar,"Survey of Region-Based Text Extraction Techniques for Efficient Indexing of Image/Video Retrieval", IJIGSP, vol.6, no.12, pp. 53-64, 2014.DOI: 10.5815/ijigsp.2014.12.08