Text Detection in Images and Videos

Marios Anthimopoulos *

[1]Department of Informatics and Telecommunications
National and Kapodistrian University of Athens

[2]Computational Intelligence Laboratory
Institute of Informatics and Telecommunications
National Centre for Scientific Research "Demokritos

anthimop@iit.demokritos.gr

Abstract. The goal of a multimedia text extraction and recognition system is filling the gap between the already existing and mature technology of Optical Character Recognition and the new needs for textual information retrieval created by the spread of digital multimedia. A text extraction system from multimedia usually consists of the following four stages: spatial text detection, temporal text detection – tracking (for videos), image binarization – segmentation, character recognition. In the framework of this PhD thesis we dealt with all the stages of a multimedia text extraction system, focusing though on the designing and development of techniques for the spatial detection of text in images and videos as well as methods for evaluating the corresponding result. Two methods for the evaluation of the text detection result were proposed that deal successfully with the problems of the related literature. Each of them uses different criteria while both of them are based on intuitively correct observations. Finally, a very efficient method was developed for the temporal detection of text which actually conduces to a better spatial detection while concurrently enhances the quality of the text image.

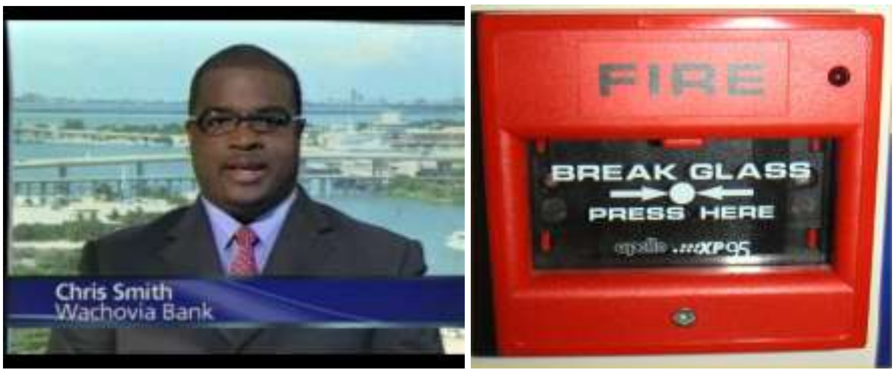
Keywords: text detection, text recognition, artificial text, scene text, natural scene images, Video OCR, multimedia information retrieval, evaluation.[1]

1 Introduction

Nowadays the size of the available digital video content is increasing rapidly. This fact leads to an urgent need for fast and effective algorithms for information retrieval from multimedia content. Textual information in multimedia constitutes a very rich source of high-level semantics for retrieval and indexing. Document image processing, after many decades of research, has reached a high level of text recognition accuracy, for traditional scanner-based images. However, these techniques fail to deal with text appearing in videos or camera-based images. The goal of a multimedia text extraction and recognition system is filling this gap between the already existing and mature technology of Optical Character Recognition and the new kinds of data.

Mainly, there exist two kinds of text occurrences in videos and images, namely artificial and scene text. Artificial text, as the name implies, is artificially added in order to describe the multimedia content while scene text is textual content that was captured by a camera as part of a scene, e.g. text on T-shirts or road signs. Figure 1(a) presents a video frame with artificial text and Figure 1(b) a natural scene image containing scene text. Text can also be classified into normal or inverse. Normal is denoted any text whose characters have lower intensity values than the background while inverse text is the opposite. In Figure 1(b), "FIRE" is considered as normal while the rest of the text is considered inverse.

[1] *Dissertation Advisors: [1]Sergios Theodoridis, Professor – [2]Dr. Basilis Gatos, Researcher



(a) (b)

Figure 1. (a) Video frame with artificial text, (b) Image with scene text

2 Related Work

Several methods for text extraction from multimedia have been proposed the last decade. Most of them are divided in the following stages: spatial text detection, temporal text detection, image binarization – segmentation and character recognition. From these stages the most crucial is the stage of spatial detection which actually concentrates the focus of most researchers in this field. The performance of the whole text extraction system depends on the accurate localization of text in an image or video frame. Much lesser work has been done for the temporal detection of text in videos while usually only static text is considered. Some methods have also been proposed for the binarization of the text image, although most researchers use state-of-the-art algorithms from the classic document analysis research area. Finally a very important aspect that has not been sufficiently studied is the development of the corresponding evaluation protocols which are necessary for optimizing the algorithms as well as comparing the several methods in literature. In this section we will outline the techniques found in literature for the different stages of text extraction from images and video frames.

In general, the existing spatial text detection methods can be roughly divided in two categories: region-based and texture-based. Region-based methods group pixels that belong to the same character based on the colour homogeneity, the strong edges between character and background or by using a stroke filter. Then, the detected characters are grouped to form textlines according to colour, size and geometrical rules. Texture-based algorithms scan the image at different scales using a sliding window and classify image areas as text or non-text based on texture-like features. Another possible categorization of text detection methods could be dividing them into heuristic and machine learning techniques. Typical heuristic, region-based approaches that rely on connected components can be found in [1-3]. Other region-based methods detect text based on edge or stroke information, i.e. strength, density or distribution [4-7]. DCT coefficients globally map the periodicity of intensity images and they have been widely used as texture features for heuristic texture-based methods [8-12]. Some hybrid methods have also been proposed [13-18]. Hybrid techniques combine the efficiency of a heuristic coarse stage with an accurate machine-learning refinement stage. Several pure machine learning, texture-based approaches have also been proposed for the detection of text areas with great success. These methods use directly machine learning classifiers to detect text [19-23]. The main shortcoming of the methods attributed to this category is the high computational complexity since a sliding window is required to scan the entire image, requiring thousands of calls to the classifier per image.

Very few researchers have focused n the problem of the temporal text detection mainly due to the objective difficulties of the task, the lack of an unbiased evaluation methodology and the great effort that is needed for the ground truth annotation. These existing attempts consider mainly artificial text and use standard object tracking methods while the result is usually evaluated optically. Lienhart et al. [24] considered text static or linearly moving and applied tracking with block matching using the least mean square criterion. Antani et al. [25] and Gargi et al. [11] use motion vectors in compressed MPEG-1 videos based on the work of Nakajama et al. [26] and Pilu et al. [27]. Li et al. [28] use affine transformation and Sum of Square Difference – SSD for the detection of moving text followed by a validation process which is based on Mean Square Error – MSE) and the moving text trajectory.

Within the framework of multimedia text extraction some binarization methods have also been proposed. Text detection methods based on connected components often performed image binarization followed by heuristic grouping of character components [2,29]. LeBourgeois et al. [30] applied the Maximun Entropy Method combined with projection analysis. Antani et al. [31] use the binarization method proposed by Kamel and Zhao [32]. Lienhart and Wernike use a global binarization method based on text and background color estimation for each bouding box [22]. Wu, Manmatha et al. [33] calculate the threshold from text chips while Chen et al. [34] propose asymmetric Gabor filters for character stroke enhancement. Chen et al. [35] propose also a Markov Random Field (MRF) based on 2x2 cliques. Finally, many researchers use classic binarization algorithms like Otsu [36], Ohya [37], Niblack [38] and Sauvola [39].

The evaluation of a text detection system is an aspect not as trivial as it might seem. It resembles the generic problem of object detection evaluation having additionally its own issues. Most researchers use for their experimentation simple boxed-based or area-based methods while very few works have focused on the specific problem of evaluation. In [40] and [41] Kasturi, Manohar et al. propose as overall measure of text detection in a frame, a box-based measure called Frame Detection Accuracy (FDA). The evaluation methods of this kind are based on the mapping between ground truth and detected objects. Especially for the text detection problem, text lines are considered to be the objects where a text line is usually defined as an aligned series of characters with a small intermediate distance relative to their height. However, this subjectively small distance can result arbitrarily in bounding box splits or merges among annotators and detectors making the object mapping inappropriate. In addition, the number of correctly retrieved boxes is not generally a measure of the retrieved textual information since the number of characters in different boxes may vary considerably. Wolf et al. [42] proposed the creation of match score matrices with the overlap between every possible pair of blocks, in order to evaluate document structure extraction algorithms. The benefit of this kind of algorithms is their ability to consider the possible splits or merges of the bounding boxes besides one-to-one matching. However, in order to match two ground truth boxes with one resulting box, the total overlap threshold has to be very low (~40%). This will have as a result accepting as correct, a box with size even higher than the double size of the ground truth box. Many researchers have used the overall overlap to compute area-based recall and precision measures ([5], [13]). However the main drawback here similarly to the box-based approaches is the fact that the number of retrieved pixels does not correspond to proportional textual information since different textlines may contain characters of various sizes.

3 Spatial Text Detection

In the framework of this thesis we proposed three algorithms for text spatial detection - localization in images or/and video frames. Firstly, we proposed an edge-based heuristic method for the localization of artificial text in video frames. The second text detection system we developed is a two-stage scheme that uses the previous method as a first stage while a second machine, learning texture-based stage follows refining the initial result. The third detector we proposed uses directly machine learning techniques to distinguish text from background using texture-like features and is capable of detecting both artificial and scene text in images and video frames.

3.1 Edge-based heuristic method

The proposed algorithm [43] (Fig. 2) exploits the fact that text lines produce strong vertical edges horizontally aligned and follow specific shape restrictions. Using edges as the prominent feature of our system gives us the opportunity to detect characters with different fonts and colors since every character present strong edges, despite its font or color, in order to be readable. First, an edge map is created using the Canny edge detector [44]. Then, morphological dilation and opening are used in order to connect the vertical edges and eliminate false alarms. Bounding boxes are determined for every non-zero valued connected component, consisting the initial candidate text areas. Finally, an edge projection analysis is applied, refining the result and splitting text areas in text lines. The whole algorithm is applied in different resolutions to ensure text detection with size variability. Experimental results prove that the method is highly effective and efficient for artificial text detection in video frames with relative smooth background.

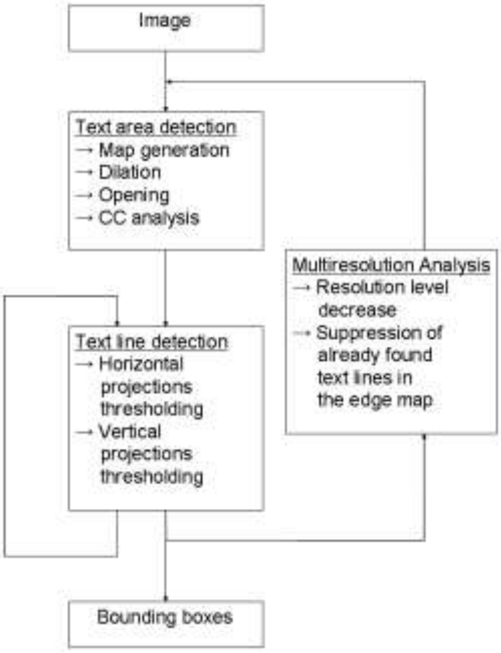


Image
Text area detection
→ Map generation
→ Dilation
→ Opening
→ CC analysis
Multiresolution Analysis
→ Resolution level decrease
→ Suppression of already found text lines in the edge map
Text line detection
→ Horizontal projections thresholding
→ Vertical projections thresholding
Bounding boxes

Figure 2. Flowchart of the proposed edge-based heuristic method for artificial text detection

3.2 Two-stage text detection scheme

This method consists of an initial coarse stage and a machine learning refinement stage [45]. In the first stage, text lines are detected based on the algorithm presented in section 3.1, leading in a high recall rate with low computational time expenses. In the second stage (Fig. 3), 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.

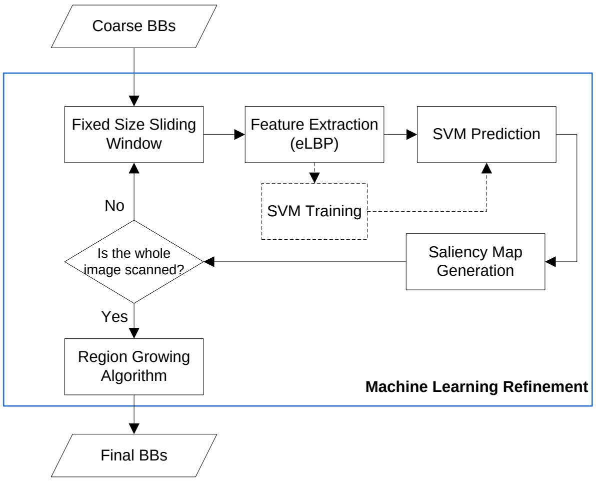

Coarse BBs
Fixed Size Sliding Window
Feature Extraction (eLBP)
SVM Prediction
SVM Training
No
Is the whole image scanned?
Saliency Map Generation
Yes
Region Growing Algorithm
Machine Learning Refinement
Final BBs

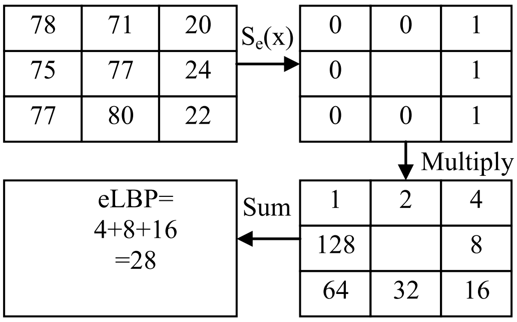
Figure 3. Flowchart of the proposed machine learning refinement stage

The feature set consists of the histogram values of the eLBP map which corresponds to each image area. In eLBP, a neighbouring pixel is represented by 0 if it is close to the center pixel or 1 if not. In order to define closeness, we require a minimum absolute distance *e* from the center to give to the pixel the binary value 1 (Figure 10).
Formally, the new eLBP operator is defined as:

$$eLBP(x_c, y_c) = \sum_{n=0}^{7} S_e(i_n - i_c)2^n \qquad (1)$$

where function $S_e(x)$ is defined by:

$$S_e(x) = \begin{cases} 1\,, & |x| \ge e \\ 0\,, & |x| < e \end{cases} \tag{2}$$



78	71	20
75	77	24
77	80	22

0	0	1
0		1
0	0	1

1	2	4
128		8
64	32	16

Figure 4. Example of eLBP computation

The value of *e* has to be large enough in order to avoid the arbitrary intensity variations caused by noise and small enough to detect all the deterministic intensity changes of texture. In [46] a value near 20 was proved to be satisfactory, after relative experimentation. Although this value was optimal for the discrimination of most text and non-text patterns there were still problems. Some text patterns of low contrast images presented edges that did not exceed the specified threshold and thus, were classified as non-text while non-text patterns of high contrast images presented strong enough edges to be classified as text. In order to solve this problem we propose the generation of multilevel eLBP edge histograms with different values for *e,* which will describe the edge distribution in different detail levels.These values for *e* are given by the quantile function of the exponential PDF which is denoted as:

$$e(i) = -M * \ln(1 - i / (L+1)) \tag{3}$$

where M is the mean value of the gradient image, i=1…L and L the number of different levels. For our experiments we used 8 number of levels (L=8) which resulted in 2048 features and forced us to do a feature reduction.

3.3 Machine learning method for artificial and scene text detection

In this work [47], we propose the use of a Random Forest within a sliding window model for the discrimination of text areas in the first stage, and then apply a gradient-based algorithm to achieve separation and refined localization of the text lines (Figure 2). This is actually a hybrid scheme combining an initial machine learning, texture-based technique with a heuristic, region-based refinement. The algorithm described above constitutes a fixed-scale detector so it is applied in multiple resolutions in order to detect text of any size. After processing all the resolutions needed, the final text line bounding boxes have been computed. Then, text lines are binarized and optionally segmented into words based on the distances between the resulting connected components. This segmentation is applied in cases where the text detection targets words instead of text lines.

The main contributions of this work is the choice of the classifier which provides efficiency and generalization capabilities, together with an improved, highly discriminative feature set that was designed particularly for reflecting the textual characteristics. Moreover, the use of a machine learning architecture for the first and most crucial stage, produces a generic and robust system for the detection of artificial and scene text in camera-based images and video frames. The description of the edge spatial distribution is done with the new MACeLBP (Multilevel Adaptive edge Local Binary Patterns) operator which considers the valuable RGB color information while it adapts the contrast levels locally to each area of the image producing an actual parameter-free feature set. The use of the SVM in [45] forced us to invent a reduced version of the feature set in order to have an efficient system. This fact resulted in loss of information for the description of the textual texture. Contrarily, in this work the use of a Random Forest and its capability to deal efficiently with high dimensional feature spaces allowed us to use the whole proposed feature set instead of a reduced version so all available information is exploited for a better description of texture. Finally, the impressive efficiency of the Random Forest gave us the capability to use it as the first and basic stage for scanning the whole image and detecting text instead of just refining the coarse results of a heuristic and unreliable stage.

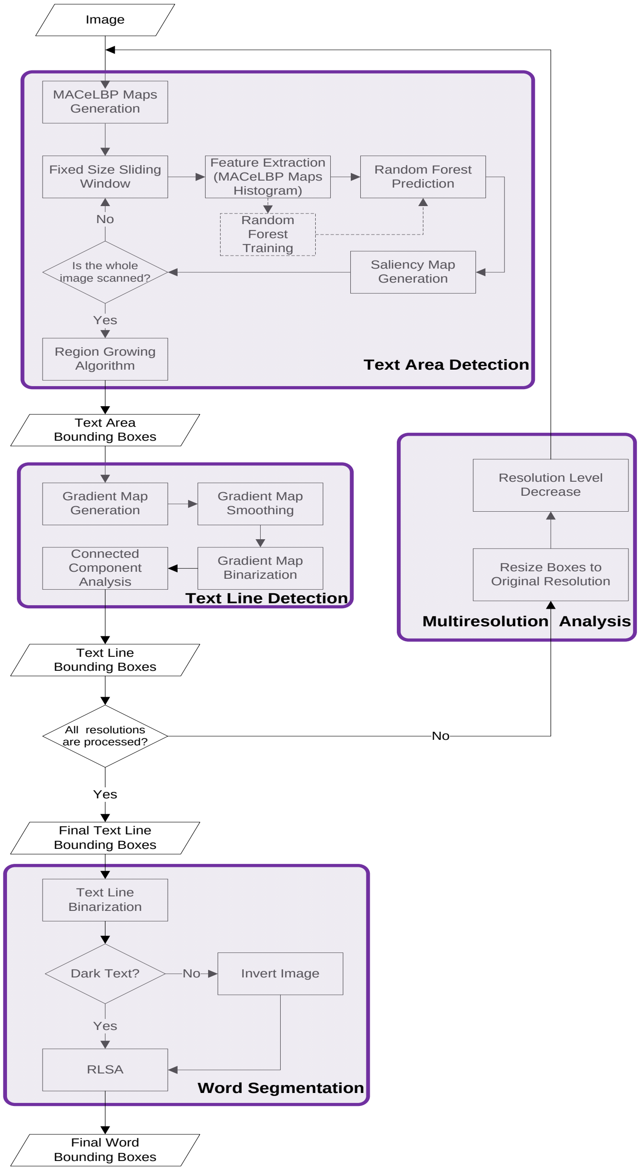


Figure 5. Flowchart of the proposed machine learning text detection algorithm

4 Temporal Text Detection

In order to obtain textual information from video, text detection and recognition in a single frame basis is not adequate. There is a giant amount of temporal information which has to be exploited. Every text line has to remain in the video for at least 2 seconds to be readable, which means at least 50 frames. Using this extra information we can remove the noise from the image and also smooth the background. Artificial text in television captured videos is usually static. Making this assumption we have the opportunity to use temporal information not only for enhancing the already detected text image but to improve text detection as well. Video frames are firstly averaged temporally and then fed to the text detection algorithm. Figure 6 presents an example of temporal text detection.

The algorithm mainly consists of three steps:

1. Temporal averaging of video (20 frames depth).
2. Text detection in the averaged video every 5 frames.
3. Matching of the detected text from frame to frame if overlap between the two boxes is over 80%.

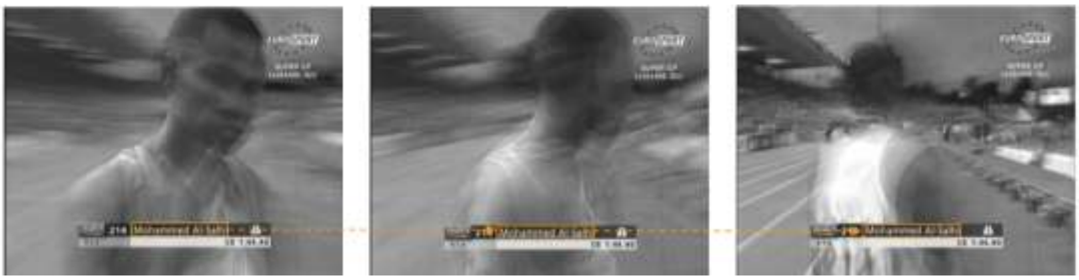


Figure 6. Example of temporal text detection

5 Text Image Binarization

The proposed binarization method consists of an Otsu thresholding [36] followed by a normal/inverse text detection. In order to classify between normal or inverse text we firstly apply a connected component analysis. The numbers of white (*WCC*) and black (*BCC*) connected components are counted, discarding components with height less than 8 pixels or less than the 40% of the box height. If |*WCC*-*BCC*|>1 then the color that corresponds to the largest number of connected components is regarded as text color. Else if the distance between *WCC* and *BCC* is less or equal to 1, the condition for the inversion is based on the pixel values of the borders of the bounding boxes. If the majority of border pixels are black then text is considered inverse. Finally, color inversion is applied for every text detected to be inverse so the final result will be normal binary text.

6 Text Detection Evaluation

6.1 Evaluation based on estimated number of characters

Ideally a text detection method as a part of a text extraction system should not be evaluated on the size of detected areas nor the number of detected boxes but on the number of the detected characters. Unfortunately, the number of characters in a bounding box cannot be defined by the algorithm but it can be approximated by the ratio width/height of the box, if we assume that this ratio is invariable for every character, the spaces between different words in a text line are proportional to its height and each textline contains characters of the same size.

In that way, the evaluation will be based on the recall and precision of the area coverage, normalised by the approximation of the number of characters for every box [45]. The overall metric will be the weighted harmonic mean of precision and recall also referred as the F-measure.

$$\text{Recall}_{\text{ecn}} = \frac{\sum_{i=1}^{N} \frac{|\text{GDI}_i|}{\text{hg}_i^2}}{\sum_{i=1}^{N} \frac{|\text{GB}_i|}{\text{hg}_i^2}} \qquad \text{Precision}_{\text{ecn}} = \frac{\sum_{i=1}^{M} \frac{|\text{DGI}_i|}{\text{hd}_i^2}}{\sum_{i=1}^{M} \frac{|\text{DB}_i|}{\text{hd}_i^2}} \qquad \text{F}_{\text{ecn}} = \frac{2*\text{Precision}_{\text{ecn}}*\text{Recall}_{\text{ecn}}}{\text{Precision}_{\text{ecn}}+\text{Recall}_{\text{ecn}}}$$

where GB_i is the ground truth bounding box number i and hg_i is its height, while DB_i is the detected bounding box number i and hd_i is its height. N is the number of ground truth bounding boxes and M is the number of detected bounding boxes and GDI, DGI are the corresponding intersections:

$$\text{GDI}_i = \text{GB}_i \cap \left(\bigcup_{i=1}^{M} \text{DB}_i \right) \qquad \text{DGI}_i = \text{DB}_i \cap \left(\bigcup_{i=1}^{N} \text{GB}_i \right)$$

6.1 Evaluation based exclusively on character pixels

The proposed algorithm [48] (Fig.7) generates two binarized images for the ground truth and the resulted bounding boxes respectively. For each case, the algorithm takes as input the image and a set of bounding boxes. The pixels contained in each box are binarized using Otsu thresholding [36] and then conditionally inverted producing black pixels for text and white pixels for the background. All pixels outside boxes are set to white. Otsu's method proved to be a very good solution for this kind of text images although the choice of the binarization method does not affect considerably the result of the proposed evaluation because of the skeletonizing that follows.

Then, the two binarized images have to be compared in order to compute the recall and precision rates. However, a straight comparison between the two images would produce evaluation rates strongly depended to the used binarization method. To overcome this problem and focus only on the evaluation of text detection we use the skeleton of each image computed by an iterative skeletonization method presented in [49]. Specifically, the recall and precision rates are defined by the equations:

$$\text{Recall} = \frac{|\text{GT\_skel} \cap \text{RS\_bin}|}{|\text{GT\_skel}|} \qquad \text{Precision} = \frac{|\text{RS\_skel} \cap \text{GT\_bin}|}{|\text{RS\_skel}|}$$

where GT_bin is the binarized ground truth image and GT_skel is its skeleton while RS_bin is the binarized image of the detection result and RS_skel is its skeleton. The operator |…| denotes the number of text (black) pixels and ∩ is the intersection of the text (black) pixels.

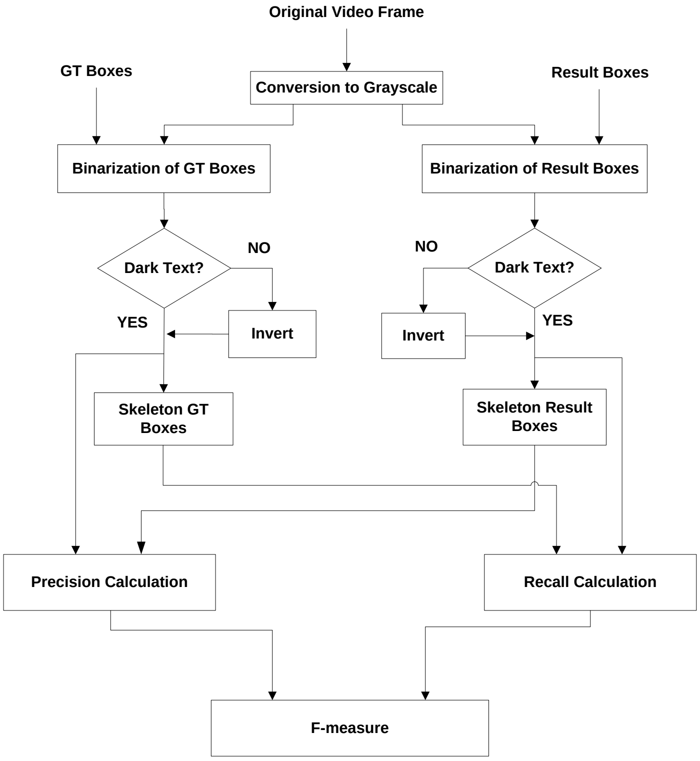


Figure 7. Flowchart of the proposed algorithm

7 Experimental results

For the experiments regarding the text detection methods we created two datasets. The first one consists of 214 frames from athletic videos while the second contains 172 video frames from news broadcasts. Table 1 shows comparative results for the two proposed text detection methods [45, 47] against three state-of-the-art algorithms. The used evaluation protocol is the one presented in section 6.1. Table 2 presents the average processing times for the same methods. Table 3 present the results of the proposed method [47] for the dataset of the competition Robust Reading of ICDAR 2003 and 2005 compared to the other methods of the competition. These experiments combined with the ones of table 1, proved the capability of the proposed method to deal with both artificial and scene text in camera-based images or video frames. Table 4 provides a feature set comparison in terms of classification accuracy combined with two classifiers, SVM and Random Forest. According to these results the best performance is achieved by reduced eLBP feature set combined with an SVM classifier as proposed in [45]. However, the difference in terms of performance between SVM and RF best results is negligible compared to the corresponding computational cost as it is shown in table 5. This exceptional efficiency of Random Forest allowed us to create a pure machine learning system [47] which outperformed any other method in literature.

Table 1. Comparative results for artificial text detection algorithms

%	**Method**	**Recall$_{ecn}$**	**Precision$_{ecn}$**	**F$_{ecn}$**
Dataset 4	[20]	66.9	66.7	66.8
	[13]	65.4	75.6	70.1
	[23]	81.1	70.5	75.4
	Proposed [45]	83.9	79	81.4
	Proposed [47]	84.1	79.8	81.9
Dataset 5	[20]	63.3	69.2	66.1
	[13]	68.2	71.1	69.6
	[23]	80.6	71.5	75.8
	Proposed [45]	82.7	83.5	83
	Proposed [47]	83.4	84.7	84

Table 2. Average processing time per frame for *Dataset 4* and *Dataset 5*

Method	**Average Processing time per frame (secs)**
[20]	8
[13]	3.35
[23]	1.5
Proposed [45]	2
Proposed [47]	1.6

Table 3. Comparative results for ICDAR2003 dataset and evaluation protocol

	Method	***p***	***r***	***f***	**T(sec)**
	Proposed [47]	0.82	0.61	0.70	4.2
	Ji [19]	0.59	0.79	0.68	-
	Epshtein [8]	0.73	0.60	0.66	0.94
Robust Reading Competition 2005	**Hinnerk Becker**	0.62	0.67	0.62	14.4
	Alex Chen	0.60	0.60	0.58	0.35
	Qiang Zhu	0.33	0.40	0.33	1.6
	Jisoo Kim	0.22	0.28	0.22	2.2
	Nobuo Ezaki	0.18	0.36	0.22	2.8
Robust Reading Competition 2003	**Ashida**	0.55	0.46	0.50	8.7
	HWDavid	0.44	0.46	0.45	0.3
	Wolf	0.30	0.44	0.35	17
	Todoran	0.19	0.18	0.18	0.3

Table 4. Classification results of *Dataset 1* for different features and classifiers

	Features	**Feature dimension**	**Recall$_{text}$**	**Precision$_{text}$**	**F$_{text}$**
RF	MACeLBP	2048	96.6	97.4	**97**
	Reduced eLBP	256	93.5	94.6	**94.1**
	DCT	287	94.2	91.1	**92.7**
	Haar	279	91.4	90	**90.1**
	Gradient	288	90.5	88.2	**89.3**
SVM	Reduced eLBP	256	98	98	**98**
	MACeLBP	2048	96	98.2	**97.1**
	DCT	287	94.2	96.2	**95.2**
	Haar	279	93.1	95.7	**94.4**
	Gradient	288	80.1	92.3	**86.3**

Table 5. Comparing SVM and RF in terms of prediction speed.

(Predictions/sec)	**SVM**	**RF**
MACeLBP	40	~150.000
Reduced eLBP	160	~500.000
DCT	180	~500.000
Haar	385	~500.000
Gradient	200	~500.000

8 Conclusion

In this thesis, novel methodologies for the extraction of textual information from images and videos have been proposed. Three different algorithms for text detection were developed covering all cases of text in all kinds of media. The first is a very efficient edge-based heuristic algorithm for the detection of artificial text in videos. The second one is a two-stage, coarse to fine algorithm which uses the previous approach as a first step while also integrates a second machine learning stage that refines the result. The success of this stage is based on a very powerful feature set, in terms of discrimination which is produced by a new operator called edge Local Binary Pattern (eLBP). The eLBP feature set describes the distribution of local edge patterns which actually distinguishes text from the background. The third proposed text detection method uses directly machine learning to locate text. The feature set is based on Multilevel Adaptive Color eLBP (MACeLBP) which constitutes an evolved eLBP operator that includes additionally color edge information and operates edge multilevel thresholding in a locally adaptive way. The use of the very efficient Random Forest classifier gave us the capability to scan with a sliding window the whole image in several multiresolution levels and develop a completely parameter-free machine learning system which is able to detect both artificial and scene text in camera-based images or video frames. Two text detection evaluation methods were also proposed. Both of them are based on intuitively correct criteria and try to measure objectively the percentage of the retrieved textual information. The first evaluation protocol computes recall and precision rates based on the estimated character number of each text bounding box. The second algorithm passes the evaluation process to the next stage of text binarization, computing the retrieval rates based exclusively on character pixels. In that way we overcome all the problems caused by the ambiguously defined text bounding boxes. In this framework we developed a text pixel segmentation system which binarizes the image and then detects and inverts cases of inverse text with a very high accuracy. Finally, a very efficient method was developed for the temporal detection of text which actually conduces to a better spatial detection while concurrently enhances the quality of the text image.

References

1. Lienhart R and Effelsberg W (2000) Automatic Text Segmentation and Text Recognition for Video Indexing. ACM/Springer Multimedia Systems, Vol. 8. pp.69-81
2. Sobottka K , Bunke H, Kronenberg H (1999) Identification of Text on Colored Book and Journal Covers. International Conference on Document Analysis and Recognition, pp. 57–63.
3. Wang K, Kangas J.A (2003) Character Location in Scene Images from Digital Camera. Pattern Recognition, Volume 36, Number 10, pp. 2287-2299(13)
4. Sato T, Kanade T, Hughes E, and Smith M (1998) Video OCR for Digital News Archives, IEEE Workshop on Content-Based Access of Image and Video Databases , pp. 52 – 60
5. Kim W, Kim C (2009) A New Approach for Overlay Text Detection and Extraction from Complex Video Scene. IEEE Transactions on Image Processing, vol.18, no.2, pp.401-411
6. Chen X, Yang J, Zhang J, Waibel A (2004) Automatic Detection and Recognition of Signs from Natural Scenes, IEEE Transactions on Image Processing, Vol. 13, No. 1. pp. 87-99.
7. Epshtein B, Ofek E, Wexler Y (2010) Detecting Text in Natural Scenes with Stroke Width Transform, IEEE Conference on Computer Vision and Pattern Recognition, San Francisco.
8. Zhong Y, Zhang H and Jain A.K (2000) Automatic Caption Localization in Compressed Video. IEEE Trans. on Pattern Analysis and Machine Intelligence, 22(4): pp.385–392.
9. Crandall D, Antani S, Kasturi R (2003) Extraction of Special Effects Caption Text Events from Digital Video. International Journal on Document Analysis and Recognition (5), No. 2-3, pp. 138-157
10. Lim Y.K, Choi S.H, and Lee S.W (2000) Text Extraction in MPEG Compressed Video for Content-based Indexing. International Conference on Pattern Recognition, pp. 409-412.
11. Gargi U, Crandall D.J, Antani S, Gandhi T, Keener R, Kasturi R (1999) A System for Automatic Text Detection in Video. International Conference on Document Analysis and Recognition, pp. 29-32
12. Goto H (2008) Redefining the DCT-based feature for scene text detection: Analysis and comparison of spatial frequency-based features. International Journal on Document Analysis and Recognition (11), No. 1, October 2008, pp. 1-8.
13. Chen D, Odobez J-M and Thiran J-P (2004) A Localization/Verification Scheme for Finding Text in Images and Videos Based on Contrast Independent Features and Machine Learning Methods. Image Communication, vol. 19(3), pp. 205-217
14. Ye Q, Huang Q, Gao W, Zhao D (2005) Fast and Robust Text Detection in Images and Video Frames. Image Vision Computing, 23(6): pp.565-576.
15. Jung C, Liu Q, Kim J (2009) A Stroke Filter and its Application to Text Localization. Pattern Recognition Letters, 30(2): pp. 114-122.
16. Ye Q, Jiao J, Huang J, Yu H (2007) Text detection and restoration in natural scene images, Journal of Visual Communication and Image Representation 18(6), pp. 504-513.
17. Ji R, Xu P, Yao H, Zhang Z, Sun X, Liu T (2008) Directional correlation analysis of local Haar binary pattern for text detection. IEEE International Conference on Multimedia & Expo, pp.885-888.
18. A. Ekin. Information Based Overlaid Text Detection by Classifier Fusion. IEEE International Conference on Acoustics, Speech and Signal Processing, (2006), pp. II-753-756
19. Jung K (2001) Neural Network-based Text Location in Color Images. Pattern Recognition Letters , 22(14): pp. 1503–1515.
20. Kim K.I, Jung K, Park S.H and Kim H.J (2001) Support Vector Machine-based Text Detection in Digital Video. Pattern Recognition, 34(2): pp. 527–529.
21. Wolf C and Jolion J-M (2004) Model Based Text Detection in Images and Videos: a Learning Approach. Technical Report LIRIS-RR-2004-13 Laboratoire d'Informatique en Images et Systemes d'Information, INSA de Lyon, France.
22. Lienhart R and Wernicke A (2002) Localizing and Segmenting Text in Images and Videos. IEEE Trans. on Circuits and Systems for Video Technology, 12(4): pp.256–268.
23. Li H, Doermann D and Kia O (2000) Automatic Text Detection and Tracking in Digital Video, IEEE Transactions on Image Processing. Vol. 9, No. 1, pp. 147-156.
24. Rainer Lienhart and Frank Stuber, (1995) "Automatic text recognition in digital videos", Technical Report / Department for Mathematics and Computer Science, University of Mannheim ; TR-1995-036.
25. S. Antani, U. Gargi, D. Crandall, T. Gandhi, and R. Kasturi, "Extraction of Text in Video", Technical Report of Department of Computer Science and Engineering, Penn. State University, CSE-99-016, August 30, 1999.
26. Y. Nakajima, A. Yoneyama, H. Yanagihara, and M. Sugano, "Moving Object Detection from MPEG Coded Data", Proc. of SPIE, 1998, Vol. 3309, pp.988-996.
27. M. Pilu, On Using Raw MPEG Motion Vectors to Determine Global Camera Motion, Proc. of SPIE, 1998, Vol. 3309, pp. 448-459.
28. H. Li and D. Doermann. Text Enhancement In Digital Video Using Multiple Frame Integration. Proceedings of ACM Multimedia 99 , pages 19-22.
29. C.M. Lee and A. Kankanhalli. Automatic extraction of characters in complex scene images. International Journal of Pattern Recognition and Artificial Intelligence, 9(1):67-82, 1995.
30. F. LeBourgeois. Robust Multifont OCR System from Gray Level Images. In Proceedings of the 4th International Conference on Document Analysis and Recognition, pages 1-5, 1997.

31. S. Antani, D. Crandall, and R. Kasturi. Robust Extraction of Text in Video. In Proceedings of the International Conference on Pattern Recognition, volume 1, pages 831-834, 2000.
32. M. Kamel and A. Zhao. Extraction of Binary Character/Graphics Images from Grayscale Document Images. Computer Vision, Graphics, and Image Processing, 55(3):203-217, May 1993.
33. V. Wu, R. Manmatha, E.M. Riseman. Textfinder: An Automatic System to Detect and Recognize Text in Images. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 21, Issue 11, pp. 1224-1229, Nov. 1999.
34. D. Chen, K. Shearer, and H. Bourlard. Text enhancement with asymmetric filter for video OCR. In Proceedings of the 11th International Conference on Image Analysis and Processing, pages 192-197, 2001.
35. D. Chen, J.M. Odobez, and H. Bourlard. Text segmentation and recognition in complex background based on markov random field. In Proceedings of the International Conference on Computer Vision and Pattern Recognition, volume 4, pages 227-230, 2002.
36. Otsu N (1979) A Threshold Selection Method from Gray-Level Histograms. IEEE Transactions on Systems, Man and Cybernetics Vol. 9, No. 1, pp. 62-66.
37. Jun Ohya, Akio Shio, and Shigeru Akamatsu. Recognizing Characters in Scene Images. IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol. 16, No. 2, Febr. 1994.
38. W. Niblack. An Introduction to Digital Image Processing, pages 115-116. Englewood Cliffs, N.J.: Prentice Hall, 1986.
39. J. Sauvola, T. Seppanen, S. Haapakoski, and M. Pietikainen. Adaptive Document Binarization. In International Conference on Document Analysis and Recognition, volume 1, pages 147-152, 1997.
40 V. Manohar, P. Soundararajan, M. Boonstra, H. Raju, D. Goldgof, R. Kasturi, J. Garofolo. Performance Evaluation of Text Detection and Tracking in Video. International Workshop on Document Analysis Systems, pp. 576-587
41 R. Kasturi, D. Goldgof , P. Soundararajan, V. Manohar , J. Garofolo , R. Bowers , M. Boonstra , V. Korzhova , J. Zhang. Framework for Performance Evaluation of Face, Text, and Vehicle Detection and Tracking in Video: Data, Metrics, and Protocol, IEEE Transactions on Pattern Analysis and Machine Intelligence (2008), 31 (2) pp. 319-336.
42 C. Wolf, J. Jolion. Object Count/area Graphs for the Evaluation of Object Detection and Segmentation Algorithms. International Journal on Document Analysis and Recognition (2006), 8(4) pp. 280-296.
43 M. Anthimopoulos, B. Gatos and I. Pratikakis, "Multiresolution Text Detection in Video Frames", 2nd International Conference on Computer Vision Theory and Applications (VISAPP 2007), pp. 161-166, Barcelona, Spain, March 2007
44 Canny J., 1986. A computational approach to edge detection, IEEE Trans. Pattern Analysis and Machine Intelligence, 8, 679-698.
45 M. Anthimopoulos, B. Gatos, I. Pratikakis, «A two-stage scheme for text detection in video images», Image and Vision Computing, Vol. 28 , Issue 9, pp. 1413-1426, 2010.
46 M. Anthimopoulos, B. Gatos, and I. Pratikakis, "A Hybrid System for Text Detection in Video Frames", 8th International Workshop on Document Analysis Systems (DAS'08), pp. 286-292, Nara, Japan, September 2008.
47 M. Anthimopoulos, B. Gatos, I. Pratikakis, “Detection of Artificial and Scene Text in Images and Video Frames”, Pattern Analysis and Applications. Accepted for publication.
48 M.Anthimopoulos, N.Vlissidis, B.Gatos, “A Pixel-Based Evaluation Method for Text Detection in Color Images” The 20th International Conference on Pattern Recognition
49. H. J. Lee, B. Chen, "Recognition of Handwritten Chinese Characters via Short Line Segments", Pattern Recognition (1992), 25 (5), pp. 543-552.