Automatic Text Extraction in Digital Video based on Motion Analysis

Duarte Palma, João Ascenso, Fernando Pereira

Instituto Superior Técnico – Instituto de Telecomunicações, 1049-001 Lisboa, Portugal
e-mail: {Duarte.Palma, Joao.Ascenso, Fernando.Pereira}@lx.it.pt

Abstract. It is well known that the text that appears in a video scene or is graphically added to it is an important source of semantic information for indexing and retrieval, notably in the context of video databases. This paper proposes an improved algorithm for the automatic extraction of text in digital video; its major strengths are its robustness in terms of text skew and its improved performance in dealing with scene text. The system is based on a segmentation approach, using geometrical and spatial analyses for text detection. After, temporal redundancy is exploited to improve the detection performance by means of motion analysis. The output of the text detection step is then directly passed to a standard OCR software package in order to obtain the detected text as ASCII characters.

1 Introduction

The technological advances seen in recent years in the area of audiovisual representation technology have led to a boom in the usage of audiovisual information, namely accessed through the Internet, by a growing number of users. The increasing amount of audiovisual information being deployed has led many relevant content players such as audiovisual content producers and television operators to show interest in creating digital libraries which should allow the efficient storage and indexing of audiovisual information for future management and retrieval. Nowadays, the task of annotation is typically performed manually, by a human operator; this process is very expensive, time consuming and many times suffers from the subjectivity associated to the human operator. To address the need and overcome the problem, it is necessary to develop systems capable of automatically processing audiovisual information, targeting its efficient indexing, storage, searching, transmission, and viewing. Thus the development of automatic and efficient systems able to analyze, describe, filter and retrieve audiovisual information is necessary. With this objective in mind, several methods and even international standards have been developed in the last few years. Much of this technology targets the textual information that exists in images and video sequences since this is a source of highly semantic information and thus, if available, would allow the filtering and searching of audiovisual data by users in a more intuitive and natural way. The objective of this paper is to propose an automatic text extraction solution for digital video based on motion analysis. The most significant novelty of this proposal regarding the solutions already available in the literature is its capacity to deal both with graphical and scene text, oriented in any direction, with a high performance, notably a small number of false detections. Graphical text is the text separately produced from the video shooting which is overlaid on the scene in a post-processing stage; scene text appears as part of the scene and is recorded together with the scene. Both types of text may be composed by characters with various sizes, fonts, colors and can appear in different directions; of course, scene text is typically more varied.

There have been several approaches proposed in the last few years for the automatic extraction of text in digital videos. Existing text extraction methods can be classified into three main categories: methods based on segmentation, on edge detection and on supervised learning. The methods based on the segmentation of the visual data treat the text as regions with special characteristics, e.g. in terms of texture, size, shape and alignment constraints. For example in [1], color is used to form homogeneous regions; after, heuristic methods are used to detect characters in these homogeneous regions. The methods based on edge detection exploit the typical high concentration of edges in characters. For exa mple in [2], the problem of video indexing is addressed assuming that text consists in strokes with high contrast; thus the method searches for vertical edges which are grouped into rectangles. The methods based on supervised learning typically use neural networks for text detection. For example in [3], a hybrid wavelet/neural network is used to locate text in videos; the high-frequency coefficients are the input to the network. In [4], text lines are identified by using a complex-valued multi-layer feed-forward network trained to detect text with a fixed scale. The output of the neural network for all scales and positions is integrated into a single text-map, serving as a starting point for the detection of candidate text lines. Many of these methods show a lower detection performance when dealing with scene text and non-horizontal text.

This paper is structured as follows: Section 2 presents the automatic text detection algorithm here proposed. Section 3 shows the experimental results and thus the performance of the proposed text detection solution for various types of text. Finally, Section 4 concludes the paper and refers to future work.

2 A Proposal for the Automatic Extraction of Text in Video

This paper intends to propose an improved algorithm for the extraction of text in digital video considering both graphical and scene text. The overall architecture for the text extraction solution proposed is shown in Fig. 1 and considers, as usual, two major phases: text detection and text recognition.

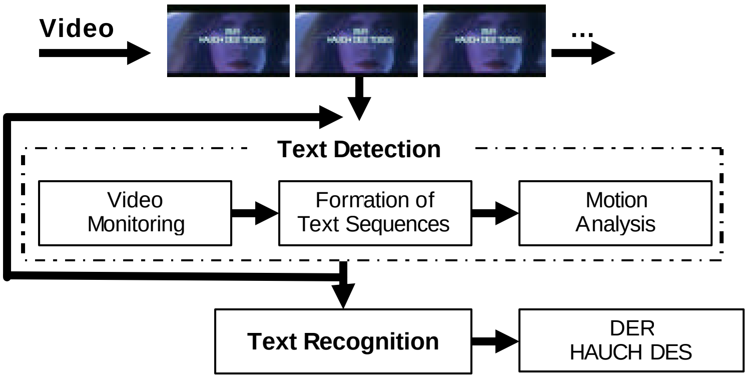


Fig. 1. Architecture of the proposed system.

2.1 Text Detection

The principal objective of the text detection phase is to identify a set of regions classified as text candidate regions. The text detection is carried out in three steps [2, 3, 4]: video monitoring, formation of text sequences and motion analysis.

2.1.1 Video Monitoring

The main purpose of the video monitoring module is to detect text in a video sequence. In order to reduce the overall computational complexity, text detection is first performed using a low temporal resolution of the video, this means only a few (periodic) frames are processed out of the full set. It is known from vision research that humans need between 2 and 3 seconds to process a complex scene [4, 5]. Therefore, it is safe to assume that in order for humans to recognize text, it needs to be visible for at least about one second; for a 25 fps video, this means that human text processing lasts for sure at least 25 frames (1 s) and thus it is enough to process one out of 25 frames in order to detect the presence of text. After the presence of text is detected at this low granularity, a finer detection phase starts targeting the frame precise detection of the text, see the formation of text sequences module.

The proposed video monitoring algorithm includes four steps: frame simplification, frame segmentation, character detection and word formation, detailed in the sequel:

1. *Frame simplification* – The purpose of this step is to decrease the influence of some unwanted effects such as noise or a too high number of colors. Frame simplification shall preserve as much as possible the original edges in each frame. So each frame is filtered with a filter combining edge detection with a median filter (3×3) [6]. The edge detector used is based on the Canny operator [7].
2. *Frame segmentation* – The purpose of frame segmentation is to split the image into homogeneous regions based on texture data (luminance). Each one of these regions has a certain probability of being a text character. The split and merge algorithm used in this paper was proposed by Cortez et al. [8] and is based on the hierarchical decomposition of each frame. However, when the frames have a significant amount of noise, this method results in over-segmentation, which degrades the quality of the characters edges. In order to eliminate this effect and improve the quality of the characters edges, it is necessary to eliminate small regions that share an edge with more than one region and that in some way contribute to the degradation of the characters boundaries. In order to identify these small regions for posterior merging and improve the characters edges, a technique presented by Lienhart [1] combining edge detection with dominant contrast local orientation is used: while edges are located using the Canny detector [7], the dominant local orientation is determined by the inertia tensor [9].
3. *Character detection* – In order to detect each text character, the regions resulting from the segmentation step are filtered based on adequate geometrical restrictions related to their height, width, height-to-width ratio and compactness *(=A_i/A_{bb})* where A_i, is the area of region *i*, and A_{bb}, is the area of the *bounding box* for region *i*. The thresholds used for each restriction depend on the range of character sizes selected for detection.

4. *Word formation* **–** To form words, it is considered that text consists of groups of characters aligned in a certain direction with characters sufficiently close to each other to form words. The word formation technique proposed here is based on the spatial analysis of each region previously classified as text. The following spatial criteria are used to form words:
   - ♦ *Proximity* – The regions shall be sufficiently close to each other;
   - ♦ *Alignment* – The regions shall be aligned in a certain direction;
   - ♦ *Height* – The regions shall have a minimum difference in height;
   - ♦ *Luminance* – The regions shall have similar luminance values;
   - ♦ *Dimension* – A word has to include at least three regions; smaller sets of related regions are discarded since they are meaningless.

The full description of the word formation process is described in [6]. Each analyzed frame is considered to contain text if at least one valid word exists, in any possible direction.

2.1.2 Formation of Text Sequences

The objective of the formation of text sequences module is to identify in the video sequence sets of contiguous frames where text exists, in order to apply later motion analysis. A text sequence is limited at the beginning and at the end by at least three contiguous frames without text. If the text monitoring phase detects some text in a frame, a finer detection of text starts for all the intermediate frames to determine the precise limits of the text sequence. This temporal finer detection is first performed backwards from the frame where the text has been initially detected and then forward in order to establish more precisely the first and the last frame where the text appears and the location of the text for each frame. This backward and forward detection ends when at least three contiguous frames without text are found in each direction. The detected interval represents a text sequence and the detected regions will be subject to motion analysis targeting the exploitation of temporal redundancy to increase the chance of detecting text and removing false alarms in individual frames since they are usually not stable throughout time. In order to detect text in the intermediate frames, the same algorithm used in the text monitoring phase is used.

2.1.3 Motion Analysis

In the motion analysis module, the correlation between frames is exploited to improve the results of the text detection phase. The major goals are to refine the text detection in terms of removing false alarms in individual frames, interpolating the location of accidentally missed text characters and words in individual frames, and temporally localizing the beginning and end of each word as well as its spatial location within each frame. The method proposed for motion analysis is based on the comparison of regions in successive frames and includes five steps:

1º Step – Text Tracking

This step is responsible for tracking the already detected text along the frames that constitute each text sequence targeting the formation of temporally related chains of characters. Each character chain represents the same character during its existence in the video sequence and consists in a collection of similar regions, occurring in several contiguous frames (although very likely with motion, e.g. scrolling). Every time a character region is detected for the first time, a position is stored and a signature is computed for that character region, using the following features: luminance, size and shape. Each frame contributes with one region classified as a character for the construction of a character chain.

The text tracking process creates the character chains through the following steps:

1. Each character region C_i, which belongs to frame n is compared with all the character chains CC_j, $j \in \{1,..., t\}$ that exist in frames n-8 to n-1. If the caracter C_i signature is similar enough to an already existing caracter chain signature and the caracter position is close to the estimated position for the already existing chain in the current frame, then that caracter is included in that caracter chain. The estimated position $\left(x_{CC_{j_k}}, y_{CC_{j_k}}\right)$ for the chain CC_j, in the frame F_k, is given by:

   $$\left(x_{CC_{j_k}}, y_{CC_{j_k}}\right) = \left(x_{CC_{j_{k-1}}}, y_{CC_{j_{k-1}}}\right) + (dx, dy) \tag{1}$$

   Where (dx, dy) is the medium displacement for CC_j, among two frames, given by:

   $$(dx, dy) = \left(\frac{x_{CC_{j_1}} - x_{CC_{j_{k-1}}}}{n}, \frac{y_{CC_{j_1}} - y_{CC_{j_{k-1}}}}{n}\right) \tag{2}$$

   Where n represents the duration of CC_j and $x_{CC_{j_1}}, y_{CC_{j_1}}, x_{CC_{j_{k-1}}}, y_{CC_{j_{k-1}}}$ represent the CC_j coordinates in the frame where it was created and in the frame before the current frame, respectively.
2. If caracter C_i is not similar to any existing character chain, a new character chain is created and initialized with the features/signature of that caracter.
3. After the processing of each frame, it is necessary to verify which character chains ended and which can still propagate to the next frame. A character chain is declared ended if it does not meet at least one of the following criteria:
   - Creation in frame n-1 but not present in frame n.
   - Not tracked in the last 0.32 seconds, i.e. 8 frames for a 25 fps.
   - Estimated position is outside the borders of the image.

   After a character chain is declared stopped, it is has to be globally validated by means of the following criteria:
   - Duration greater than 0.25 seconds, i.e. 6 frames for a 25 fps, and effective appearance in more than 3 frames;
   - Motion smaller than 250 pels per second, i.e. 10 pels/frame at 25 fps.

The thresholds for the above criteria were obtained empirically after extensive tests. Character chains classified as invalid are discarded in order to avoid false detections for individual frames. The output of the text tracking phase is a group of character chains, each one corresponding to a single character; they represent the temporal evolution of each character present in the video.

2º Step– Text integration

The text integration step is responsible for grouping the character chains in order to form words. The word formation technique proposed here is based on the spatial and temporal analysis of each character chain. Besides the spatial criteria presented in Section 2.1.1, the following temporal criteria are used to form words:

- *Temporary coexistence* – The character chains shall exist in about the same frames, more precisely in at least 75% of the smaller character chain duration;
- *Duration* – The character chains shall have a minimu m difference in duration;
- *Motion* – The character chains shall have a similar motion;

The chains not included in words at this phase are considered as noise and are discarded. The full description of the word formation process is described in [6].

3º Step– Character Recovery

The character recovery step explores the video temporal redundancy to complete the words with missing characters at least for some frames, e.g. due to noise or too textured background. In order to complete these words, they are extended to the size of their biggest chain of characters and the characters missing in the chains are recovered by means of temporal interpolation with motion compensation. Thus, by using temporal redundancy, the text detection for each frame is improved by 'comp leting' the words with missing characters for some frames.

4º Step– Elimination of Overlapped Words

This step is important to improve the performance for scene text; it performs the elimination of words which are overlapped with other words, i.e. words which bounding boxes overlap. Usually, overlapping of words occurs when false words are detected, e.g. due to shadows or three-dimensional text. Every time two or more words overlap, more precisely their bounding boxes overlap at least for one frame, the words with the smaller areas are discarded.

5º Step– Text Rotation

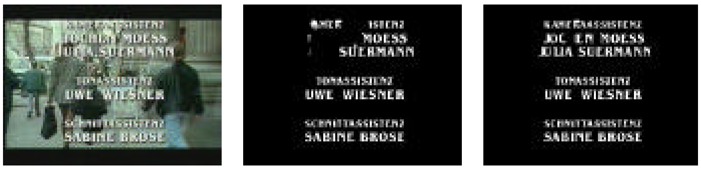
This step performs the rotation of the text to the horizontal position in order to be (better) recognized by the OCR system. Text rotation is performed in two steps:

1. *Calculation of the rotation angle* – The rotation angle is defined as the angle between the *xx* axis and the line that passes through the center of the regions where the word in question begins and ends. Considering the word $P = \{C_1, ..., C_n\}$ with bounding box of coordinates $\{(x_{C_1}, y_{C_1}), ..., (x_{Cn}, y_{Cn})\}$, its rotation angle is given by:

$$q_T = \tan^{-1}\left(\frac{b}{a}\right) \tag{3}$$

   Where $a = \max\{x_{C_1}, ..., x_{Cn}\} - \min\{x_{C_1}, ..., x_{Cn}\}$, $b = y_{\max\{x_{C_1}, ..., x_{Cn}\}} - y_{\min\{x_{C_1}, ..., x_{Cn}\}}$;
2. *Text rotation* – To rotate the text, the rotation angle ? computed is used along with the algorithm proposed by Alan Paeth [10].

Fig. 2 illustrates the effect of exploiting temporal redundancy for the extraction of text in video sequences. The result of extracting text from a single frame is shown in Fig. 2 (b) while Fig. 2 (c) shows the result of extracting text for the same frame but exploiting the temporal redundancy in the context of a video sequence.



(a) (b) (c)

Fig. 2. Example of the benefits of motion analysis for text extraction: (a) original image; (b) text detection for a single frame; (c) (better) text detection for the same frame but using temporal redundancy for recovering lost characters.

2.2 Text Recognition

The text bitmap produced using the words formed in the previous phase is recognized by standard OCR engines. In this paper, the OCR system used was OmniPage Pro 12.0.

3 Experimental Results

In order to evaluate the proposed automatic text extraction solution, a group of 13 video sequences with a total of 14.2 minutes and with plenty of text was used. These sequences where selected from several TV channels and were acquired using a Pinnacle Linx Video Input Cable, video capture card, with resolutions ranging from 352×208 pels to 384×288 (luminance) pels. The video sequences contain both scene and graphical text, aligned in any direction, with multiple fonts and sizes, in a total of 5852 characters. For example, the test video sequences were obtained from movie opening sequences and titles where graphical text is common, as well as TV commercials and news programs where scene text is common. The sequences with graphical text contain still as well as moving text (with scrolling motion).

3.1 Text Detection

Before processing each video sequence with the proposed text detection algorithm, a ground truth in terms of characters to be detected and recognized was determined. After, the correct detection or not of the ground truth characters has been determined by visual inspection of the images created by the text detection algorithm. For evaluating the text detection performance, the precision and recall metrics have been used: *Recall=CCD/CGT; Precision=CCD/TCD where CCD* represents the number of characters correctly detected by the algorithm, *CGT* represents the number of ground truth characters and *TCD* represents the total number of characters detected by the algorithm. The results in terms of detection for all the text defined as ground truth, i.e. horizontal, vertical and skewed text, for each type of text, can be observed in Table 1.

Table 1. Detection results for all the video sequences.

Type of text	Frames in test set	Characters in test set	Recall	Precision
Scene text	8319	907	0.879	0.836
Graphic text	12979	4926	0.953	0.958
Total text in the ground truth	**21298**	**5833**	**0.941**	**0.938**

The best detection results, both in terms of recall and precision, were obtained for the videos with graphical text Recall results of approximately 94% indicate that only 6% of the text that was considered relevant according to the defined conditions was not detected. Precision results of approximately 93% indicate that only 7% of the characters detected were falsely detected.

Fig. 3 illustrates the result of text extraction for a video sequence. The result of extracting text from the individual frames is shown in Fig. 3 (b) while Fig. 3 (c) shows the final integrated result for the full video sequence.

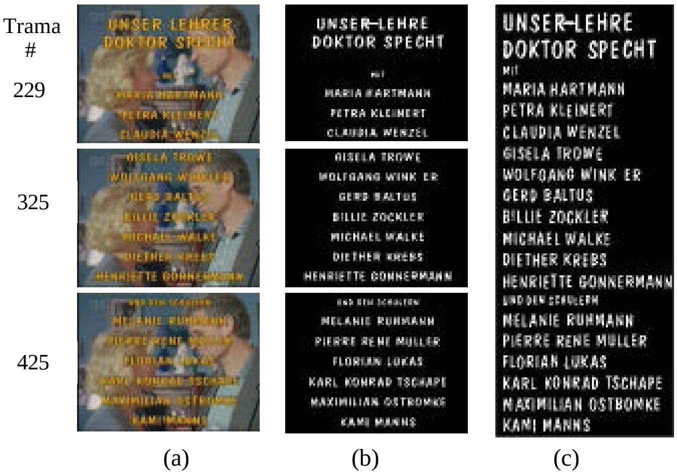


Fig. 3. Text extraction result for a video sequence: (a) video sequence; (b) text detection for individual frames (c) image resulting from the integration of the text in the full video sequence.

3.2 Text Recognition

For evaluating the text recognition performance, also precision and recall metrics have been used: *Recall=CCR/CGT; Precision=CCR/CR where CCR* represents the number of characters correctly recognized by the OCR system, *CGT* represents the number of ground truth characters and *CR* represents the number of characters output by the OCR system. The results in terms of recognition are presented in Table 2.

Table 2. Detection and recognition results for all the video sequences.

Type of text	**Recall**	**Precision**
Scene text	0.803	0.811
Graphic text	0.920	0.922
Total text in the ground truth	**0.902**	**0.905**

Similar to the detection results, the best recognition results were obtained for the videos with graphical text, both in terms of recall and precision. The characteristics of graphical text are the main justification for the different performance: normally, graphical text has a higher contrast and is better defined than scene text; on the other hand, scene text shows a larger number of fonts and sizes, making it more difficult to discriminate from other structures present in the scene.

4 Conclusions and Future Work

In this paper an improved automatic text extraction solution for digital video has been proposed. The major novelty of the solution is its improved performance in dealing with both types of text, oriented in any direction, with a small number of false detections. The performance of the proposed algorithm has been evaluated using various types of videos and text, notably title sequences of feature films, commercials and newscasts that contain both scene and graphical text, aligned in any direction with multiple fonts and sizes. Although it is difficult to compare the performance of the proposed algorithm with the few results available in the literature, notably because different sets of test material are used, it is possible to state that the present solution performs at least as well as the available state-of-the-art solutions and this for more varied text, notably in terms of scene text and text orientation. Our future research will concentrate on text with fewer constraints, e.g. text with random movements, characters may touch each other, and the font size may be smaller.

5 References

1. R. Lienhart e W. Effelsberg, “Automatic Text Segmentation and Text Recognition for Video Indexing”, Multimedia Systems, Vol. 8, Nº 1, (2000) 69 – 81.
2. T. Sato, T. Kanade, E. K. Hughes, M. A. Smith e S. Satoh, “Video OCR: Indexing Digital News Libraries by Recognition of Superimposed Captions”, Multimedia Systems, Vol. 7, Nº5, (1999) 385 – 395.
3. H. Li, D. Doermann e O. Kia, “Automatic Text Detection and Tracking in Digital Video”, IEEE Transactions on Image Processing, Vol. 1, Nº 1, (2000) 147 – 156.
4. R. Lienhart e A. Wernicke, “Localizing and Segmenting Text in Images and Videos”, IEEE Transactions on Circuits and Systems for Video Technology, Vol.12, Nº4, (2002) 256 – 268.
5. P. H. Lindsay e D. A. Norman, ”Introduction into Psychology – Human Information Reception and Processing”, Springer – Verlag (1991).
6. D. Palma, “Automatic Text Extraction in Digital Video Sequences”, Instituto Superior Técnico, Lisboa, Master Thesis, (2004).
7. J. Canny, “A Computational Approach to Edge Detection”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 8, Nº 6, (1986) 679 – 698.
8. D. Cortez, P. Nunes, M. Sequeira, F. Pereira, “Image Segmentation Towards New Image Representation Methods”, Signal Processing: Image Communication, Vol. 6, Nº 6, (1995) 485 – 498.
9. B. Jähne: Digital Image Processing, “Concepts, Algorithms, and Scientific Applications”, Springer (1997).
10. A. Paeth: A Fast Algorithm for General Raster Rotation. Graphics Interface, (1986) 77 – 81