Detection and Extraction of the Text in a video sequence

Hichem KARRAY and Adel ALIMI
REGIM: Research Group on Intelligent Machines, ENIS, University of Sfax, Tunisia
hichemkarray@yahoo.fr; adel.alimi@ieee.org

Abstract : **The extraction of the text in a video sequence has become controversial since the emergence of data processing and represents a revolution in the world of the multimedia. This article entitled "Detection and Extraction of the Text in a video sequence" consists achieving a system that extracts the inclusive text in video while relying on hypotheses and in trying to take advantage of the previous research this field.**

Key-words: Video source, Indexation of frames, Text Extraction, Segmentation.

I. INTRODUCTION

The video is a topic of paramount importance which continues to be dealt with directly as a basic (non-decomposable) object in multimedia documents. Its contents remain rarely clarified and it is often very difficult to classify and extract any knowledge.

In many applications, such as the indexing and research through content, we are asked to reach the internal structure of the video, and to lie out or handle data of finer granularity, such as the text or visual objects.

Classification and the annotation are usually carried the out manually according to a list of keywords chosen by user. This technique is tiresome and the automation of the indexing process is of great interest. The extraction of relevant information like the text can really provide us with additional data regarding the semantic contents of these videos. Nevertheless, the detection and the text recognition encounter several problems. If it is often relatively well contrasted in relation to its environment, the text can be found superimposed at a heterogamous and complex bottom. Moreover the text can be multicolored and heterogeneous. These characteristics make its extraction difficult.

The paper is organized as follows. Section 2 presents the characteristics image and text in video Section 3 presents the text detection, localization and tracking approach. Section 4 describes experimental results obtained for a set of video sequences. Section 5 concludes the paper and outlines areas for future research.

II. IMAGE AND TEXT CHARACTERISTICS IN VIDEO

A video sequence is a succession of images so to store a video on a computing support means to store a sequence of images which will have to be perfectly presented to the user at sufficient intervals (25 images per second).

- Image gray level: each, image element (pixel) represents the light intensity included 0-255 i.e. 256 gray level .

- Image color: The restitution of an image color is based on the additive synthesis of three primary colors: red , green and blue. All the other tones and nuances are obtained by the linear combinations of these three colors.

The system (R, V, B) is the representations system of the color in which the images are digitized. The values of the components (R, V, B) depend on the used acquisition system. When the system components are coded upon 8 bits, their values range from 0 and 255 and thus no coding to realize. On the other hand, it is possible to convert the system (R, V, B) corresponding to the acquisition device of another system.

One of the drawbacks of (R, V, B) system is the fact that three components are strongly correlated. Indeed, they have a strong reflectance factor distributed on each one of them [1], [2].

The texts included in images and the video orders represent a rich information source for contained applications based on indexing and recovery. Yet, the text characters are too difficult to be detected and identified due to their various sizes, gray level values and complex milieu.

Generally we distinguish two classes of texts: a scenic text and a graphic one.

The scenic text appears in a filmed scene. It forms an integral part of the image and can be regarded as an element of the real world. Elements like a traffic sign, a billboard, or a number plate are regarded as scenic texts. Generally this type of text is not very informative except when it intends to identify people or vehicles.

The graphic text is an element which is manually added in order to go along with the audio-visual support. Therefore; it is often structured and related with the global subject of the sequence. News headlines, time and place indications peoples names are examples of graphic texts.

1) Size

A text appears in a variety of sizes in video data. As the text is intended to be readable at a limited time, there is always a minimal size of text characters. However, the upper bound on character sizes is unwelcome.

The text can often be as large as half the frame height or more.

2) Color and Intensity

Color is a strong feature for use in visual information indexing. Text characters tend to have a perceptually uniform color and intensity over the character stroke.

While the character stroke appears to have the same color, in reality it is usually composed many different colors. In cases where the color varies across the caption, it does so in a gradual way so that adjacent characters or character segments have very similar colors.

3) Alignment

The characters in the caption text appear in clusters and usually lie horizontally, although sometimes they can appear as non-planar texts as a result of special effects. This does not apply to scenic texts, which can have various perspective distortions. Scenic texts can be aligned in any direction and can have geometric distortions.

4) Inter-character distance

Characters in a text line have a uniform distance between them.

5) Motion

The position of an artificial text can be changed, as in a scrolling text but does so in a very uniform way, either vertically or horizontally.

An important aspect is that the available text in video may be many different languages and scripts. For each script, the nature of text may differ. For instance, an Arabic script may have inter-character characteristics much different from English one.

.

III. TEXT EXTRACTION FROM A VIDEO

The extraction of a text includes four main tasks: detection, localization, follow-up and binarisation in order to send the results towards an OCR software.

1) State of the art of text detection

According to the various approaches described in the literature, we can mention three classes for detection regarding the used image: the color [3], [4], the texture [5], [6], and contours [7], [8].

However, the text characters are too difficult to be detected and identified to their various sizes, their gray level values and complex mediums

Various methods are used to tackle this problem. Among those methods, we can mention Jain and al [9] who proposed a method based on the uniformity of color. Lienhart and Al relied on the search for local contrasts. Soto, Kanade and Al [10] relied on the zone texts revealing a strong density of vertical contours, Wu, Manmatha and Riseman combined the research of vertical contours with a texture filter. LeBourgeois [11] relied on the accumulation of the horizontal gradients while relying on the statistical rules on the projection aspects of gray level. C.Wolf and J.M. Jolion[12] proposed an improvement of Niblak [13] and Sauvola's [14] methods in the level of threshlding criteria.Li and Doermann [15], Y. Hao and Z. Yi[16] used algorithms based on training by the use of a neuronal network.

In the field of compressed video, Zhang and Jain [17] relied on the DCT coefficients of MPEG glow to detect the text.

2) Our text extraction method

The approach that we propose relies on the recursive analysis of the component histograms of the color representation.

RVB components are strongly correlated .We have used space I1, I2, I3 which significantly improve the segmentation quality [18]. These three components are defined by the following equations:

I1=(R+V+B)/3, I2=(R-B)/2, I3= (2V-R-B)/4

We notice that the first one, which is also most discriminating, represents brightness. The two other ones represent respectively a blue-red contrast and a magenta-green contrast .To sum up we may regard the system (I1, I2, and I3) as a brightness-chrominance system.

2.1 Determination of the histograms of the color components

- Representation of the histograms of the components I1, I2, I3 for column 635

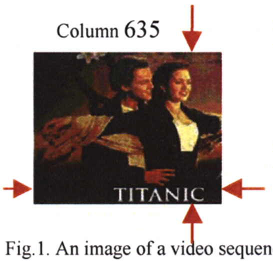

Fig.1. An image of a video sequence

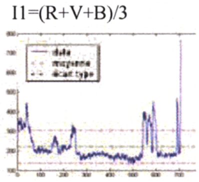

Fig.2. Histogram of the I1 component

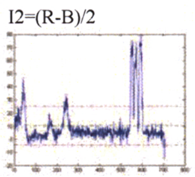

Fig.3. Histogram of the I2 component

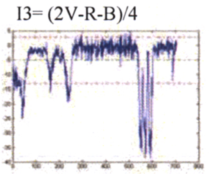

Fig.4. Histogram of the I3 component

- Representation of the histograms of the components I1, I2, I3 for line 650

Fig.5. An image of a video sequence

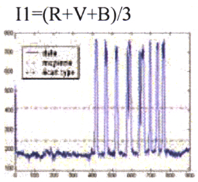

Fig.6. Histogram of the I1 component

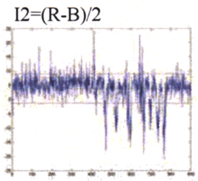

Fig.7. Histogram of the I2 component

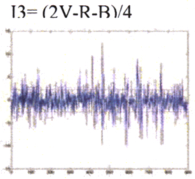

Fig.8. Histogram of the I3 component

2.2 Determination of the most representative color component

We are looking for the most representative color of the image content [19], and that which separates best the classes from the pixels.

According to expertise, we have noted the two components I2, I3 reveal less information than the third component I1. Thus we retain this color component as the most representative one.

2.3 Detecting and localizing texts

The idea is to draw lines across the image, then draw columns and try to analyze the variation of the curves color for each zone.

After the examination of curves, we can conclude that a pixel in an image belongs to a text if its value according to the I1 component is higher than the sum of the average with the standard deviation of the line or lower than the difference between the average and the standard deviation of this line (respectively for the columns).

This selection can be stated as:

-Analysis of image I by line: zone probably containing the text

$$\left| I_{(i,j)} - 1/n\sum_{k=1}^{n} I_{(i,k)} \right| > \sigma_i \text{ with } \sigma_i = (1/n\sum_{k=1}^{n} I_{(i,k)}{}^2 - 1/_n{}^2 (\sum_{k=1}^{n} I_{(i,k)})^2)^{1/2}$$

σi : represents the standard deviation on line I (1)

$1/n\sum I_{(i,k)}$: represents the average on line I

$I_{(i,j)}$: is the value of the pixel according to the component I1=(R+V+B)/3

-Analysis of image I by column: zone probably containing the text

$$\left| I_{(i,j)} - 1/n\sum_{k=1}^{n} I_{(k,j)} \right| > \sigma_j \text{ with } \sigma_j = (1/n\sum_{k=1}^{n} I_{(k,j)}{}^2 - 1/_n{}^2 (\sum_{k=1}^{n} I_{(k,j)})^2)^{1/2}$$

σ_j : represents the standard deviation on the column J (2)

$1/n\sum I_{(k,j)}$: represents the average on the column J

$I_{(i,j)}$: is the value of the pixel according to the component I1=(R+V+B)/3

-The intersection of the two analyses is a binary matrix $I'_{(i,j)}$

$$I'_{(i,j)} = \begin{cases} 1 & \left| I_{(i,j)} - 1/n\sum_{k=1}^{n} I_{(i,k)} \right| > \sigma_i \text{ and } \left| I_{(i,j)} - 1/n\sum_{k=1}^{n} I_{(k,j)} \right| > \sigma_j \\ 0 & \text{else} \end{cases} \quad (3)$$

2.4 Segmentation by area

For the improvement of the obtained results, we will apply a segmentation method (A. Miene method [20]) which relies on the image segmentation of areas, then according to the criteria based on these segments, we can asses whether a segment belongs to a text or not.

Once the text is localized in the video frame, it needs to be segmented against the rest of the frame. All the segments which do not comply with these criteria will be removed. These criteria are summarized as follows:

-The length ratio by width must range from 0.33 and 15.

-The filling ratio must lie between 0.15 and 1.

-The area size must not exceed 50% in height and 25% in width.

- Absolute minimum height of the area is 5 px.

- Absolute minimum of the width of the area is 1 px.

2.5 Text Region Refinement

After having determined these stages, some of the bottom elements persist by complying with the already mentioned criteria randomly. In order to remove them, we gather the remaining, areas in words and we remove the rest of areas in the image.

The criteria of grouping the areas are as follows:

-A minimal number of areas associated together are 3.

-Area of almost equal height (a 20% margin will be acceptable)

-The distance between two characters of the same word will be small (the maximum distance is relative to the size of the characters)

Then, the candidate word zones are them selves analyzed in order to remove the area groups of candidate characters belonging to the bottom of the image. If the overlapping degree of two candidate words exceeds 80%, the smallest one will be regarded as the bottom part and will be removed

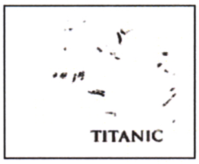

Fig.9. Image extracted after refinement

IV. RESULTS AND DISCUSSION

The method introduced in this paper was tested upon a corpus of color video sequences. This method was evaluated with that of Niblak[13] , Sauvola[14] and Wolf[12] independently in term of effectiveness and quality, respectively by measuring the computing time and by manually estimating the rates of Tr recall and The precision rate. Table 1 summarizes average measurements of effectiveness and of quality. We point out that the rates of recall and precision can be measured in the following way

$$T_r = \frac{N_c}{N_c + N_m}$$
$$T_p = \frac{N_c}{N_c + N_f} \quad (4)$$

With N_c, N_m, and N_f respectively representing the number of correct detections, the number of missed detections, and the number of false detections.

Approach	Time	Rate of Tr recall and Tp precision
Niblak	Not expensive in term of time	Tr weak Tp high
Sauvola	Not expensive in term of time	Tr weak Tp high
Wolf	Expensive in term of time	Tr average Tp average
Our Approach	Expensive in term of time	Tr average Tp high

Table 1: Assessment of time and quality measures

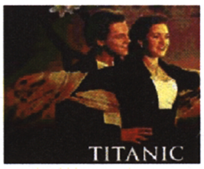

Fig10. An image of a video sequence

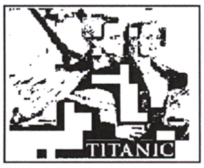

Fig11. Image extracted by the method of Sauvola

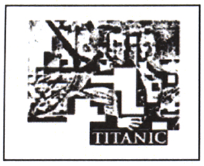

Fig12. Image extracted by the method of Niblak

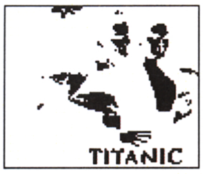

Fig13. Image extracted by the method of Wolf

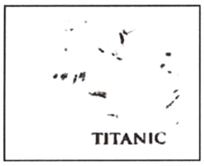

Fig13. Image extracted by our method

V. CONCLUSION

According to the experiments carried out on a broad video scene and concluding results from the comparison with other methods we can prove the validation of our approach.

Knowing that it is possible to improve the results of our approach by relying on post-processing so as to reduce false alarms.

The approach presented in this paper, is a humble contribution to the development of dubbing tools for film or visual cultural and scientific documentaries.

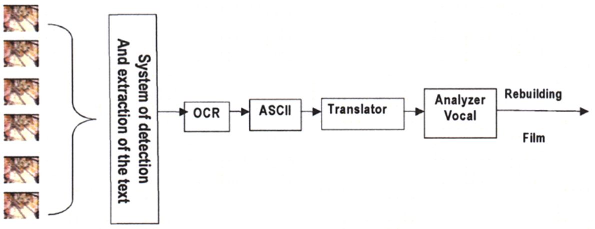

Fig14 Diagram of the total system considered

For the moment we are in the phase of detection and extraction. Besides our final goal is to manage to finalize the system schematized below.

So far we have noticed in the detection phase a potential improvement. The continuity of this work is in the recognition phase. We have already carried out tests upon the main commercial tools of character recognition (OCR). In this study, we can deduce the minimal characteristics which have to validate the texts included in video to be recognized.

REFERENCES

[1] Y.I. Ohta,T. Kanade,et T.Sakai.Color information for region segmentation.*Computer Graphics and Image Processing*, 13:222–241, 1980.

[2] R.K. Kouassi, J.-C. Devaux, P. Gouton, et M. Paindavoine. Application of the Karhunen-Loeve transform for natural colour images analysis. In *Irish Machine Vision and Image Processing Conference*,volume1, pages 20–27, Londonderry, 1997.

[3] ZHONG Y.,KANT K., JAIN A.,Locating text in complex color image, *Pattern Recognition*,vol.28, no10, pp1528–1535,1995

[4] KIM H., Efficient automatic text location method and content-based indexing and structuring of video database, *Journal of Visual Communication and Image Representation*, , no 4, 1996, pp. 336–344.

[5] WU V., MANMATHA R., RISEMAN E., TextFinder: an automatic system to detect and recognize text in images, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no 11, 1999, pp. 1224–1229.

[6] ZHONG Y., ZHANG H., JAIN A., Automatic caption localization in compressed video, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no 4,2000,pp.385–392.

[7] DIMITROVA N., AGNIHOTRI L., DORAI C., BOOLE R., MPEG-7 Videotext description scheme for superimposed text in images and video, *Signal Processing : Image Communication*, vol. 16, 2000, pp. 137–155.

[8] LIENHART R., WERNICKE A., Localizing and Segmenting Text in Images and Videos, *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 12, no 4, 2002, pp. 256-268.

[9] A.K. Jain et B. Yu. Automatic Text Location in Images and Video Frames. Technical Report MSU-CPS-97-33, PRIP Lab., Department of Computer Science, 1997.

[10] T. Sato, T. Kanade, E.K. Hughes, M.A. Smith, et S. Satoh. Video OCR : Indexing digital news libraries by recognition of superimposed captions. *ACM Multimedia Systems : Special Issue on Video Libraries*, 7(5) :385–395, 1999.

[11] F. LeBourgeois. Robust Multifont OCR System from Gray Level Images.Dans Proceedings of the 4th International Conference on Document Analysis and Recognition,pages 1–5, 1997.

[12] C.Wolf, M.J. Jolion, Text Localization, Enhancement and Binarization in Multimedia Documents *Proceedings of the International Conference on Pattern Recognition (ICPR),* vol 4, pages 1037-1040, 2002

[13] W. Niblack, an introduction in digital image processing, pages 115,116, Englewood Cliffs, N.J. ; Prentice Hall,1986

[14] J.Sauvola, T.Seppanen, S. Haapakoski , et M.Pietikainen. *Adaptative ocument binarization.* International Conference on document analysis and recogntion, volume 1,pages 147-152,1997

[15] H. Li et D. Doerman. A Video Text Detection System based on Automated Training.Dans Proceedings of the International Conference on Pattern Recognition 2000, pages 223–226, 2000.

[16] Yan Hao, Zhong Yi, Hou Zang-guang, Tan Min , automatis text detection in video frames dased on Bootstrap Artificial Neural Network and CED, journal of WSCG, vol 11, No 1, 2 003.

[17] Y. Zhong, H. Zhang, et A.K. Jain. Automatic Caption Localization in Compressed Video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(4):385–392, 2000.

[18] D. Crandall et R. Kasturi. Robust Detection of Stylized Text Events in Digital Video. Dans Proceedings of the International Conference on Document Analysis and Recognition, pages 865–869, 2001.

[19] R. Ohlander, K. Price, et D. R. Reddy. Picture segmentation using a recursive region splitting method. *Computer Graphics and Image Processing*, 8 :313–333, 1978.

[20] A. Miene, Th. Hermes and G. Ioannidis Extracting Textual Inserts from Digital Videos In Proc. of the Sixth International Conference on Document Analysis and Recognition (IDCAR'01), pp. 1079-1083, Seattle, Washington, USA, IEEE Computer Society, September 10-13, 2001.