Caption Localisation in Video Sequences by Fusion of Multiple Detectors

Sébastien Lefèvre
LSIIT, University of Strasbourg I
Parc d'Innovation, Bd. Brant, BP 10413
67412 Illkirch Cedex, France
lefevre@lsiit.u-strasbg.fr

Nicole Vincent
CRIP5, University of Paris V
45 rue des Saints Pères
75270 Paris Cedex 06, France
nicole.vincent@math-info.univ-paris5.fr

Abstract

In this article, we focus on the problem of caption detection in video sequences. Contrary to most of existing approaches based on a single detector followed by an ad hoc and costly post-processing, we have decided to consider several detectors and to merge their results in order to combine advantages of each one. First we made a study of captions in video sequences to determine how they are represented in images and to identify their main features (color constancy and background contrast, edge density and regularity, temporal persistence). Based on these features, we then select or define the appropriate detectors and we compare several fusion strategies which can be involved. The logical process we have followed and the satisfying results we have obtained let us validate our contribution.

1. Introduction

Nowadays, the amount of multimedia data is so high that some indexing tools are required to let users browse among the available informations and look for pertinent samples. Among these indexing tools, shot and scene change detectors help to break down the video sequences [7], tracking techniques give successive positions of the objects, while keyframe extractors can be used to publish a visual index of the video content. Another information of main interest is the caption text present in video frames.

Once the text has been extracted from video sequences, it may then be used to build some data textual annotations which are easy to index. So detection and analysis of caption text in video sequences is a main problem of multimedia data indexing. In this paper, we focus on the caption extraction for which we propose an efficient approach. We have to precise that scene text detection is considered as another problem and is out-of-scope of our study. Before describing the main aspects of our contribution, we will briefly recall related works. For more information, a recent review has been done by Jung *et al.* [5], whereas Hua *et al.* [3] focus on evaluation of caption detection methods.

The approaches described in literature may be classified depending on the image properties used: color [4], texture [12], motion [1], or contours [10]. Some methods use the temporal constancy of caption areas [9]. Most of the methods share the same processing synopsis: a main detector is first used (usually based on one of the features) and a post-processing is then required to increase the quality of the results. This post-processing *ad hoc* is often characterized by a high computational complexity. However, it seems that no method can be used stand-alone to give satisfying results [5]. So, as in [8, 2], we have rather decided to build our method as a combination of several detectors in order to take into account the advantages of all features and to avoid a post-processing step. Moreover we have determined the optimal detectors from the caption area features.

We will first describe the goal of caption areas in video sequences and study how these areas are represented in video frames. This study will then help us to determine the main features from which we can build the most appropriate detectors, which will be explained in the next section. We will also present the fusion strategies to combine the different detectors. Finally we will describe and comment the results obtained, which let us evaluate our contribution.

2. Study of caption features

In order to determine which are the features the most adapted to detect caption areas in video sequences, we have first to study these areas. Differents reasons can be put forward by the audiovisual production team to artificially insert text in video frames. This text can have several goals and be related to: commerce to mention companies or individuals taking part to a movie, sport to resume the game evolution (elapsed time, score) and associated data (player names, statistics), news to describe the current report (journalist name, place, abstract), law to indicate the rights attached to a document, *etc*. However, the caption text always plays a specific role in the video sequence, and so has to be easily visible by any watcher.

More precisely, from the study of a representative video corpus, we can notice that text areas are clearly separated from the other parts of image content, as the text would be easily read. The contrast between text area and text or scene background is relatively high. Moreover, these areas are displayed in front and are never hidden: so they are always completely visible. We can also observe that most of the time characters are monochrome, and the letters belonging to an only one word are usually displayed with unique color and texture. The last idea related to color constancy of these caption areas is the planar properties of the text: while the image content usually represents 3-D data, the caption areas are located on a 2-D plan.

Another important feature of text areas concerns the character shapes. These characters have most of the time constant size, shape, and orientation. The character size is usually determined by following some legibility and readibility rules: all characters have the same size, interspaces are fixed, the number of words per line is not greater than 5, *etc*. Depending on the geographic area, the text can be read from left to right, from right to left, or from top to bottom. Another interesting point is that areas usually contain many contours, which are also regular. It is the concatenation of characters from a similar font, and these characters would be easily distinguished from the background.

Finally, we can also notice the temporal constancy of caption areas in video sequences: these areas move only rarely in the spatial plane from one frame to the next one, and the associated motion is low. Moreover, the characters contained in the text appear in the following frames.

The preliminary conclusions of our study let us consider that caption area features can be determined *a priori* in order to define related optimal detectors. These features are linked to area color and texture constancy, contrast with other parts of the image, regular shapes of text characters and high contour density, and temporal persistence of caption areas. From these conclusions, we are now able to define some appropriate caption detectors. In order to ensure the highest possible efficiency to our method, we will mainly select detectors with a low computational cost.

3. Description of selected detectors

From the previous observations, we can conclude that the expected areas are parts of the image which are characterized by uniform color and texture, high contrast with background, dense and regular contours to delimit the different letters of the text, and a temporal persistence on several frames. As we take into account the computation time, we will focus on fast and complementary detectors. We have decided to retain three kinds of detectors, based respectively on color, texture, and contour information, and to consider temporal persistence as a possible filtering step.

3.1. Color-related detector

The first detector we will use is based on color constancy of caption areas. As the contrast between these areas and the background is high, we can assume that in a caption area at least two colors are displayed, which are related to text and background pixels. So we can discard all uniform areas.

We start our analyse by dividing the image I into blocks. For each block and each color component c, we compute the local histogram H_c. In order to increase the robustness of this analysis, we have decided to reduce the number of color values to be used. We then assume that a block B_k contains text if there is no high value in the histograms:

$$B_k = \left\{ \begin{array}{ll} 1 & \text{if } \forall c\, \forall v \quad H_c(v) < S_c \\ 0 & \text{otherwise} \end{array} \right. \tag{1}$$

where $H_c(v)$ is the vth bin of the histogram for color component c, and S_c is a predefined threshold.

3.2. Texture-related detector

In order to locate the caption areas based on their texture constancy, we used Haar wavelets [9]. However, we have decided to limit our analysis to the first level to keep the areas with regular texture.

A decomposition of the image I following the horizontal, vertical and diagonal directions let us obtain three images I^{LH}, I^{HL} and I^{HH}, which are then summed up in a single image I' of reduced size, so for a pixel I'_p we have:

$$I'_p = \left\{ \begin{array}{ll} 1 & \text{if } \frac{1}{3}(I_p^{HL} + I_p^{LH} + I_p^{HH}) > S_{t_1} \\ 0 & \text{otherwise} \end{array} \right. \tag{2}$$

The result is summed up at block level, allowing then to locate the areas with regular texture:

$$B_k = \left\{ \begin{array}{ll} 1 & \text{if } \left(\sum_{p \in B_k} I'_p \right) > S_{t_2} \times \omega(B_k) \\ 0 & \text{otherwise} \end{array} \right. \tag{3}$$

where $\omega(B_k)$ is the number of pixels from block B_k and S_{t_2} the rate used in the comparison.

3.3. Contour-related detectors

Two main caption features related to contours have been identified: density and regularity of edge pixels. So here we propose to use two appropriate different detectors.

The edge density is inspired from [11]. The goal is here to label an image region as a caption area if it contains many edge pixels. First we identify the edge pixels with a binarisation of the gradient image obtained with Sobel operator.

A block-based processing helps then to estimate locally the edge density (*i.e.* the number of edge pixels):

$$\delta(B_k) = \sum_{p \in B_k} E_p \tag{4}$$

where:

$$E_p = \begin{cases} 1 & \text{if } I_p^{\text{Sobel}} > S_{c_1} \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

and I^{Sobel} the Sobel gradient image obtained from original image I. We consider that a block B_k of high density $\delta(B_k) > S_{c_2}$ belongs to a caption area.

To detect caption areas based on contour regularity, we assume that, as the caption areas are composed of printed text, they usually contain some line segments of predefined (mainly vertical and horizontal) directions. So we use a fast block-based line segment detector [6] which is of particular interest for horizontal and vertical directions. The results consist for each block in the presence and the position of one or several line segments of predefined direction. An area characterized by an important number of line segments of similar direction with close but non connected positions is assumed to be a caption area.

Here we have introduced four detectors based on identified caption features and an efficiency criterion. In order to ensure more efficiency and quality, these detectors can involve the temporal invariance feature.

3.4. Involving the temporal invariance

The temporal invariance feature is involved in each of the different detectors, following two alternative principles.

The first way is to consider that, for a given detector, the input image contains only the areas extracted by this detector on the previous video frame. A reset to the complete image is performed periodically. This principle helps to limit the computation time, and can be formulated as:

$$I'(t) = \begin{cases} D(I_t) & \text{if } t \mod \Delta = 0 \\ D_{I'_{t-1}}(I_t) & \text{otherwise} \end{cases} \tag{6}$$

where $D(I_t)$ represents the application of detector D to the image I_t, $D_{I'_{t-1}}$ defines the restriction of the detector D to detected areas on the previous frame at time $t-1$ (and noted I'_{t-1}), and Δ measures the reset step.

The second way consists in assuming that a caption area will be kept only if it has been extracted by the same detector for a given number of successive frames, *i.e.*:

$$I''(t) = \bigwedge_{k \in [t-\lambda, t]} I'(k) \tag{7}$$

where the final image I'' at time t is obtained from a combination of the results I' on a range of λ successive frames.

We can then obtain some local segmentation results using the different detectors with a temporal invariance principle. We will now show how to fusion the different results in order to obtain a global segmentation decision.

4. Fusion strategies

Each of our detectors works on pixel blocks. In every case, the blocks have same size and represent an image partition. A fusion is then required, and in our case it will be applied at block level. In order to formalize our fusion strategy, we represent each detector by a function defined on an image of various size and with values in a binary set:

$$\begin{array}{rccl} D: & \mathcal{I} & \longrightarrow & \{0,1\} \\ & I & \longmapsto & D(I) = \begin{cases} 1 & \text{if } \mathcal{C}(\mathcal{I}) \\ 0 & \text{otherwise} \end{cases} \end{array} \tag{8}$$

with the condition :

$$\mathcal{C}(\mathcal{I}) : \text{the block } I \text{ is labelled as text} \tag{9}$$

To qualify a block I we have defined two strategies for combination of detectors.

The first strategy considers a parallel processing of every detector. The results are then merged using weighting coefficients with detectors. These coefficients can be defined *a priori* from a learning step or set on line. A region is kept only if its global score (the sum of the weighted individual scores) is higher than a predefined threshold. Using the notations introduced previously, we have:

$$\begin{array}{rccl} D_{\text{final}}: & \mathcal{I} & \longrightarrow & \{0,1\} \\ & I & \longmapsto & D_{\text{final}}(I) = \begin{cases} 1 & \text{if } \sum_{i=1}^{k} p_i D_i(I) > S_f \\ 0 & \text{otherwise} \end{cases} \end{array} \tag{10}$$

for k detectors where p_i represents the different weights associated with detectors D_i, and S_f the global threshold.

The second strategy considers a sequential processing of the different detectors. Moreover, this processing can be seen as hierarchical. The detectors are sorted and indexed depending on their efficiency and their tolerance when applied on the complete frame. The first is the most tolerant, whereas the last is of best quality but of worse efficiency. The strategy can then be expressed for a block I as follows:

- D'_i is defined for $i > 1$ by:

$$\begin{array}{rccl} D'_i: & \mathcal{I} & \longrightarrow & \{0,1\} \\ & I & \longmapsto & D'_i(I) = \begin{cases} D_i(I) & \text{if } D_{i-1}(I) = 1 \\ 0 & \text{otherwise} \end{cases} \end{array} \tag{11}$$

- D_{final} is defined from this sequence of operators $D_1, \ldots, D_k$ by $D_{\text{final}} = D'_k$

The two strategies introduced here have pros and cons. Whereas the first one can be based on a learning step to determine the optimal weights, it requires the processing of all detectors on complete frames. On the opposite, the second strategy is faster, particularly for a monoprocessor system, but some caption areas can stay undetected.

5. Results and discussion

The method introduced in this paper has been tested on a varied corpus of color video sequences. The detectors described in section 3 have been evaluated independently in terms of efficiency and quality, respectively by measuring the average computation time and by estimating the recall rate T_r and precision rate T_p. We have also evaluate the effects of the additional temporal persistence criterion and of the two fusion strategies. Finally, we have compared our results with those from Wolf and Jolion [11].

One of the contributions of this paper was to propose a fast caption detection method. Table 1 gives the average computation time T_μ of our detectors, based on a Java-based implementation on a PC workstation (3 GHz CPU and 512 MBytes RAM) with RGB images of 320×240 pixels. The cost of edge-related detectors is high due to the edge detection step. The first temporal persistence criterion helps to greatly decrease the computation time (about 80-95 %), whereas the second criterion implies a low additional cost (about 0.02 ms per frame for $\lambda = 2$). We have then estimated the quality rates. In order to ensure the robustness and genericity of our method, we have used a single parameter set for all our tests. These optimal parameters have been obtained following an analysis of the recall/precision curves on various images, and are given in the table 2. Using these parameters, we have computed the average recall rate T_r and precision rate T_p, also shown in table 1.

As we can notice, the parallel strategy returns better average quality, contrary to the sequential strategy which is faster. However, the latter can give better results with a parameter set ensuring a maximal recall rate.

In order to evaluate our contribution, we have also compared our results (with parallel strategy) with Wolf and Jolion method [11]. As illustrated in figure 1, we can notice that for a same number of false negatives, our method generally returns less false positives.

Figure 2 shows the application of the sequential strategy. The limit is that a caption area will be missed if at least one single detector does not retrieve it.

6. Conclusion

In this article we have introduced a new method for caption areas detection in video sequences. Contrary to most of the other approaches, we do not rely on a single detector followed by an *ad hoc* and costly post-processing but we rather consider several detectors simultaneously. In order to determine the detectors to be used, we have first made a study of caption areas in video sequences and we have identified their main features, related to color, texture, contours, and temporal invariance. From these features we have defined or selected the appropriate detectors. We have then introduced two different strategies to fusion the results obtained with each of the detectors in a global decision, either in a parallel or a hierarchical way. We have finally compared the detectors and the strategies on a various video corpus, which let us validate our contribution.

Among the perspectives we consider, we can mention the use of robust detectors as the flat morphological operators to detect areas with uniform colors. We think also to adapt our method to compressed video data in order to process the video frames directly in the compressed domain.

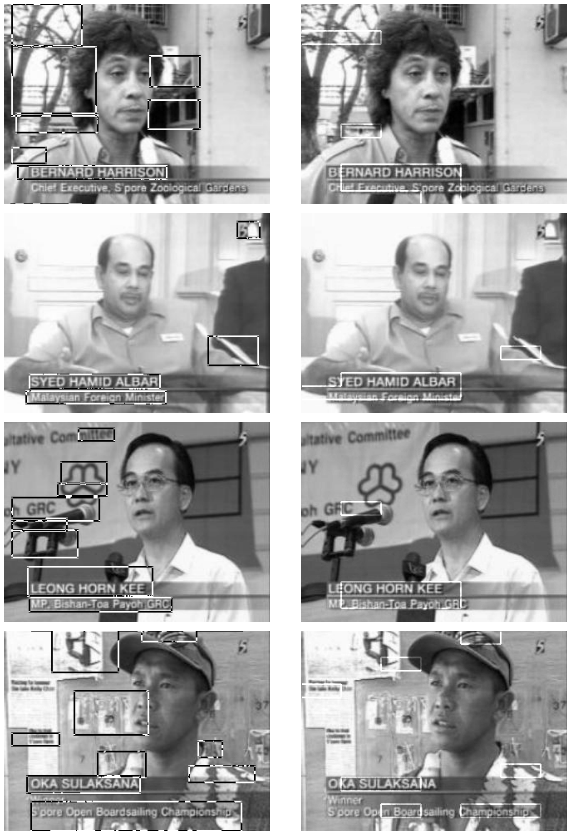
BERNARD HARRISON
Chief Executive, S'pore Zoological Gardens
BERNARD HARRISON
Chief Executive, S'pore Zoological Gardens
SYED HAMID ALBAR
Malaysian Foreign Minister
SYED HAMID ALBAR
Malaysian Foreign Minister
LEONG HORN KEE
MP, Bishan-Toa Payoh GRC
LEONG HORN KEE
MP, Bishan-Toa Payoh GRC
OKA SULAKSANA
S'pore Open Boardsailing Championship
OKA SULAKSANA
Winner
S'pore Open Boardsailing Championship

Figure 1. Results obtained with Wolf and Jolion method [11] (left) and with our method (right).

References

[1] T. Gandhi, R. Kasturi, and S. Antani. Application of planar motion segmentation for scene text extraction. In *IAPR International Conference on Pattern Recognition*, volume 1, pages 445–449, Barcelona, Spain, September 2000.

[2] X. Hua, L. Wenyin, and H. Zhang. Automatic performance evaluation for video text detection. In *International Conference on Document Analysis and Recognition*, pages 545–550, Seattle, USA, September 2001.

[3] X. Hua, L. Wenyin, and H. Zhang. An automatic performance evaluation for video text detection algorithms. *IEEE Transactions on Circuits and Systems for Video Technology*, 14(4):498–507, 2004.

[4] K. Jung. Neural network-based text location in color images. *Pattern Recognition Letters*, 22:1503–1515, 2001.

[5] K. Jung, K. Kim, and A. Jain. Text information extraction in images and video: a survey. *Pattern Recognition*, 37:977–997, 2004.

[6] S. Lefèvre, C. Dixon, C. Jeusse, and N. Vincent. A local approach for fast line detection. In *IEEE International Conference on Digital Signal Processing*, volume 2, pages 1109–1112, Santorini, Greece, August 2002.

[7] S. Lefèvre, J. Holler, and N. Vincent. A study of real-time segmentation of uncompressed video sequences for content-based search and retrieval. *Real-Time Imaging*, 9(1):73–98, 2003.

[8] C. Li, X. Ding, and Y. Wu. Automatic text location in natural scene images. In *International Conference on Document Analysis and Recognition*, pages 1069–1074, Seattle, USA, September 2001.

[9] H. Li, D. Doerman, and O. Kia. Automatic text detection and tracking in digital video. *IEEE Transactions on Image Processing*, 9(1):147–156, 2000.

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

[11] C. Wolf and J. Jolion. Extraction and recognition of artificial text in multimedia documents. *Pattern Analysis and Applications*, 6:309–326, 2003.

[12] Y. Zhong, H. Zhang, and A. Jain. Automatic caption localization in compressed video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(4):385–392, 2000.

Detector	T_μ (in ms)	T_r (in %)	T_p (in %)
Color	4.17	65	32
Texture	7	86	85
Edge density	30	93	74
Edge regularity	41.17	85	85
Parallel strategy	82.33	92	76
Sequential strategy	36.5	60	90

Table 1. Efficiency and quality measures.

Parameter	Value
Block height	$h = 16$
Block width	$v = 48$
Number of quantified colors	$V = 6$
Color threshold	$S_c = 0.45 \times h \times v$
Texture first threshold	$S_{t_1} = 15$
Texture second threshold	$S_{t_2} = 0.2 \times h \times v$
Contour first threshold	$S_{c_1} = 75$
Contour second threshold	$S_{c_2} = 0.15 \times h \times v$
Weights in fusion process	$\forall i \quad p_i = 1$
Threshold in fusion process	$S_f = 2$

Table 2. Parameters and selected values.

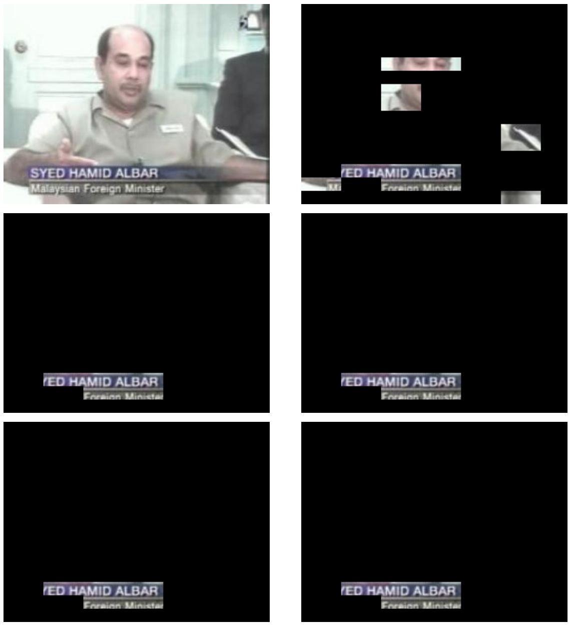

Figure 2. Results obtained with the sequential strategy (from left to right and top to bottom): original image, result from detector based on color, texture, edge density, edge regularity and final result.