REGION-BASED CAPTION TEXT EXTRACTION

Miriam Leon, Veronica Vilaplana, Antoni Gasull, Ferran Marques

Technical University of Catalonia (UPC), Barcelona, Spain
mleon@tsc.upc.edu, {veronica.vilaplana,antoni.gasull,ferran.marques}@upc.edu

ABSTRACT

This paper presents a method for caption text detection. The proposed method will be included in a generic indexing system dealing with other semantic concepts which are to be automatically detected as well. To have a coherent detection system, the various object detection algorithms use a common image description. In our framework, the image description is a hierarchical region-based image model. The proposed method takes advantage of texture and geometric features to detect the caption text. Texture features are estimated using wavelet analysis and mainly applied for *Text candidate spotting*. In turn, *Text characteristics verification* is basically carry out relying on geometric features, which are estimated exploiting the region-based image model. Analysis of the region hierarchy provides the final caption text objects. The final step of *Consistency analysis for output* is performed by a binarization algorithm that robustly estimates the thresholds on the caption text area of support.

1. INTRODUCTION

The text present in a scene can provide relevant information for scene semantic analysis. This is specially true for caption text which is usually synchronized with the contents in the scene. Caption text is artificially superimposed on the video at the time of editing and it usually underscores or summarizes the video content. This makes caption text particularly useful for building keyword indexes [1]. Text detection algorithms can be divided into three phases [2]:

1. *Text candidate spotting:* where an attempt to separate text from background is done.
2. *Text characteristics verification:* where text candidate regions are grouped to discard those regions wrongly selected.
3. *Consistency analysis for output:* where regions representing text are modified to obtain a more useful character representation as input for an OCR.

In this paper, we focus on caption text detection, extending the technique presented in [3]. Note that the proposed caption text detector is to be included in an already existing global system. This system aims at two main goals [3]: (i) off-line enrichment of the current annotation of very large video databases and, (ii) instantiation of new descriptors for future annotation of new semantic concepts. These goals impose two requirements to the caption text detector:

- Analysis of the video at the temporal resolution provided by the key frames that are currently stored .
- Use of an image representation and description compacting in the smallest possible number of elements all the information in the scene, while being as generic as possible in order to reuse the representation in different contexts [4].

This work was partially founded by the Catalan Broadcasting Corporation (CCMA) and Mediapro through the Spanish project CENIT-2007-1012 i3media and TEC2007-66858/TCM PROVEC of the Spanish Government

Given these requirements, we proposed in [3] a method for caption text extraction in still images using a hierarchical region-based image representation. Here, improvements for the first two phases (*Text candidate spotting* and *Text characteristics verification*) and a solution for the third phase (*Consistency analysis for output*) are proposed. The presentation of these concepts is structured as follows. For the sake of completeness, Section 2 summarizes the image model concept [5] as well as its extension to the case of object detection [4] and, specifically, text detection. In Section 3, the region-based caption text detection approach is detailed. This Section is structured in three subsections in which every phase of the text detector is described. Subsection 3.1 discusses the use of wavelet information to spot the text candidates in the image [6]. In concrete, the use of the Haar transform in the color domain is presented and allows to extract those text candidates that, not being contrasted enough in the luminance component, still present a text-like texture pattern. In Subsection 3.2, geometrical descriptors are used to confirm the spotted candidates and discard false positives [7]. In that case, we take advantage of the region-based representation to estimate the geometrical descriptors [3] and of the hierarchical image description to obtain the best set of text caption representatives. In turn, Subsection 3.3 describes the proposal for the final step. It is performed by a binarization algorithm that robustly estimates the thresholds on the area of support of the caption text candidate and provides the final input to the OCR. Section 4 discusses the results obtained by this technique. Finally, conclusions are outlined in Section 5.

2. HIERARCHICAL REGION-BASED IMAGE MODEL

The Binary Partition Tree (BPT) [5] reflects the similarity between neighboring regions. It proposes a hierarchy of regions created by a merging algorithm that can make use of any similarity measure. Starting from a given partition, the region merging algorithm proceeds iteratively by (1) computing a similarity measure for all pair of neighbor regions, (2) selecting the most similar pair of regions and merging them into a new region and (3) updating the neighborhood and the similarity measures. The algorithm iterates steps (2) and (3) until all regions are merged into a single region. The BPT stores the whole merging sequence from an initial partition to the one-single region representation. The leaves in the tree are the regions in the initial partition. A merging is represented by creating a parent node (the new region resulting from the merging) and linking it to its two children nodes (the pair of regions that are merged).

3. CAPTION TEXT DETECTION APPROACH

Caption text can be described as text added inside a rectangular bar, horizontally aligned, highly contrasted regarding the bar background and with textured aspect. These features are commonly translated into two types of descriptors: texture and geometric de-

scriptors which are typically used for text candidate spotting and text characteristic verification, respectively.

Textured areas can be detected using wavelet analysis. However, this approach produces many false positives (that have to be filtered out using geometric descriptors) and some misses in low contrast areas. On the other hand, given the generic framework of our application, the BPT has been created combining color and contour homogeneity criteria [4]. Due to their homogeneous background and regular shape, caption text objects are likely to appear as single nodes in the BPT. Hence, we propose to combine the two approaches.

3.1. Text candidate spotting

As proposed in [6], texture descriptors such as DWT coefficients give enough information to determine where textured areas can be found in an image. In [3] we proposed to take advantage of the power of the LH and HL subbands in a Haar transform (Fig. 1.b) analyzed over a sliding window of fixed size HxW (typical values are 6x18; W>H to consider horizontal text alignment):

$$P^l_{LH}(m,n) = \frac{1}{WH}\sum_{i=1}^{W}\sum_{j=1}^{H} LH^l(m+i, n+j) \qquad (1)$$

where l denotes the decomposition level and an analogous expression should be used for P^l_{LH}. The window is moved over subbands of the transformed image with an overlapping of half the window size in both directions. Both subbands are analyzed because DWT power in windows containing text should present high values in at least one of them and relevant enough values in the other subband. This way, all pixels in a window are classified as text candidates if the power in the window fulfills the following condition:

$$((P^l_{LH} > T_1) \wedge (P^l_{HL} > T_2)) \vee ((P^l_{LH} > T_2) \wedge (P^l_{HL} > T_1)) \quad (2)$$

where T_1 and T_2 are two thresholds, T_1 being more restrictive than T_2.

As previously commented, the wavelet analysis may produce misses in low contrasted areas. In the case of caption text, such misses are commonly related to text over a background that share a similar luminance value but whose chrominance values are different enough to be separately perceived by the human visual system. Taking into account this observation, the previous technique has been separately applied to the three YUV image components.

The final mask marking all the text candidates is obtained by performing the union of the (upsampled) masks at each decomposition level (Fig.1.c) and at each image component. For the results presented in Section 4, $l = 2$, $T_1Y = 1200$ and $T_2Y = 400$ for luminance, and $T_1uv = 18$ and $T_2uv = 10$ for chrominance.

Finally, regions in the search partition (Fig. 1.d) are selected if they contain any text candidate pixel. Moreover, texture-based selection is propagated through the BPT so that all ancestors of the candidate regions are selected as well (Fig. 1.g). This is a very conservative policy but, at this stage, it is important not to miss any possible region containing text (Fig. 1.e).

3.2. Text characteristic verification

For every selected node, descriptors are estimated to verify if the region represents a caption text object. Initially, a region-based texture descriptor is computed as in eq.(1) but now the sum is performed over interior pixels to avoid the influence of wavelet coefficients due to the gradient in the region boundary. Mainly, this descriptor is used to filter out regions that have been selected due to the presence in the mask (see Fig. 1.c) of a few wrong candidate pixels in the surroundings of textured areas.

To complete the verification process, geometric descriptors are calculated for every remaining candidate node. Before computing these descriptors, the area of support of candidate nodes is modified by a hole filling process and an opening with a small structuring element (typically, 9x9). This stage is needed to eliminate small leaks that the segmentation process may introduce due to the interlacing or to color degradation between regions. Such leaks result in very noisy contours that bias the geometric descriptor estimation. Finally, since the opening may split the region into several components, the largest connected component is selected as the area of support for computing geometric descriptors.

Given the regular shape (close to rectangular) of caption text objects, the geometric descriptors used in this work are often compared to those of the bounding box (BB) of the node area of support. Descriptors and the thresholds they should accomplish (following a restrictive policy) are listed in the sequel. Values in brackets indicate the thresholds used for the experiments presented in Section 4 for standard PAL format 720x576 images.

- *Rectangularity (R):* Occupancy in [3] has been replaced by the calculation of the rectangularity discrepancy. R must be greater than T_R; the nearer to 1, the more similar to a rectangle ($T_R = 0.85$).
- *Aspect ratio:* ($AR = Width_{BB}/Height_{BB}$) must be in the range $[T_{AR_1}, T_{AR_2}]$, the superior limit is not strictly necessary but helps avoiding line-like nodes ($T_{AR_1} = 1.33$, $T_{AR_2} = 20$).
- *Height:* must be in the range $[T_{H_1}, T_{H_2}]$ ($T_{H1} = 13$ pixels for character visibility and $T_{H_2} = 144$, a quarter of PAL format height).
- *Area:* must be in the range $[T_{A_1}, T_{A_2}]$ ($T_{A_1} = 225$, the area of a node with minimum height and minimum aspect ratio, and $T_{A_2} = 138.240$, a third of the PAL format image area).
- *Compactness:* ($CC = Perimeter^2/Area$) must be smaller than T_{CC}, to avoid nodes with long, thin elongations commonly due to interlacing ($T_{CC} = 800$).

The result of applying these descriptors is presented in Fig. 1.g, where the verified nodes are marked in orange.

At this stage, verified nodes may present two problems. First, as shown in Fig. 1.h, it is common that several verified nodes are in the same subtree; that is, several (maybe partial) instances of the same caption text object may be represented in a subtree. Second, if the image contains a collection of caption text bars laying close enough, they may be gathered into a single node; that is, a single subtree may represent several caption text objects that, due to their proximity, can be understood as a single one.

The first problem leads to the presence of unnecessary verified nodes, actually representing the same caption text object, that are to be processed in the *Consistency analysis for output* step. In that case, the best node in the subtree has to be selected. The straightforward solution of selecting the highest node in the subtree may lead to non-optimal solutions, as discussed in [3]. In that work, a confidence value was estimated for each node and those nodes in the subtree leading to the highest confidence value were finally preserved.

Nevertheless, a second problem has been detected due to the presence of several caption text objects in the image, which can be understood as a single one given its global texture and geometric features. Such configurations are very common in, for instance, sport events where the data of several participants are jointly presented. In that case, the problem is more severe due to possible differences in the colors of fonts and backgrounds used in the neighbor caption

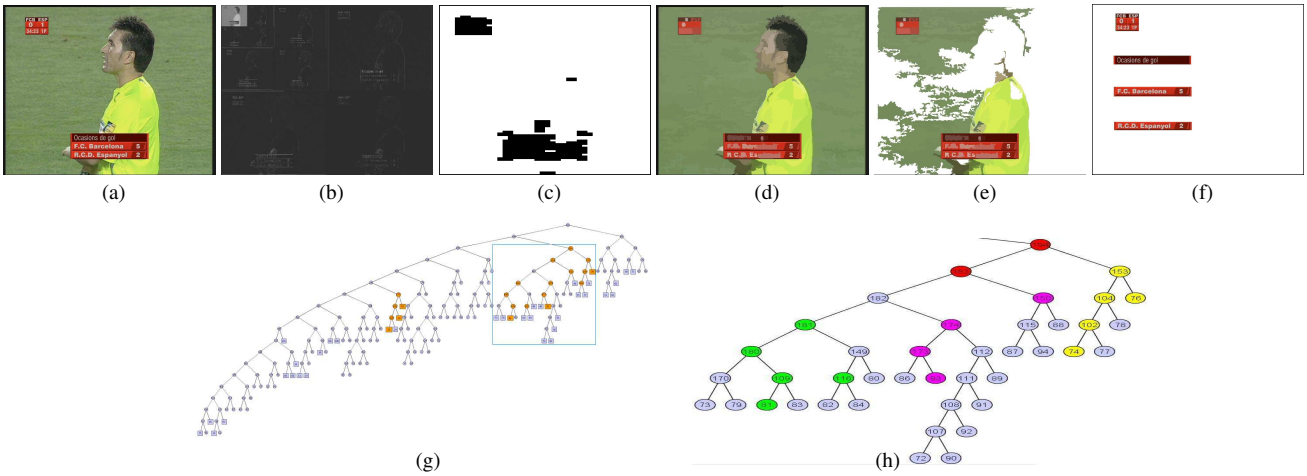

Fig. 1: Example of caption text detection. (a) Original image, (b) Wavelet transform, (c) Text candidate pixels, (d) Search partition (e) Text candidate regions, (f) Set of final selected regions, (g) BPT showing the selected leaves (squared nodes) and the candidate nodes (orange nodes), (h) Detail of the BPT (rectangle in g)) showing the final selected nodes for each text bar (green, lilac and yellow nodes) and the discarded nodes (red nodes)

text bars. When selecting the set of caption text bars as a single object, these differences result in a decrease in the performance of the subsequent *Consistency analysis for output* step. This step relies on a binarization of the validated caption text bar area of support and if the two classes to be detected (character and background) are not homogeneous in color, the classification cannot be correctly done.

Having in mind these two problems, we propose here a new strategy to handle jointly both situations in a more robust manner. Subtrees are traversed in postorder. For each subtree, a list of possible caption text objects is produced. Verified nodes in the subtree are analyzed and they are compared with the previous caption text objects in the list. If the geometrical features of the verified node under analysis allow us to assume that this node belongs to a caption text object already in the list, the verified node under analysis is assigned to this caption text object and the characteristics of this caption text object are updated. If the verified node under analysis cannot be assigned to any already existing caption text object, it is included in the list as a new caption text object.

All these comparisons are performed only using simple geometrical descriptors previously extracted from the tree nodes. In particular, the features that are compared between a verified node under analysis and an already existing caption text object are the coordinates of its center of mass as well as the height and the width of the modified node bounding box (see Subsection 3.2). Combining these three elements, the following situations can be detected:

1. The node completes an already existing caption text object: This is the case of a caption text object that is mostly represented by a single node in the BPT but some parts of it (for instance, of its interior) are missing. In that case, neither the y-component of the center of mass nor the height or the width of the BB present a substantial change. The node is assigned to this caption text object and the parameters are updated.
2. The node horizontally extends an already existing caption text object: This is the case of a caption text object that has been split in the BPT into two horizontal-neighbor regions. The y-component of the center of mass and the height of the bounding box do not present a substantial change, whereas the width of the bonding box increases. The node is assigned to this caption text object and the parameters are updated.

For other situations, the overlap between the node under analysis and the extension of the area of support of the caption text object is analyzed. If they overlap, the node is assumed to be part of the caption text object and its parameters are updated. If they do not overlap, a new caption text object is defined.

In the example of Fig. 1.a the biggest text box is detected as three separated text bar. Figure 1.h shows a subtree whose root node represents this text box. The search algorithm detects the nodes with the same color as nodes which are part of the same text bar, obtaining each text bar independently, see Fig.1.f.

3.3. Consistency analysis for output

For every caption text object, a binarization step is carried out. Given the specific characteristics of caption text bars, the binarization is performed by analyzing a few lines in the image. N (typically N=3) equidistant horizontal line segments are selected within the area of support of the caption text object. The mean and the variance of the pixels in each line segment are computed. Those line segments presenting high variance are assumed to be formed by text and background and are used to estimate the binarization threshold. In turn, low variance line segments are supposed to only represent the background and can be used to characterize it. An example illustrating this process is presented in Fig. 2. As it can be seen (and it will be further discussed in next section), this approach leads to good results. Other approaches have been also tested (as applying k-means clustering with k = 2) leading to lower performance.

This way, the output of the method is directly used as input for the OCR system. In this work, we have used the open-source **tesseract-ocr** system (http://code.google.com/p/tesseract-ocr/) which can be trained for a specific language and vocabulary.

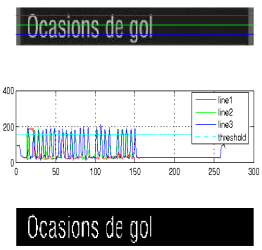

Fig. 2: Illustration of the caption text binarization process for N=3.

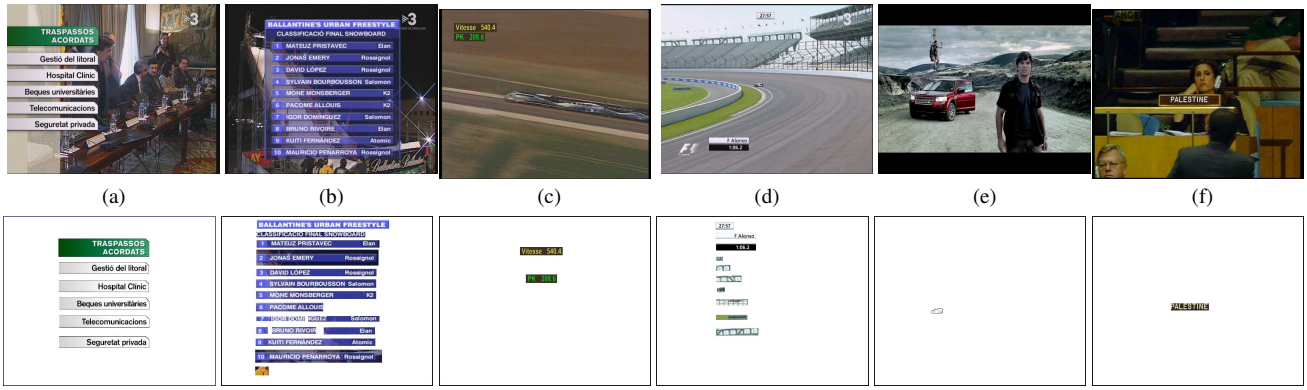

Fig. 3: Illustration of the caption text detection process. First row: Original images; Second row: Final selected regions

4. RESULTS

The technique has been tested in two corpus formed mainly by news and sport event videos, and only sport event videos, respectively[1]. There is a total of 249 caption text objects extracted from a set of 150 challenging images with text of different size and color, and complex background textures in the first corpus. Results, classified as correctly detected, partially detected, false positives and false negatives, are summarized in Table 1 and illustrated in Fig. 3.

	Detected Objects	**% over 249 objects**
Correctly detected	215	86.35%
Partially detected	22	8.83%
False negative	12	4.82%

Table 1: Detection results related to the number of objects in the 1st database

	Detected Objects	**% over 2063 objects**
Correctly detected	1758	85.21%
Partially detected	40	1.93%
False negative	265	12.84%

Table 2: Detection results related to the number of objects in the 2nd database

If these values are expressed in terms of recall and precision, partially detected objects (PDO) can be considered as false negative or as detected objects since they represent good anchor points for the following step (see Table 3). The number of false positives is 24. Results do not differ significantly from [3] but text bars are detected separately instead of together in a single text box.

PDO	**as outlier**	**as correct**	**as outlier**	**as correct**
Recall	0.863	0.950	0.8521	0.871
Precision	0.885	0.894	0.692	0.697

Table 3: Detection results presented as precision and recall for the first and second database, respectively.

In the second corpus there is 2063 caption text objects extracted from 649 key frames. The most remarkable result is that the number of false positives is very high, both public and advertising panels make this value increasing, whereas the number of detected text bar is satisfying (see Tables 2 and 3). But some of these elements will be discarded in the third step (see Figs.3.d and e).

Figure 3 illustrates these results with some examples. Fig. 3.a presents an example of non-perfectly rectangular caption text objects. It is common that caption text objects present some modifications to make information more attractive for the viewer.

Fig. 3.b is an example to illustrate false negative and partial detections. The similarity between caption text background and objects around mislead the segmentation process and, in some cases, the caption text object is not correctly represented in the BPT. Moreover, we can illustrate as well an example of partial detections: caption text object marked with a "7" has been reported as partial detection since it has not been fully extracted as a single node. Fig. 3.c shows an image where exploiting color information (see subsection 3.1) provides good results, letters in fluorescent green would be discarded in this step due to low contrast if only luminance information had been used. Fig. 3.d and Fig. 3.e are representative examples of typical outliers. They correspond to highly textured, rectangular nodes. Nevertheless, they are removed in the following phase of consistency analysis for output. Finally, Fig. 3.f shows an example of scene text. This text can be detected by means of this caption text algorithm when text is placed in a rectangular bar and highly contrasted.

5. CONCLUSIONS

We have presented a new technique for caption text detection. This technique is to be included in a global indexing system and, therefore, takes advantage of a common hierarchical region-based image representation. The technique combines texture information (through Haar wavelet decomposition) and geometric information (through the analysis of the regions proposed by the hierarchical image model) to robustly extract caption text objects in the scene.

6. REFERENCES

[1] D. Crandall, S. Antani, and R. Kasturi, "Extraction of special effects caption text events from digital video," *Int. Journal on Document Analysis and Recognition*, , no. 2, pp. 138–157, April 2002.

[2] K. Jung, K. Kim, and A.K. Jain, "Text information extraction in images and video:a survey," *Pattern recognition*, vol. 37, pp. 977–997, 2004.

[3] M.Leon, V.Vilaplana, A.Gasull, and F.Marques, "Caption text extraction for indexing purposes using a hierarchical region-based image model," *IEEE ICIP 2009, El Cairo, Egypt*, November 2009.

[4] V. Vilaplana, F. Marqués, and P. Salembier, "Binary partition trees for object detection," *IEEE Transactions on Image Processing*, vol. 17, no. 11, pp. 2201–2216, November 2008.

[5] P. Salembier and L. Garrido, "Binary partition tree as an efficient representation for image processing, segmentation and information retrieval," *IEEE Trans. on Image Processing*, vol. 9, no. 4, pp. 561–576, 2000.

[6] H. Li, D. Doermann, and O. Kia, "Automatic text detection and tracking in digital video," *IEEE Transactions on Image Processing*, vol. 9, no. 1, pp. 147–155, January 2000.

[7] T. Retornaz and B. Marcotegui, "Scene text localization based on the ultimate opening," *Proc. ISMM*, vol. 1, pp. 177–188, January 2007.

[1] All images used in this paper belong to TVC, Televisio de Catalunya, and are copyright protected. These key-frames have been provided by TVC with the only goal of research under the framework of the i3media project.