CAPTION TEXT EXTRACTION FOR INDEXING PURPOSES USING A HIERARCHICAL REGION-BASED IMAGE MODEL

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 technique for detecting caption text for indexing purposes. This technique is to be included in a generic indexing system dealing with other semantic concepts. The various object detection algorithms are required to share a common image description which, in our case, is a hierarchical region-based image model. Caption text objects are detected combining texture and geometric features, which are estimated using wavelet analysis and taking advantage of the region-based image model, respectively. Analysis of the region hierarchy provides the final caption text objects.

Index Terms— Image segmentation, feature extraction, object recognition, text recognition

1. INTRODUCTION

Semantic image indexing relies on the annotation of the presence in the scene of some a priori defined semantic concepts. At a first level of abstraction, semantic concepts are commonly associated to objects. However, object detection is a non-solved problem in a general framework and the extraction of the text in the scene can provide additional relevant information for scene semantic analysis [1]. 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 [2]. For example, when recognizing a given location (e.g.: a street), in addition to the information obtained by recognizing the buildings in the image, a caption text associating the scene with a given city may help confirming the location.

As proposed in [3], text detection algorithms can be classified in two categories: those working on the compressed domain and those working on the spatial domain. Independently of their domain, algorithms can be divided into three phases: (i) *Text candidate spotting*, where an attempt to separate text from background is done; (ii) *Text characteristics verification*, where text candidate regions are grouped to discard those regions wrongly selected; and (iii) *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 the text candidate spotting and text characteristic verification phases within the context of caption text.

The caption text detector presented in this work will be included in a more generic indexing system. Actually, the global application is that of off-line enrichment of the current annotation of very large video databases (for instance, the whole repository of TV broadcasters) as well as of creation and instantiation of new descriptors for future annotation of new semantic concepts (for example, searching in the database for a person who previously did not require being explicitly annotated).

Two of the requirements imposed by this application are (i) analysis of the video at the temporal resolution provided by the key frames that are currently stored and (ii) 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 (e.g.: searching efficiently in the same image for different objects at different moments) [4].

Given the first constraint, we concentrate on the problem of caption text extraction in still images. Caption text presents some features that are typically used by text extraction algorithms. The horizontal intensity variations produced by the text are exploited in techniques that analyze the image in the transform domain, either using the DCT [5] or the wavelet transform [6]. Also spatial domain techniques take advantage of this feature by proposing edge detectors to spot the areas with high probability of containing text [7]. Next, spatial cohesion features, such as size, fill factor, aspect ratio or horizontal alignment, are applied to check if these regions are consistent with its neighborhood and to discard false positives [8].

Note that all these techniques are specific for text detection and commonly independent of the approaches dealing with the detection of other semantic concepts. In the case of using text in a global indexing system, it is interesting to have a common image representation and a (as much as possible) common set of descriptors.

Regarding the image representation, region-based image representations provide a simplification of the image in terms of a reduced number of representative elements, which are the regions. In a region-based image representation, objects in the scene are obtained by the union of regions in an initial partition. To reduce the number of possible region unions, it is useful to create a hierarchy of regions representing the image at different resolution levels. The idea is to have not only a single partition but a universe of partitions representing the image at various resolutions. In this context, object detection algorithms (and specifically text detection algorithms) only need to analyze the image at those positions and scales that are proposed by the regions in the hierarchy [4].

In a previous work, the tree of maxima (and minima) [9] was proposed as hierarchical region-based image model for text detection [10]. Nevertheless, in order to reuse the representation to detect other objects, the Binary Partition Tree (BPT) [11] is used in this work since its suitability for generic object detection has recently been demonstrated [4].

After this introduction, Section 2 briefly describes the image model. In Section 3, the region-based caption text detection approach is detailed. Section 4 discusses the results obtained by this technique. Finally, conclusions are outlined in Section 5.

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

2. HIERARCHICAL REGION-BASED IMAGE MODEL

The BPT [11] 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 (merging criterion) 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).

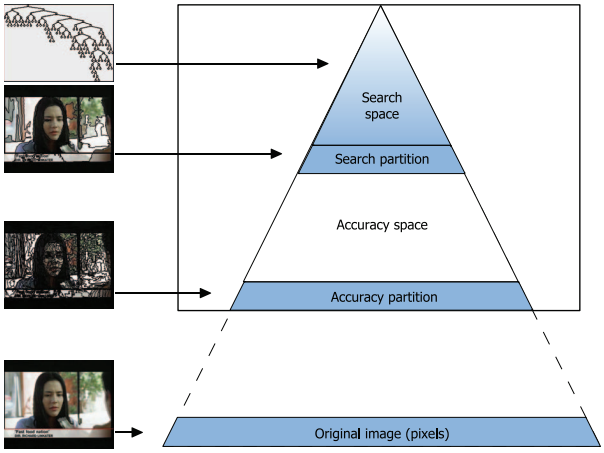

Fig. 1: Region-based hierarchical representation

The BPT represents a set of regions at different scales of resolution and its nodes provide good estimates of the objects in the scene. Using the BPT representation in object detection, the image has to be analyzed only at the positions and scales that are proposed by the BPT nodes. Therefore, the BPT can be considered as a means of reducing the search space in object detection tasks.

In object detection applications, the use as initial partition of a very accurate partition with a fairly high number of regions is appropriate [4]. Since this partition is used to ensure an accurate object representation, it is called the *accuracy partition* (see Fig. 1). Moreover, in the context of object detection, it is useless to analyze very small regions because they cannot represent meaningful objects. As a result, two zones are differentiated in the BPT: the accuracy space providing preciseness to the description (lower scales) and the search space for the object detection task (higher scales). A way to define these two zones is to specify a point of the merging sequence starting from which the regions that are created are considered as belonging to the search space. The partition that is obtained at this point of the merging process is called the *search partition* (see Fig. 1).

In the case of caption text detection, text bars are assumed to be the objects to be detected, and they are extracted by the analysis of the search space. In turn, the subsequent detection of the characters forming the text (consistency analysis for output) requires a more accurate image representation and it is performed analyzing the nodes in the accuracy space. The analysis is restricted to the subtrees defined by the nodes detected as forming the text bars. This last phase will not be detailed since it is not under the scope of this paper.

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. As we have commented in Section 1, these features are commonly translated into two types of descriptors: texture and geometric descriptors 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]. Caption text objects, due to their homogeneous background and regular shape, are likely to appear as single nodes in the BPT. Hence, we propose to combine the two approaches.

In a first stage, texture is estimated over the whole image by means of a multi-resolution analysis using a Haar wavelet decomposition. Texture information is used to prompt highly textured regions (candidate regions) in the BPT. Candidate regions are anchor points for caption text detection. In order to correctly estimate useful descriptors to evaluate candidate regions, their area of support is conformed to the object shape characteristics. Region evaluation is carried out combining region-based texture information and geometric features. Among those nodes in the BPT that have fulfilled this text characteristic verification, final caption text nodes are selected analyzing the various subtrees. The complete process is shown in Fig. 2, where we have selected a simple image (see Fig. 2.a) to illustrate the different steps.

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. Our proposal is to use the power of the LH and HL subbands in a Haar transform (Fig. 2.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. The final mask marking all the text candidates is obtained by performing the union of the (upsampled) masks at each decomposition level (Fig.2.c). For the results presented in Section 4, $l = 2$, $T_1 = 1200$ and $T_2 = 400$. Finally, regions in the search partition (Fig. 2.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. 2.g). This is a very conservative policy but, at this stage, it is important not to miss any possible region containing text (Fig. 2.e).

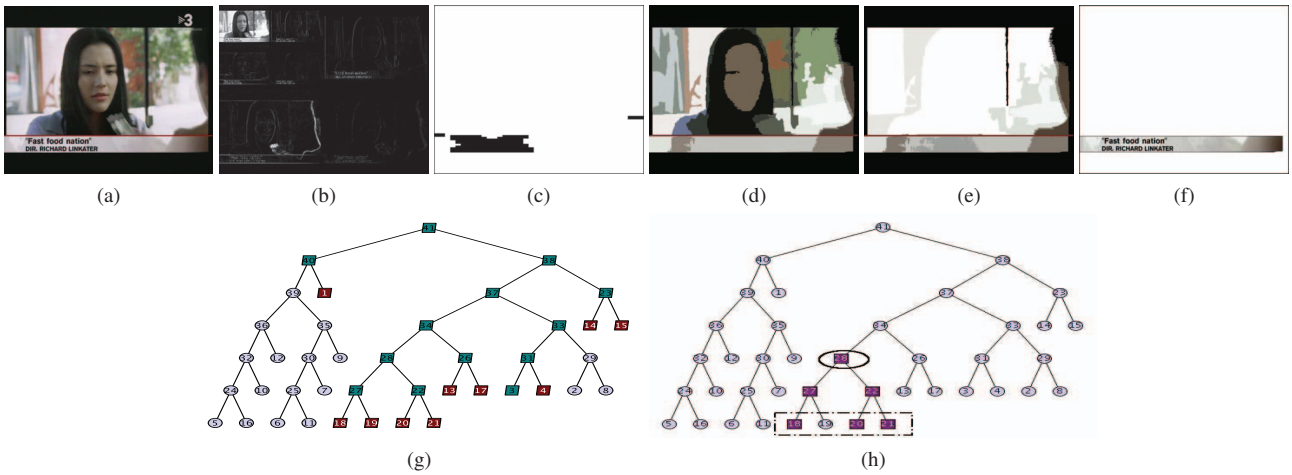

Fig. 2: Example of caption text detection. (a) Original image, (b) Wavelet transform, (c) Text candidate pixels, (d) Search partition (e) Text candidate regions, (f) Final selected region, (g) BPT showing the selected leaves (square nodes in maroon) and their ancestors (square nodes in dark green), (h) BPT showing the verified nodes (square nodes in dark purple), the final selected node marked (with an ellipsis) and the regions to be further analyzed (with a dashed rectangle)

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. This descriptor is used, mainly, to filter out regions that have been selected due to the presence in the mask (see Fig. 2.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.

- *Occupancy:* ($O_{cc} = Area_{Node}/Area_{BB}$) must be greater than $T_{O_{cc}}$; the larger the value, the more similar to a rectangle ($T_{O_{cc}} = 80$).
- *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. 2.h, where the verified nodes are marked in dark purple.

3.3. Tree analysis

As shown in Fig. 2.h, it is common that several verified nodes are in the same subtree; that is, represent the same caption text object. In order to reduce the amount of validated regions to be processed in the consistency analysis for output step and to select the best node, subtrees have to be analyzed.

For each verified node, the branch starting at its position and going up to the root node is analyzed. A confidence value is assigned to every verified node in this branch and, for each branch, the verified node with highest confidence value is selected. Note that conflicts may arise when analyzing two (or more) branches that end in the same node. Conflicts appear when through one branch the highest verified node is selected as the best one whereas, through another branch, the best node is not the highest one but one of its descendants. In those cases, all nodes selected as best ones are kept. Again, this is a conservative policy that ensures keeping all possible information in the process.

The confidence value relies on geometric information:

$$Conf_i = Occ_i^m + \frac{Occ_i^{orig}}{Occ_i^m} - \frac{CC_i - CC_j}{max(CC_i, CC_j)} - \frac{CC_i}{CC_i^{BB}}, \quad (3)$$

where Occ_i^m and Occ_i^{ori} are the occupancy in the modified and original node i, respectively. In turn, CC_i is the compactness of node i and CC_j the compactness of the closest descendant node in the branch. This way, the third term takes into account the evolution of the compactness through the branch. Finally, the last term estimates how far the actual node is from being a rectangle since CC_i^{BB} is the compactness of the bounding box associated to node i.

In the example of Fig. 2 the root node of the subtree is the final selected node. The region associated with this node is presented in Fig. 2.f whereas the final selected node is highlighted in Fig. 2.h.

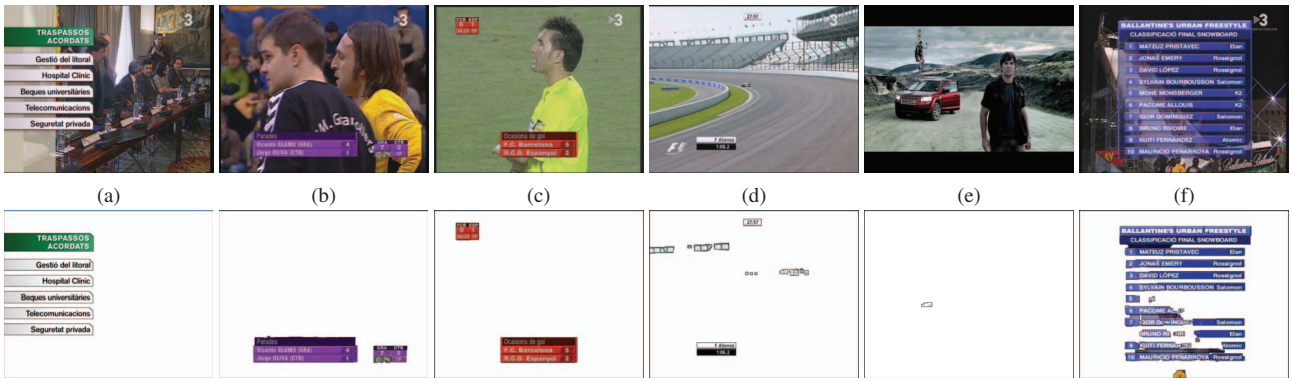

Fig. 3: Illustration of the caption text detection process. First row: Original images; Second row: Final selected regions

4. RESULTS

The technique has been tested in a corpus formed mainly by news and sport event videos[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 (100 images with caption text and 50 without text). 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 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. These two cases are analyzed in Table 2. The number of false positives is 24.

	PDO as outlier	PDO as correct
Recall	0.863	0.950
Precision	0.885	0.894

Table 2: Detection results presented as precision and recall

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 shows a difficult example since region growing yields nodes with leaks due to the color similarity between the caption text object background and the objects around it. The adequation of the area of support previous to computing the geometric descriptors is fundamental in this case. Moreover, in this example a caption text object counted as partial detection is presented. This is the case of one number in the chronometer which is partially missed.

Fig. 3.c illustrates the usefulness of the tree analysis step. Caption text objects are grouped so that, for each group of three objects, there is a single node in the tree merging them. Such nodes have been selected as candidates (rectangular area containing text) but, through the search of the best representative node(s), we are able to detect the three objects separately as single ones.

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 when analyzing the accuracy area in the BPT.

Finally, Fig. 3.f 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.

[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.

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 objets in the scene.

6. REFERENCES

[1] J. Assfalg, M. Bertini, C. Colombo, and A. Del Bimbo, "Extracting semantic information from news and sport video," *Proc. of the 2nd ISPA*, pp. 4–11, June 2001.

[2] 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.

[3] 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.

[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] Y. Zhong, H. Zhang, and A.K. Jain, "Automatic caption localization in compressed video," *IEEE Transactions PAMI*, vol. 22, no. 4, pp. 385–393, April 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] S. Tekinalp and A.A. Alatan, "Utilization of texture, contrast and color homogeneity for detecting and recognizing text from video frames," *IEEE ICIP 2003, Barcelona, Spain*, September 2003.

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

[9] P. Salembier, A. Oliveras, and L. Garrido, "Anti-extensive connected operators for image and sequence processing," *IEEE Transactions on Image Processing*, vol. 7, no. 4, pp. 555–570, April 1998.

[10] M. León, S. Mallo, and A. Gasull, "A tree structured-based caption text detection approach," *Proc. 5th IASTED VIIP*, pp. 220–225, Sept. 2005.

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