AUTOMATIC CAPTION LOCALIZATION IN VIDEOS USING SALIENT POINTS

M. Bertini, C. Colombo, A. Del Bimbo

Dipartimento Sistemi e Informatica
Via S. Marta 3
50139 Firenze, Italy
{bertini,colombo,delbimbo}@dsi.unifi.it

ABSTRACT

Broadcasters are demonstrating interest in building digital archives of their assets for reuse of archive materials for TV programs, on-line availability, and archiving. This requires tools for video indexing and retrieval by content exploiting high-level video information such as that contained in super-imposed text captions. In this paper we present a method to automatically detect and localize captions in digital video using temporal and spatial local properties of salient points in video frames. Results of experiments on both high-resolution DV sequences and standard VHS videos are presented and discussed.

1. INTRODUCTION

Broadcasters are demonstrating interest in building large digital archives of their assets for reuse of archive materials for TV programs, on-line availability to other companies and the general public. To satisfy this request there is need of systems that are able to provide efficient indexing and retrieval by content of video segments based on the extraction of content level information associated with visual data. While effective content-based retrieval of visual information of images is accomplished by supporting content representation through low-level image features, the same does not apply to content-based retrieval of videos, except for very limited application contexts. Instead, effective retrieval of videos must be based on high-level content descriptors [1].

The appeal of digital video archives is motivated by the need to use live feeds and reuse of archive materials: an example of this trend is the EU ASSAVID (Automatic Segmentation and Semantic Annotation of Sports Videos) project the authors participate in (http://www.bpe-rnd.co.uk/assavid/). On the one hand, reuse of archive sport materials is identified as one key method of improving production quality by bringing added depth, and historical context, to recent events (*posterity logging*). On the other hand the broadcaster use footage registered few hours before, that may be even recorded by a different broadcaster, and thus is not indexed, to annotate it in order to edit and produce a sport news program (*production logging*). An example of posterity logging is the reuse of shots that show the best actions of a famous athlete: they can be reused later to provide an historical context. An example of production logging is the reuse of sport highlights, such as soccer goals or tennis match points, to produce programmes that contain the best sport actions of the day.

An effective reuse of archive materials is possible if the shot description is semantically rich enough to allow retrieval by content: a thorough description of the contents allows searching the shots that fit into a request of the video producer. If present, super-imposed text captions, provide a good deal of high–level (semantic) video information, and should therefore be used for automatic annotation of video content. Movies do not usually contain super-imposed text captions, except when they are not dubbed. Captions are frequently used in sport and news videos to annotate the who, where and when of the reported events (news videos) or important highlights (sport videos). Furthermore the analysis of the extracted text may help the automatic classification of the video content: for example a game point of 40-0 is much more probably related to a tennis match than to a soccer match.

In this paper we present a method to automatically detect and localize captions in digital video using temporal and spatial local properties of salient points in video frames. Caption text regions are separated from background images using the local characteristics of corner points; they can be processed further so as to perform traditional OCR processing. Since the sources of videos can vary greatly, as well as the techniques used to produce the captions, independence from any caption style and model is of paramount importance. The method has been tested using several sport videos containing a wide range of different caption styles, containing *background* and *highlights* information about several Olympic games and football matches. The algorithm has also been tested on foreign movies with subtitles. Experiments with both digital and analog video demonstrate

0-7695-1198-8/01/$10.00 (C) 2001 IEEE

the feasibility of our approach for high–level semantic annotation of video materials.

2. PREVIOUS WORK

The problem of automatic text extraction from videos has been investigated by several authors. A method for the extraction of captions and scene texts (e.g. street names or shop names in the scene) from movies has been presented in [5]. Techniques for the extraction and OCR of caption text for news video indexing have been examined in [8]. The first problem that must be solved for effective text extraction is to determine which frames contain captions and then to localize text in the frame. The method presented in [8] is based on searching rectangular regions composed by elements with sharp borders appearing in sequences of frames; it is also based on the assumption that the captions have a high contrast on the background. In [10] a multi–resolution approach was presented that uses directional and overall image edge strength features, which are then classified by a neural feed–forward network. An algorithm that identifies potential text line segments from horizontal scan lines, expanding or merging them to form text blocks was presented in [11]. False text blocks are discarded based on their geometrical properties. In [9], a method was presented to detect and extract text in images through a first step of texture segmentation and a second heuristic step based on the analysis of significant edges and regions. The algorithm proposed in [4] operates on isolated video frames and relies on image decomposition using wavelets classification using with neural network and integration of detected text regions over multiple scales. In [3] RGB color clustering of pixels is performed; connected components are then extracted from the image and heuristic restrictions are used to identify text lines. A method to extract captions from partially decompressed MPEG videos was expounded in [12], using DC components or both DC and AC components, and detecting large inter–frame differences which is quite vulnerable to fast moving objects. [13] has presented a method to detect text in JPEG images and I–frames of MPEG compressed videos, using texture characteristics; this method is relatively insensitive to fast moving objects.

3. THE CAPTION LOCALIZATION SYSTEM

Without loss of generality, let us refer to the broad category of sport videos. In such a category, captions may appear everywhere within the frame, even if most of the time they are placed in the lower third or quarter of the image. Also the vertical and horizontal ratio of the caption zones varies, e.g. the roster of a team occupies a vertical box (Fig. 2), while usually the name of a single athlete occupies a horizontal box (Fig. 1). Character fonts may vary in size and typeface, and may be super-imposed on opaque background as well as directly on the video. Caption often appear and disappear gradually, through dissolve or fade effects. These properties call for automatic caption localization algorithms with the least amount of heuristics and possibly no training. Broadcast quality requirements also impose the use of full TV resolution when archiving videos (720x576 pixels for PAL).

Several features such as edges ([6],[5],[8],[11]) and textures ([9],[13]) were used in the recent past as cues of super-imposed captions. Such features represent global properties of images, and require the analysis of large frame patches. Moreover also natural objects such as woods and leafs, or man-made objects such as buildings and cars may present a local combination of such features that can be wrongly classified as caption [2].

In order both to reduce the visual information to a minimum and to preserve local saliency, we have elected to work with image corners, extracted from luminance information of images. Corners are computed from luminance information only; this is very appealing for the purpose of caption detection and localization in that prevents from many of misclassification problems arising with color based approaches. In fact, since all the television standards require a spatial sub-sampling of the chromatic information, the borders of captions are affected by color aliasing. Therefore, to enhance readability of characters the producers typically exploit luminance contrast, since luminance is not spatially sub–sampled and human vision is more sensitive to it than to color contrast. Another distinguishing feature of our approach is that it does not require any knowledge or training on super-imposed captions features.

The rest of the paper is devoted to the presentation of the technical and experimental details of the approach, starting from corner extraction through corner analysis and caption localization and extraction.

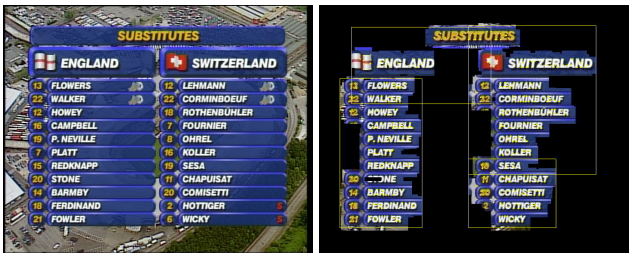
a) b)

Figure 1: (a) Source frame. (b) Detected captions. Olympic games

Corner extraction. The salient points of the frames, that are to be analyzed in the following steps, are extracted using the Harris algorithm ([7]), briefly reported in this section.

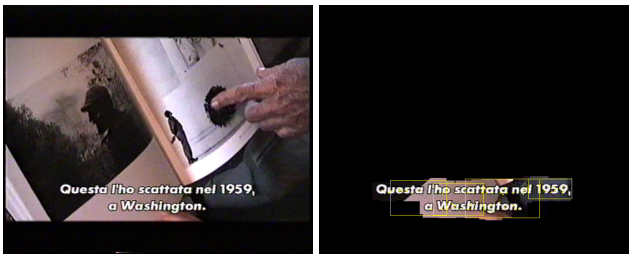
The luminance bitmap I is extracted from each DV frame,

0-7695-1198-8/01/$10.00 (C) 2001 IEEE



a) b)

Figure 2: (a) Source frame. (b) Detected captions. European football championship



a) b)

Figure 3: (a) Source frame. (b) Detected captions. Buena Vista Social Club

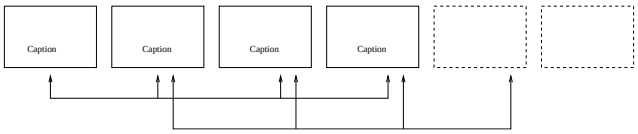
that is stored in *YCbCr* color space. An image location is defined as a corner if the intensity gradient in a patch around it is not isotropic, i.e. it is distributed along two preferred directions. Corners are image points with large and distinct eigenvalues of the gradient auto-correlation matrix, where subscripts denote partial differentiation with respect to the coordinate axis, and brackets indicate Gaussian smoothing:

$$\mathbf{A} = \begin{pmatrix} \langle I_x^2 \rangle & \langle I_x I_y \rangle \\ \langle I_x I_y \rangle & \langle I_y^2 \rangle \end{pmatrix}$$

$$c(x,y) = \det\mathbf{A} - k\,\mathrm{tr}^2\mathbf{A} \text{ with } k = 0.04$$

A corner is detected if $c(x,y)$ is above a predefined threshold. The first term of the equation will have a significant value only if both eigenvalues are different from zero, while the second term inhibits false corners along the borders of the image. The corner extraction greatly reduces the number of spatial data to be processed by the text detection and localization system.

Temporal features. The most basic property of super-imposed captions is the fact that they must remain stable for a certain amount of time, in order to let people read and understand them. This property is used in the first step of caption detection. Each corner is checked to determine if it is still present in the same position in at least 2 more frames within a sliding window of 4 frames (Fig. 4).



a)

Figure 4: Temporal features: sliding window

Each corner that complies with this property is marked as *persistent corner*, and is kept for further analysis, while all the others are discarded. Every 8^{th} frame is processed to extract its corners, thus further reducing the computational resources needed to process a whole video. The choice of analyzing every 8th frame is due to the PAL refresh rate of 25 fps, and means that within a time range of 1 and 1/3 seconds the caption must be stable for 1 second. This feature makes this method completely insensitive to moving objects. Concerning the video standards used, we found it that, since the spatial resolution of S-VHS is about half of the DV source, and the analog source introduced stability noise, when analyzing the video digitized from S-VHS source a corner has been considered persistent if it was present within a 3x3 pixels box (Fig. 3).

Spatial features. For each corner its considered a surrounding patch, and if there are not enough neighbor corners, whose patch intersect it, the corner is discarded from further processing. This process, which is repeated a second time in order to eliminate the corners becoming isolated after the first processing, avoids that isolated high contrast background objects contained within static scenes are recognized as possible caption zones; this also implies checking a second basic property of captions: not having a dominant size within the image.

Noise reduction in still images. An unsupervised clustering is performed on the corners that comply with the temporal and spatial features described above. The bounding boxes are shown in all the figures, and particularly in Fig. 5.

a) b)

Figure 5: (a) Source frame. (b) Detected captions with noise removal. Olympic games

For each bounding box the percentage of pixels that belong to the corner patches is calculated, and if it is be-

0-7695-1198-8/01/$10.00 (C) 2001 IEEE

low a predefined threshold the corners are discarded. This strategy reduces the noise due to high contrast background during static scenes, that typically produce small scattered zones of corners that can not be eliminated by the spatial feature analysis. This strategy is the only heuristic used in our method.

4. SETUP, DATA AND EXPERIMENTS

The system used is an SGI O2; DV video has been acquired using SGI DV Link card and S-VHS video has been acquired using SGI AV1 card. Source videos are acquired from DV source since the DV standard has the lowest acceptable standard for broadcast professionals, and is the video exchange standard adopted within the ASSAVID project. The test set used is composed of 19 sport videos acquired from PAL DV tapes at full PAL resolution and frame rate (720x576 pixels, 25 fps) resulting in 47014 frames (31'20") and about 6 GB of disk space. The videos contained sequences from several Olympic sports (Fig. 1) and European Football Championship (Fig. 2), and are BBC archive material used by members of the EU ASSAVID project. The Buena Vista Social Club video (Fig. 3) has been acquired from S-VHS video (720x576 pixels, 25 fps) resulting in 4853 frames (3'14") and 257 MB. To test the robustness with respect to text size, since some of the captions become 9 pixels high, a short half PAL video (368x288, 25 fps) of 425 frames (17") has been acquired from a S-VHS source.

Evaluation of results takes into account text event detection (whether the appearance of text is detected and text box detection for each text event (Tab. 1). Text event detection has a precision of 80.6%, and recall of 92%, due to missed detections of the VHS video, and to only one missed detection of the DV videos.

	Occurrences	Misdetection	False detection
Text event	63	5	9
Text boxes	100	5	37

Table 1: Text event detection and text boxes localization

The text box miss rate is 5%, due to the above mentioned misses of the VHS video, and 1 missed block of the DV video. 6 false detections are due to scene text.

5. CONCLUSIONS AND FUTURE WORK

This paper presents a novel method to automatically detect and localize captions in digital video using local temporal and spatial properties of salient points in video frames. Caption text regions are separated from background images using corner characteristics. The system does not require any training or information about the caption style, and is able to cope with captions with different styles such as with and without opaque background, without need of heuristics. The system is also capable of detecting very small captions. Future work will deal with further improvements of rejection of falsely detected caption zones, use of local properties of corners to enhance OCR, and OCR pre-processing of the extracted captions.

Acknowledgment. This research was supported in part, by the European Union under contract IST-13082 (ASSAVID project).

6. REFERENCES

[1] A. Del Bimbo C. Colombo and P. Pala. Semantics in visual information retrieval. *IEEE Multimedia*, 6(3), 1999.

[2] David Doermann Huiping Li and Omid Kia. Text extraction, ehnancement and ocr in digital video. *Proceedings of the International Workshop on Document Analysis Systems*, 1998.

[3] Anil K. Jain and Bin Yu. Automatic text location in images and video frames. *Pattern Recognition*, 31(12):2055–2076, Dec. 1998.

[4] Huiping Li and David Doermann. Automatic identification of text in digital video key frames. *Proceedings ICPR 1998*, 1, Aug. 1998.

[5] R. Lienhart. Indexing and retrieval of digital video sequences based on automatic text recognition. *Fourth ACM International Multimedia Conference*, 1996.

[6] T.Kanade M.Smith. Video skimming and characterization through the combination of image and language understanding techniques. *IEEE CVPR*, 1997.

[7] J.M. Pike and C.G. Harris. A combined corner and edge detector. *Proceedings of the fourth Alvey Vision Conference*, pages 147–151, 1988.

[8] T. Sato, T. Kanade, Ellen K. Hughes, and M. A. Smith. Video ocr for digital news archive. *IEEE International Workshop on Content–Based Access of Image and Video Databases CAIVD' 98*, pages 52–60, 1998.

[9] Raghavan Manmatha Victor Wu and Edward M. Riseman. Textfinder: An automatic system to detect and recognize text in images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(11):1224–1229, Nov. 1999.

[10] Axel Wernicke and Rainer Lienhart. On the segmentation of text in videos. *Proceedings ICME 2000*, Aug. 2000.

[11] Edward K. Wong and Minya Chen. A robust algorithm for text extraction in color video. *Proceedings ICME 2000*, Aug. 2000.

[12] B.L. Yeo and B. Liu. Visual content highlighting via automatic extraction of embedded captions on mpeg compressed video. *SPIE Digital Video Compression: Algorithms and Technologies*, Feb. 1995.

[13] Hongjiang Zangh Yu Zhong and Anil K. Jain. Automatic caption localization in compressed video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(4):385–392, Apr. 2000.