Text Detection and Recognition for Person Identification in Video

Johann Poignant, Franck Thollard, Georges Quénot, Laurent Besacier

Text Detection and Recognition for Person Identification in Videos

Johann Poignant, Franck Thollard, Georges Quénot and Laurent Besacier

UJF-Grenoble 1 / UPMF-Grenoble 2 / Grenoble INP / CNRS, LIG UMR 5217, Grenoble, F-38041 France

Firstname.Lastname@imag.fr

Abstract

This article presents a demo of person search in audio-visual broadcast using only the text available in a video and in resources external to the video. We also present the different steps used to recognize characters in video for multi-modal person recognition systems. Text detection is realized using the text features (texture, color, contrast, geometry, temporal information). The text recognition itself is performed by the Google `Tesseract` *free software. The method was successfully evaluated on a broadcast news corpus that contains 59 videos from the* France 2 *French TV channel.*

1. Introduction

Accessing information in videos is challenging due to the so-called "semantic gap". Addressing this problem can be done by taking advantage of different modalities that are, explicitly or implicitly, present in the video. Recognition of persons in video documents was initially a mono-modal problem (face recognition, speaker recognition). As other information can help recognizing persons, the problem is now addressed through a multi-modal fusion strategy. Following this trend, we aim at extracting the text available in the video frame with a focus on named entities.

The recognition of text embedded in a video – like *e.g.* a name, a place, a position – can provide information about the presence or citation of a person.

The availability of this text can help us to disambiguate a name that was cited or vice versa. It can help to understand the context of the video, or the main topic of a report. Taking advantage of this extracted text could be done using a classical text based Information Retrieval (IR) system, that could be enhanced using external information such as electronic newspapers, web pages related to specific keywords (*e.g.* people, locations, ...).

Several steps are needed to obtain this information. First, the detection of the text area must be done accurately, second, an Optical Character Recognition (OCR) system is applied on the parts of the images selected and filtered during the previous step. A final post-processing step is then performed. As the processes are cascaded, the quality of a given one (*e.g* the text detection) has a great influence of the following ones (*e.g.* the recognition step). We concentrate in this paper on the first step.

The paper is organized as follows: section 2 presents related works. Our contribution is then detailed: text detection, Optical Character Recognition, a light post-processing, use of external information. Section 4 addresses the problem of the experimental evaluation on a broadcast news corpus taken from the *France 2* French TV channel. The last section is dedicated to the conclusion.

2. Related works

All the methods share three steps: text detection, text recognition and post-processing.

Regarding text detection two kinds of texts has to be considered: the scene text (*e.g.* a text written on a T-shirt), and the overlaid text. When detecting the latter, the technique will change if the text is known to be horizontal [6], or can be in any orientation; in such a case corner detection can be used as in [5]. In our case, the text is known to be horizontal and the closest work is the one from Wolf *et al.* [8] which propose a detection scheme's pressing the cumulative measure of directional gradient. Our detection procedure will follow quite closely this work.

Regarding the text recognition problem, two solutions are possible: using a conventional OCR system with image adaptation or produce a specific video one. For the second, we refer the interested reader to [1]. In our case, we will rely on an external OCR system as the text we aim at recognizing follows some strict rules: horizontal overlaid text, fixed font and fixed position. We will thus rely on the `Tesseract` free software from Google. In order to optimize the performances of `Tesseract`, images have to be

scaled to a resolution similar to the one of scanned images.

For the last step, namely the post-processing, the string edit distance can be computed between the obtained string and a pre-defined corpus [7]. Another post-processing can be done using external resources: Zhao *et al.* [9] have worked on multimodal fusion for people recognition using, not only conventional recognition but also OCR and Automatic Speech Recognition (ASR) systems. In addition to these various methods external resources were used to improve recognition: a corpus of text (namely AQUAINT), some news site and search engine. Merging was done using the RANKBOOST algorithm [2].

3. Our contribution

We focus our study on the transcription of named entities from the video frames. Proper names and positions of a person have been our first target to highlight the interest of video text to recognize people. To do this, we had to design a detection system with several steps (Fig.1): From a frame (Fig.1a), a Sobel filter is applied to detect character edges. The obtained images are then thresholded (Fig.1b), then treatment of dilatation and erosion connects the characters together (Fig.1c). Most of the noise is then suppressed using an erosion followed by a dilatation (Fig.1d).

(a) Original image (b) Edge detection

(c) Character connection (d) Noise suppression

Figure 1. Images taken from the France 2 news of February 1, 2007, INA source

After this coarse filter, a more accurate corner detection is performed on each connected component: a horizontal (Fig. 2b) (resp. vertical (Fig. 2c)) dilatation allows to detect the box height (resp. the box width). The connected components that do not hold a mandatory geometry are filtered. Fig. 2d shows a resulting image.

(a) A box after first detection

(b) find abscissa with vertical dilatation

(c) find ordinate with horizontal dilatation

(d) Image with boxes

Figure 2. precise detection of the box coordinates

We perform the detection on each frame in order to have a follow up of each text box (Fig. 3). Only the boxes sufficiently stable over time are kept. We thus have, for each kept box, a range of presence (starting/ending frame). The box images are then processed by the Otsu algorithm for binarization. Every 10 frames, an average images is computed. Since the text lasts more than 10 frames, we have many candidates images for the same text. Each of these candidate image is sent to `Tesseract`, thus leading to many transcriptions for the target text. Note that `Tesseract` being quite sensible to the resolution of the input images, a Bi-cubic interpolation is applied to obtain the required resolution.
Last but not least, an average image over the whole appearing / disappearing range is computed. This image will serve as a reference for the combination step: a weighted lattice is built using all the transcriptions provided by `Tesseract`. A viterbi search outputs the path with the maximum weight, thus giving the most likely transcription.

4. Evaluation issues

The corpus used for evaluation consists of 59 videos of France 2 TV News from 1 February to 31 March 2007. The average length of these videos is about 38 minutes, which represents an overall of 37 hours of video. 29,166 key frames were extracted by segmentation of videos. The texts extracted from these key frames have been manually annotated. We also annotated if the text corresponds to a person's name or to the person's position and if the person was present or not on the image.

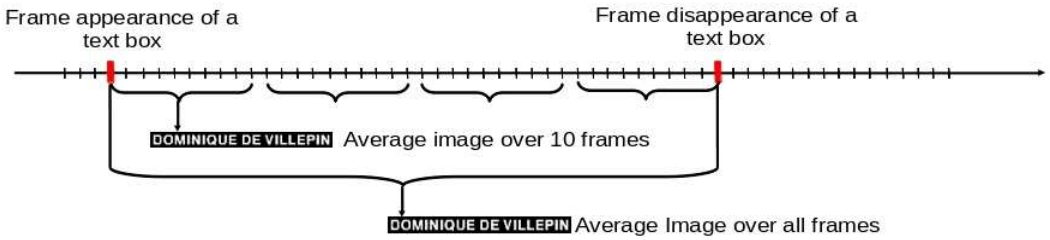


Figure 3. Temporal detection

We conducted our evaluation on 29k frames, with focus on person name and person position written on the screen (excluding those credited late reporting). Spatial position of the detected text has not been evaluated as it requires a more complex annotation setting. We nevertheless annotated presence/absence of the boxes in the annotated frames. From the 29k images assessed, 4,414 frames containing text with 9,257 text boxes. The performance of our detection text system is 91% recall. Although a bit weak, this result can be improved by a better tuning[1] of the system; an annotation that specifies the coordinates of the text boxes should also allow us to improve this result. A simple ad-hoc post-processing was applied to the recognized texts to correct some errors (for example, changing "ii" into "m" in the `Tesseract` output, ...).
We used as a metric for assessing the recognition of texts, the Levenshtein distance, given by the `Sclite`[2] tool, on the words and characters. This distance is calculated between the reference text (in our case, the human post-edition of the recognized text) and the hypothesis (the text automatically recognized by our system).

Results are presented in Table 1 and examples of such texts in Fig 4. For sake of completeness we provide the word error rate even through the words were not yet post-processed. The error mentioned is thus over-estimated. Consequently, we focus our analysis on the character error rate. As can be seen, the names of people tend to be more readable by our system (line FT, *vs* N). In the corpus we use, the names are often written with a font of 10 pixels high where the person position is usually written with a font of 7 pixels high. This can explain the difference in performance between the name only category (N: 3.7%), and the position (P: 8.4%). Moreover, one can see that the NO performances are lower than the N ones (5.0% *vs* 3.7%). It appears that the places in the video where the name does not appear alone are mainly in the news summary and are consequently written in lowercase. Moreover, in these video parts, the image is much less stable than in the main track.

[1] At the moment a non systematic tuning of the system is manually performed.

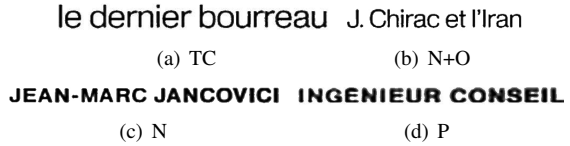
[2] Sclite: http://www.icsi.berkeley.edu/Speech/docs/sctk-1.2/sclite.htm

Table 1. Results for character recognition

Type	boxes	words	character	err rate words	err rate character
FT	9257	30912	154941	21.1%	9.8%
N	1440	3230	19248	8.3%	3.7%
NP	1394	3126	18589	8.5%	3.7%
NO	1683	4263	23984	11.9%	5.0%
NOP	1491	3566	20484	10.7%	4.6%
P	1394	5794	32472	27.5%	8.4%
PP	1360	5657	31717	27.7%	8.5%

FT: Full Text;
N: Name appears alone (without credit).
NP: N and person present a priori (alone or accompanied) in the frame.
NO: Name appears alone or with other words (without credit).
NOP: NO and person present a priori (alone or accompanied) in the frame.
P: Person position present.
PP: P and person present a priori (alone or accompanied) in video.

When a name is written alone on a box text, the person is present in 96,8% of cases.



(a) TC (b) N+O

(c) N (d) P

Figure 4. Example of localized texts

5. Demonstration

We developed a search person system in video based only on the text. When a user type a person name, or a position (job, title ...) and a date (between February 2 and March 31, 2007) the system localizes a part of a newscast from the France 2 corpus, where the person is likely to appear on the screen.

At indexing time, the system will cascade different steps: text detection followed by character recognition. This pro-

(a) (b)

Figure 5. Text localization / recognition in a video frame

vides a set of words used to find out complementary information from external ressources. We end up with a list of persons, functions, dates and places which will be used to accurately localize persons in the video. At query time, the system will be able to display relevant video parts related to the text query.

6. Conclusion and further work

This work allows us to assess the contribution that OCR can provide person name detection from video. Subsequently, the study of other types of named entities (place, date, ...) can still provide useful information for recognizing people. In our future work, we also plan to integrate some control of the `Tesseract` software (like post-processing using word lists and language models) in order to improve the quality of the transcriptions. The last question to address is the use of the obtained information. Currently, the text cannot be used directly by a classical IR system as a document (*i.e* the text in the box) has a non-classical size (only a few words). Indexing such small documents can be done using methods inherited from Speech Information Retrieval, in which trigram of characters are indexed instead of full words (as for example in [7] or [4]). Another strategy can be to directly match the query terms with the text output by our system. In order to integrate some tolerance, the match can be done fuzzily using, for example, the Levenshtein distance through some efficient Nearest Neighbor [3].

Acknowledgements: This work was partly realized as part of the Quaero Programme, funded by OSEO, French State agency for innovation.

References

[1] F. Einsele, R. Ingold, and J. Hennebert. A HMM-Based Approach to Recognize Ultra Low Resolution Anti-Aliased Words. In Springer, editor, *In Second International Conference on Pattern Recognition and Machine Intelligence (PReMI 2007)*, pages 511–518. Dcembre 2007.

[2] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. In *Proc. of EuroCOLT'95*, pages 23–37, 1995.

[3] E. Gòmez-Ballester, L. Micò, F. Thollard, J. Oncina, and F. Moreno-Seco. Combining Elimination Rules in Tree-Based Nearest Neighbor Search Algorithms. In E. R. Hancok, R. C. Wilson, T. W. Ilkay, and F. Escolano, editors, *Structural, Syntactic, and Statistical Pattern Recognition*, pages 80–89, Cesme, Turkey, Aug 2010. Springer.

[4] S. Harding, W. B. Croft, and C. Weir. Probabilistic Retrieval of OCR Degraded Text Using N-Grams. In *European Conference on Digital Libraries*, pages 345–359, 1997.

[5] X.-S. Hua, X. rong Chen, L. Wenyin, and H.-J. Zhang. Automatic Location of Text in Video Frames. In *Proceeding of ACM Multimedia 2001 Workshops– Multimedia Information Retrieval (MIR2001)*, pages 24–27. ACM Press, 2001.

[6] R. Lienhart. Automatic text recognition for video indexing. In *MULTIMEDIA '96: Proceedings of the fourth ACM international conference on Multimedia*, pages 11–20, 1996.

[7] K. Takeuchi and Y. Matsumoto. Japanese OCR Error Correction Using Stochastic Morphological Analyzer and Probabilistic Word Ngram Model. *International Journal of Computer Processing of Oriental Languages*, 13(1):62–82, Mars 2000.

[8] C. Wolf, J.-M. Jolion, and F. Chassaing. Text Localization, Enhancement and Binarization in Multimedia Documents. In *ICPR*, pages 1037–1040, 2002.

[9] M. Zhao, S.-Y. Neo, H.-K. Goh, and T.-S. Chua. Multi-faceted contextual model for person identification in news video. In *Multi Media Modeling*, 2006.