Conference & Workshop on Assistive Technologies for People with Vision & Hearing Impairments
Assistive Technology for All Ages
CVHI 2007, M.A. Hersh (ed.)

OCR-ALGORITHM FOR DETECTION OF SUBTITLES IN TELEVISION AND CINEMA

Morten Jønsson[1], Hans Heinrich Bothe[2]

[1]Informatics and Mathematical Modelling,
Technical University of Denmark (DTU).
Tel: (+45) 21204719. Email: s021724@student.dtu.dk
[2]Centre for Applied Hearing Research (CAHR), Oersted DTU,
Technical University of Denmark (DTU).
Tel: (+45) 45253954. Fax: (+45) 45880577. Email: hhb@oersted.dtu.dk

Abstract: The OCR-algorithm (optical character recognition) described in this paper is a module in the assistive device SubPal, which should be able to read subtitles from television and camera aloud. The SubPal system is described in detail in (Nielsen & Bothe, 2007). By sampling the television signal (PAL) a binary image is created. This binary image is analysed using the OCR-algorithm for generating text-strings that can be passed on to a speech synthesis box. The requirements and the implementation of the OCR are discussed and some initial results are presented. The algorithm is developed with the purpose of later being implemented in hardware (FPGA).

Keywords: optical character recognition, subtitles, visually impaired people, dyslexic

1. Introduction

Interviews with both visually impaired people and dyslexic have revealed, that a large group is cut off from full understanding of the visual media (television, DVD, VHS, cinemas), when this is given in a language that they are not comfortable with. Even if a visual impaired person is able to grasp the overall content of the screen they are unable to read the subtitles. A solution to this is presented in the paper (Nielsen & Bothe, 2007), from this it is clear that a versatile, fast and robust OCR (optical character recognition) is necessary. For a sufficient detection speed this OCR should be implemented in hardware (FPGA). Since a commercial hardware OCR is not on the market, we will in this paper show the initial steps towards such an OCR algorithm.

2. OCR Requirements

Before going into detail with the modules of the OCR, we asses the overall requirements that should be considered with respect to the application (described in (Nielsen & Bothe, 2007)).

- Robustness – each character in the subtitles should be detected with high detection rate. When a character is not detected correctly the word should still be detected by using look-up into a dictionary and selecting the best match.
- Speed – the detection rate for one word should be proportional to the response-time of the speech synthesizer.[1]

[1] Implying that the speech synthesizer meets the time requirements imposed on the system.

- Adaptive – the font differ from channel to channel depending on the subtitling company. The algorithm should be versatile enough to encounter for this. Noisy backgrounds (e.g. white T-shirt behind subtitles), should not reduce the detection rate significantly.
- Orientation – the spatial orientation of the image should be transparent to the algorithm.

Commercial OCR-solution have been surveyed, but the majority of the available OCR-solutions targets the software market and depends on a specific operating system and hardware-architecture, which imposes additional overhead with respect to performance. The OCR discussed in this paper is intentioned for the portable device described in (Nielsen & Bothe, 2007), and must comply with the necessary speed of response implying that a special purpose hardware is feasible e.g. a FPGA[2]. Since we have not found such on chip OCR-solution, we will instead develop this from scratch by first looking into optimal character recognition algorithms which is the objective of this paper. Further studies are left for implementing the algorithm on an FPGA. Although the existing commercial OCR-solutions are not relevant for this device, we can use them as a benchmark to compare how effective the developed algorithm is.

In the following sections we will look into the design of such an OCR.

3. Overview of OCR Modules

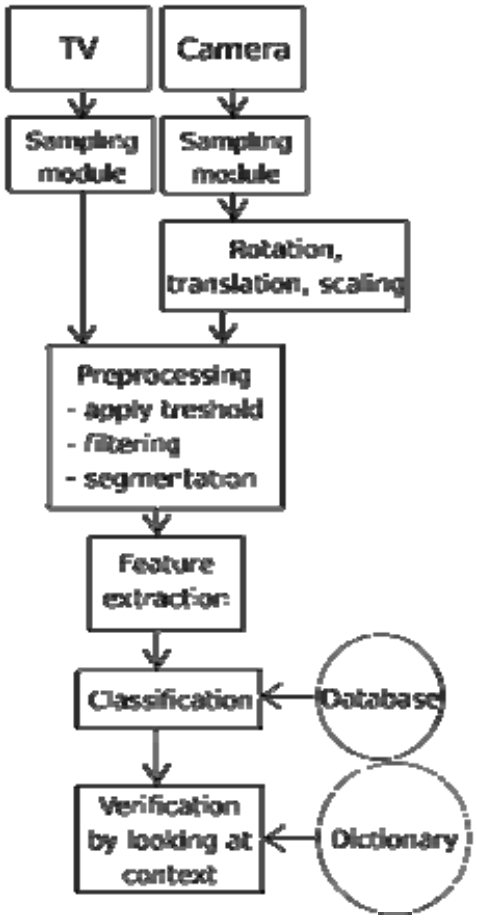

Figure 1:Modules in OCR system

The task of recognising characters in television as well as cinema can be divided into the modules illustrated in figure 1. This division is consistent with the standard approach used in OCR-systems (Trier & Torfinn, 1996).

First of all the raw signal from the composite video signal is sampled to create a binary image, from which the subtitles can be extracted. When using the signal from the camera it is also necessary with some spatial adjustment to ensure that the lines of text are horizontal in the image, which is a precondition for our OCR-algorithm. Next the binary image is prefiltered to remove noise and enhance characteristic features. After the optimal filtering the image can be divided into separate lines, words and letters. Each of these characteristic features, statistical or semantic, are detected and compared with an already existing database (based upon training set). After choosing the most likely letters in a given word, the word can be compared with a dictionary lookup, to verify if the letter combination is likely.

Each processing step will be explained in more detail in the following sections.

[2] Field Programmable Gate Array

4. Sampling Module

The images used for the character recognition are created by sampling the composite signal from the television/video camera using a Tektronix TDS1002 oscilloscope and applying a threshold. This is described in more detail in (Nielsen & Bothe, 2007). An example of the resulting binary image is shown figure 2.

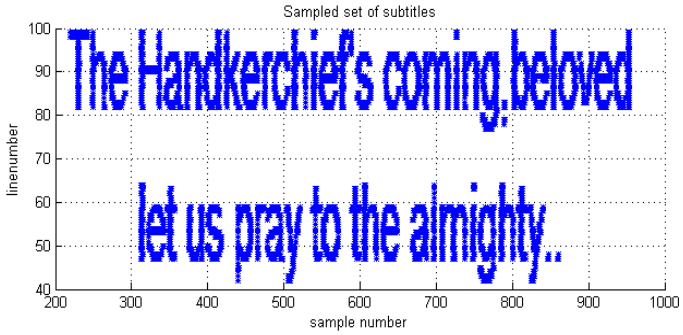

Figure 2: When sampling and applying a binary threshold the binary image is created

5. Spatial Adjustments

In the case where the images come from the CCD video camera, it will often be necessary to carry out some minor spatial adjustments, since our later described feature extraction is not rotational invariant. The subtitles consist of one or two lines of densely packed letters with the same orientation. This a'priori knowledge can be used for finding the rotation of the image, which maximizes the horizontal sum in the frame (corresponding to a horizontal orientation of the subtitles). Another approach would be to use a rotation invariant feature extraction, such as Transformation Ring Projection (Tang, 1991).

6. Preprocessing

The success of the OCR-algorithm depends on the initial filtering. The aim of this filtering and segmentation is to separate the text into separate letters and at the same time make sure that each letter is as characteristic as possible. To begin with we lowpass-filter the image to remove high frequency noise which was not removed in the sampleprocess, further more we use dilation[3] to avoid that letters are being divided into several regions (see (Carstensen, 2002), (Horn, 1986) for further details and figure 4 and 5 for illustration). The dilation is done with a structuring element consisting of two horizontal pixels. This ensures that errors that would separate letters into several regions are corrected, without influencing the characteristic appearance of the letter considerably (illustrated in figure 3).

With the optimal preconditions given the lines can then be separated by simply detecting minimas in the horizontal projections. By looking at the vertical projections instead each line can be separated into words and letters (see figure 6). After the segmentation each letter is mapped into fixed height and width, thereby making it comparable with the letters in the database, when extracting the features.

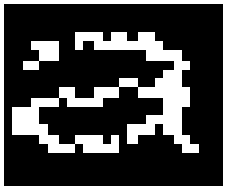

Figure 5:Letter before filtering

Figure 4:Letter after lowpass-filter and dilation

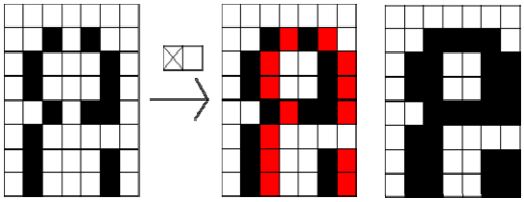

Figure 3: Principle behind dilation using 2 pixel structural element

[3] A binary morphology method

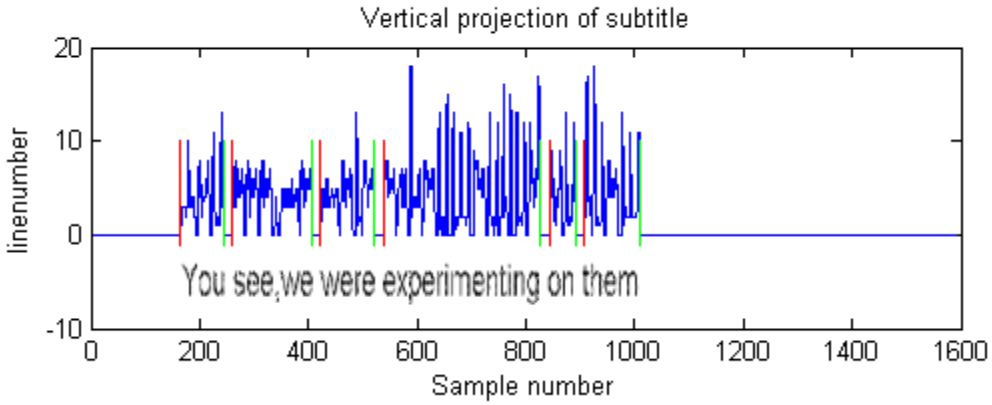

Figure 6: Example of how a line can be separated into words using vertical projection

7. Feature Extraction

To compare each region with our database we need to extract relevant features. We use a combination of simple statistical and semantic features, which minimize the amount of calculations. The statistical features are relations between area, width/height, background/foreground and the 1st order moment. Furthermore the horizontal and vertical projections (see figure 9) are used. A way to reduce the data from the horizontal and vertical projection is to do a Fourier transform of the rowsums as described in (Bourbakis, 1991), but with our low resolution it is sufficient to use the sums directly. Before calculating the projections (row- and column-sums) the letter is mapped into a fixed size using bicubic interpolation (Carstensen, 2002).

Our semantic method detects holes, feet, heads and arms in the letters as illustrated in figure 8. An improvement could be to further more use feature point extraction for detection of intersections and corners (see (Brown, 1992) for details).

8. Classification

The horizontal and vertical projections are compared with the database using crosscorrelation. When comparing the statistical features such as height/width-ratio, foreground/background, area and first order moment, a normal distribution is assumed. For each letter the mean of this distribution is given in the database, while the variance is chosen for optimal detection. An extracted feature such as the pixelarea can now be checked towards the distribution of each letter and a probability can be calculated (see figure 7). The results from the projections and the other statistics is finally combined and the most likely letter is chosen. Afterwards the semantical features are used for correction of the most likely misclassifications.

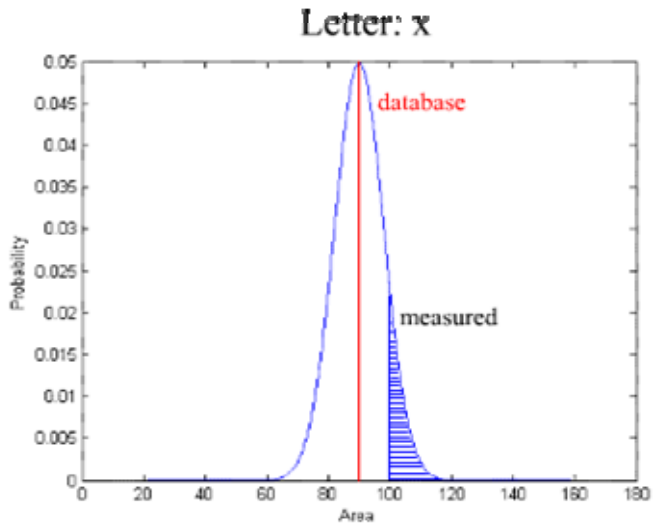

Figure 7: Example of how the area-feature of a unknown letter is compared with the letter "x" of the database

9. Verification

Even with a good detection rate of each letter, errors will occur from time to time. To cater for these, each word is checked with a dictionary, containing the most common words. If the word is not present, the most likely match is chosen, e.g. by finding the word with most letters in the correct position. It is then evaluated which of the 2 words is the most likely, based on the probabilities from the classification. The dictionary lookup is done using the large English database WordNet® (see reference (Anon, 2006) for information on license). It contain more than 200.000 nouns, verbs, adjectives and adverbs in all conjugations. e.g. Run, ran, running, runs, house, houses, housing ... We combine this with a database containing words not included in WordNet®, such as pronouns, prepositions etc.

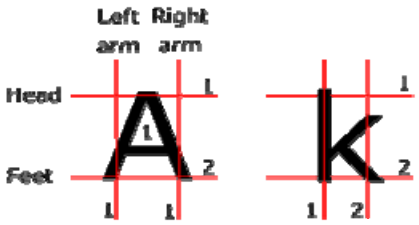

Figure 8: Illustration of semantic features

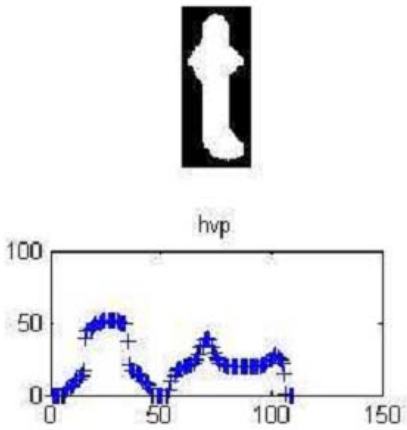

Figure 9: Plot of horizontal and vertical projection for a t

10. Conclusions and Future Work

With the binary images created by sampling of the television signal using oscilloscope, we get a letter detection rate of 95%, when using the same type of font in the trainingset and the testset. By using a commercial OCR[4] on a resized version of the subtitles we obtain a detectionrate of 96%. With the present solution the database is created using only one type of subtitle font and is therefore only optimal when the same type of font is presented, while the commercial approach produces similar results independent on the font.

The aim of these considerations is to obtain an implementation in hardware for optimal response times and optimal power consumption. Further work is necessary to make the OCR more versatile, fulfilling the requirements laid out in section 2. A possibility for doing this could be to implement the classification by using an associative neural network.

References

Anon (2006). *WordNet 3.0*, Princeton University, http://wordnet.princeton.edu/

Brown, E.W. (1992). Character recognition by feature point extraction, Northeastern University internal paper

Bourbakis, N.G and A. T. Gumahad, II (1991). Knowledge-based recognition of typed text characters, *International Journal of pattern Recognition*, vol. 5(1-2), pp. 293-310.

Carstensen, J.M (2002). Image analysis, vision and computer graphics, Technical University of Denmark

Horn, B.K.P (1986) *Robot Vision*, The MIT Press

Nielsen, S. and H.H. Bothe (2007) SubPal: A device for reading aloud subtitles from television and cinema, *CVHI conference*.

Tang, Y.Y (1991) Transformation-ring-projection (TRP) algorithm and VLSI implementation, *International Journal of Pattern Recognition*, vol. 5(1-2), PP. 25-56.

Trier, Ø.D, A.K. Jain and T. Torfinn (1996). Feature extraction methods for character recognition – a survey, *Pattern Recognition,* vol. 29 (4), pp. 641-662.

[4] Abbyy FineReader 8.0 Professional edition. http://buy.abbyy.com/content/frpro/default.aspx