Q3 2013

VIDEO BASED OCR: A CASE STUDY OF REAL TIME IN-SCREEN SUBTITLE RECOGNITION

MARCO SLIK, HANS JONGEBLOED & MARK VAN STAALDUINEN
NPO

ABSTRACT

NPO, the Netherlands public broadcasting organization, has implemented an Optical Character Recognition (OCR) solution for real time conversion of in-vision subtitles to spoken subtitling audio.

In the Netherlands, television programmes in foreign languages are subtitled. NPO offers access services for the visually impaired in the form of spoken subtitling: speech-synthesized audible text. For this to work correctly in the past, pre-produced subtitle files or embedded data within the live studio feeds were necessary. Programmes are now being contributed more and more with subtitles already present in the active video - the so-called ‘title-copies’ or ‘burned-in subtitles’ in former technical jargon. The percentage of spoken subtitled programmes has thus clearly shown a decreasing trend, just when the political call for access services for the visually impaired has been growing louder.

NPO needed to find a solution and teamed up with Dutch research organisation TNO, the postal-market technology provider Prime Vision, and NPO’s playout facility provider Ericsson Broadcast Services, to work towards solving this problem. With success…

INTRODUCTION

NPO, the Netherlands public broadcasting organization, has implemented a solution for real time conversion of in-screen subtitles to closed caption data. This means that 'burned-in' subtitles in the video picture are converted to a separate data stream containing the subtitle text.

THE USE CASE: SPOKEN SUBTITLING IN THE NETHERLANDS

In the Netherlands, about 430,000 people have serious sight problems. As a result of the ageing population, this number is growing; approximately 76,000 people are now blind or near blind.

For the visually impaired, television is slightly subsidiary to other forms of media consumption. However, it plays an equally important role, as it does for the seeing audience. Television is an important way to keep up with current affairs. The social value of watching together and the social interaction produced by a programme must also not be underestimated. Despite a different dimension and way of 'watching' (which is the term blind people use themselves), the viewing habits of the visually impaired audience nearly match those of the non-visually disabled.

On Dutch television, foreign language programmes or programme-items are mostly subtitled. Post-synchronization/audio-dubbing is only common practice for children's programming up to twelve years of age. For the viewing audience, the feeling with the original programme is completely maintained. For the hearing impaired, subtitling is an ideal way of watching TV. For the visually disabled audience however, dubbing would have been more helpful.

Though not required by legislation, NPO as a public broadcasting organization, is under political pressure to broadcast the programmes on NPO's main channels with subtitle data streams[1]; in order to allow visually impaired viewers to hear spoken Dutch language text through a speech synthesis solution (i.e. automatic conversion of text to speech). This automated so-called 'spoken subtitling' service has been operational in the Netherlands since 2001. It is a cost-effective way of serving the sight disabled audience, compared with the 'audio description' method that is used in other countries and which necessitates manual authoring before broadcast. If certain preconditions are met, spoken subtitling can achieve 100% programme coverage.

The end-user can receive the spoken subtitling audio in the following ways:

- As a DVB language track in digital distribution, offering a spoken subtitle audio overlay over the stereo programme audio. The programme audio is lowered in volume in a rather neat way when subtitle text is presented; the so-called 'broadcast mixed' approach. This signal is suitable for users capable of handling the relatively complex (from target audience perspective) set-top box operation. This is especially useful for the large group of dyslectics or viewers with other reading disabilities.
- Through a visual aid; e.g. a separate receiver box, decoding the subtitle data offered through teletext page 889 or an internet data feed. Currently, the costs of two commercial aids are reimbursed through the Dutch medical insurance system. These aids also offer other functionality like spoken newspapers, books and magazines. Ease of operation, including audio feedback, makes this the most suitable solution for seriously visually impaired 'viewers'. The spoken subtitle audio is provided from a separate loudspeaker, enabling other members of the family to listen to the regular programme sound.

The timing difference between speech and subtitles is normally less than a few TV frames' duration. Whereas, for the internet solution, where the hard timing relation with the broadcast video

[1] For the hearing impaired, the Dutch Media Act requires 95% visible subtitling for Dutch spoken word programs. This is offered through Teletext page 888 and DVB subtitling, in addition to the in-vision foreign language translating subtitles. Political expectations for spoken subtitling are (without particular legal grounds) set on the same basis as the legislation on subtitling for the hearing impaired.

distribution is lost, a user-adjustable, streamside delay is used to compensate for the difference in distribution paths (satellite, terrestrial, cable and IPTV).

THE ISSUE

Unfortunately, the subtitles needed for data distribution and speech conversion are available as separate subtitle files for only 70% of programmes. For the following reasons, this percentage is even decreasing:

- Daily news shows and live current affair programmes tend to imprint the subtitles in the video signal more and more for reasons of flexibility in last-minute-before-air editing.
- Manual subtitle operation in live studio production is being reduced to a minimum, because of budget restrictions. File-based items produced within the studio are nowadays expected to be fully broadcast ready.
- There is a tendency towards cross-media delivery during or before broadcast. Dependency on live broadcasting has to be minimized, and maximum flexibility in versioning for different distribution outlets is desired.

OPTICAL CHARACTER RECOGNITION (OCR) SOLUTION APPROACH

The solution for which NPO has been searching, is how to recognize subtitle text automatically from within the final video product. NPO cooperated with TNO (the largest independent scientific research organization in the Netherlands since 1932) and Prime Vision (a company specializing in OCR, mainly for postal automation), to work towards an automated conversion tool. Ericsson Broadcast Services, NPO's playout facility supplier, joined the team in the proof of concept phase and streamlined the practical implementation within the playout system.

The requirements during the project definition phase were formulated as follows:

- The system should support SD and HD resolution video.
- File-based and live programme workflows should both be covered.
- The solution's architecture should be simple and efficient. Complicating the (digital) workflow should be avoided as much as possible.
- As it is expected that OCR cannot match a 'good old' text based data feed, the data stream from the subtitling equipment should take precedence if present.
- 99% of the OCR-ed words should be correctly converted into text.
- Only Dutch subtitles should be converted
  (thus excluding subtitles for minority languages, such as Arabic, Turkish, etc., that would demand extra font and language sets and substantially complicate the project).
- Punctuation marks should be read as well.
- The OCR solution output data should be available within 300ms after the first frame in which the subtitle appears, to retain acceptable speech-to-subtitle synchronization throughout the full broadcast chain.

Because of the live programme requirement, the choice was quickly made for a near real time system approach in the playout video stream.

SELECTION OF IMPLEMENTATION PARTNER

TNO started by looking for the right business and implementation partner. Given the fact that video OCR is rather new, TNO was not expecting to find a commercial off-the-shelf product that would fit the target requirements. A request for information to ten potential vendors resulted in four answers, from which only two showed a proper understanding of the challenge.

After the subsequent request for a project proposal, Prime Vision turned out to be the best fit. The company is experienced in developing tailor-made solutions, understanding highly critical operational processes, processing of large amounts of data and has in-house development facilities in the Netherlands. This would keep communication lines short, which later turned out to be an important success factor in the project. A potential pitfall was the fact that they were new to streaming media solutions; this risk was diminished by a convincing demonstration of early results on some representative SD test material.

Prime Vision has 60 years of experience reading handwriting and machine printed characters on items of mail, parcels, money transaction and similar forms, etc. They supply OCR based postal automation services to companies worldwide. Apart from the postal market, Prime Vision has the ambition to profile itself as a solution provider in other innovative domains as well. An example of this is license plate recognition for road traffic applications. The functionality requested by NPO/TNO therefore fitted well within their project portfolio.

DEFINING AND MEASURING QUALITY

With reference to the following test images in Figures 1 to 6, it can be seen that although the subtitles seem to be well-structured, these examples illustrate the variance of appearance and some differences in font face which could be fatal if not given due attention.

Figure 1: Near title captions, close to or within the subtitling area

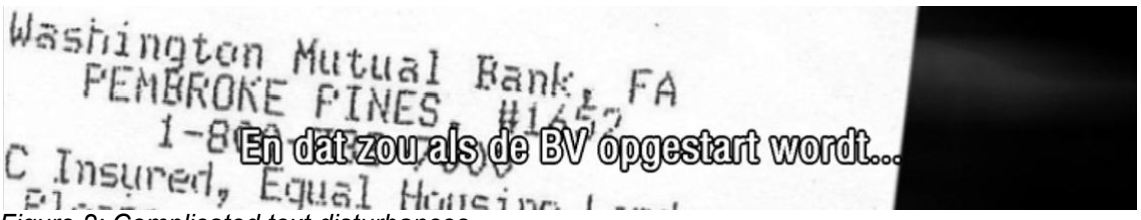

Figure 2: Complicated text disturbances

Figure 3: Presentation differences (e.g. italic text)

Figure 4: Foreign language text; these should automatically be ignored by the OCR

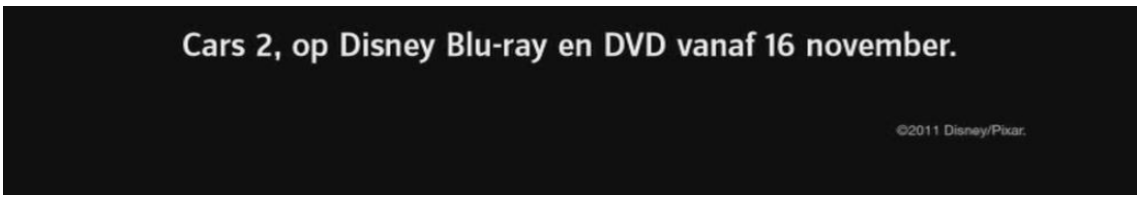

Figure 5: Non-subtitle text within the subtitle area, accepted false positive

Figure 6: Character overlap, non-consistent border

The proposed OCR solution should be robust to these variances.

TNO therefore developed a test plan to measure the quality of the recognized subtitles in a structured way and secure the process of determination whether the requirements would be met.

Two quality parameters were defined:

- detection quality: defining if an OCR supposed subtitle is really a subtitle.
- recognition performance: the measure of correct pattern matching.

Measuring detection quality is not as trivial as it may seem. For example, not every frame contains a subtitle; while if a subtitle is present it always appears in a number of consecutive frames. In the end, the subtitle data should be exported only once. ‘Subtitle Sequence’ detection was therefore an important requirement.

After correct subtitle detection, the character errors should be reduced to a minimum to give reasonable and understandable text output for spoken subtitling. After the initial video-OCR development phase, a group of testers compared a batch of captured frames and OCR subtitle results in an automation assisted procedure. The recognized subtitles were checked and eventually corrected manually. This resulted in a long list of recognized subtitles and their correct appearance. These subtitles were compared to each other and structural errors were identified.

A technique called Levenshtein distance [1] was used for this. It compares two strings and determines the number of edits needed to adjust one string to be the same as the other. In this case, the number of edits to change a recognized subtitle to its correct form were determined. If both subtitles were identical, the number of edits was zero. If one character was missed or miss-

spelled, then one edit would be needed to solve the difference - an addition or replacement, respectively. Needless to say, the lower the Levenshtein distance, the better the recognition quality.

Quality testing in the different development phases thus proved not only the need for optimization, but also provided input towards the specific resolution direction. In this article, we now look into the matter from the OCR perspective.

OCR 101

OCR is a pattern recognition technology that is able to read machine- or handwritten characters. It uses statistical pattern matching algorithms to learn the characteristics from a large number of labeled examples (where labeled means: a logical human-given character representation of the assumed image pixel groups) and uses these characteristics for recognition. OCR is a widely used and successful technique for automatically converting scanned documents (images) to editable text.

As a reference, the human brain is very good at recognizing highly complicated pattern forms. We can recognize deviations and are able to consider the context of patterns in our matching and labeling. This is part of our ongoing conscious and unconscious natural learning processes, starting from birth; for example, think of recognizing faces or objects (identification, whether it is partly covered, facing away or even hidden). Sometimes our learning is actively stimulated in a specific direction, like learning to read our home language in primary school.

Computers are still far away from doing all the pattern recognition tasks that humans can do. However, OCR works quite well in a conditioned environment (e.g. documents, items of mail, parcels); most characters in a mainstream presentation format are well recognized (> 99% correct) by a well-taught OCR system. However, OCR systems are not faultless. Possible causes of errors may be: near-similar characters, image noise, low quality print, unknown fonts, characters not included in the stored set (e.g. Greek, math symbols, currency signs, etc.).

The OCR applied by Prime Vision includes a number of smart technologies that overcome many of the well-known OCR challenges:

1. A best recognition try is done on punctuation marks and diacritics (e.g. é, ë). This is non-trivial, as both punctuation marks and diacritics are rather small (only a few pixels) and could also be background noise.
2. Statistical language models trained for Dutch are used to incorporate context knowledge to disambiguate between characters that look visually similar for OCR - like ‘l’ (upper case i) and ‘l’ (lower case L); ‘O’ and ‘o’ vs. ‘0’ (zero). A typical example in Dutch is to distinguish between the ‘I’ in ‘Ik’ (meaning ‘I, myself’), and the ‘l’ in ‘elkaar’ (each other).
3. Characters are not always nicely separated: typical combinations like ‘f’ and ‘t’ (as in gift), ‘r’ and ‘t’ (as in porto), can be so close together, that the OCR recognizes them as one character (the closest character being the ‘n’). To resolve this problem the OCR training set has been extended and trained on these so-called ‘fused characters’, such that upon encountering a fusion, the OCR output identifies the two correct characters. An example of the fused character problem is shown in Figure 7.

Figure 7: Example of fused characters

VIDEO BASED OCR FOR SUBTITLES

To apply OCR to video, one needs static rather than moving pictures. Therefore, the incoming video is decomposed into its separate picture frames through a 'frame grabbing' process. The result: de-interlaced, uncompressed picture files with 1920 x 1080 pixel resolution and 24 bit pixel depth. This gives a data storage of about 1 GB (gigabyte) per minute or over 1 TB (terabyte) per day. Consequently, requirements for internal RAM memory and storage space, and a limitation to subtitle processing time were set for the development and rollout of the final solution. Picture images have to be saved to disk to some extent for logging, evaluation and tuning purposes. To keep up with the video stream, the requirement is to give a recognition result within 300 ms after the first frame containing a subtitle. Therefore the implemented algorithms needed to be relatively fast, and could not deal with using extensive dictionary lookups or complicated language models.

Next to the above-mentioned technology considerations, the following specific process steps were implemented to achieve optimal subtitle-OCR results and to meet some required key performance indicators, which follow on from the three OCR challenges listed above:

4. Configurable recognition areas: to focus only on the picture area that are likely to contain a subtitle (i.e. the lower third in NPO's case; subtitling in the higher picture area is rare).
5. Image processing techniques: to combine multiple frames to detect stable stationary areas where the subtitles could be.
6. Intelligent segmentation algorithms: to separate the foreground (i.e. the subtitle) from the background (the video image). This is especially useful if the video background contains text as well. See the example image in Figure 8 below.
7. Advanced edge detection: to detect where the subtitle characters are in the video image by taking advantage of the black borders of subtitle characters and the clustering of color regions.
8. N-gram language models: to filter out non-Dutch subtitles (e.g. Turkish subtitles in a double language subtitle for programmes in which the police might be asking the public for help in tracking a missing person). The model contains statistics about a language; in this example, the typical use of character combinations was a main focus of the filtering process.
   The decision whether the OCR output text has valid Dutch wording or not, is based on a threshold of the language model score and the optimal threshold. Thus balancing between false accepts and false rejects, and basing the decision on evaluating a large number of example subtitles.
   The language models also prevent nonsensical OCR results that originate from white/black patterns in the image, that could lead to recognizing something illogical like 'I j ii J L l'.

Figure 8: Video fragments can have text by itself, challenging the OCR.

THE FINAL SUBTITLE CHALLENGE

The so-called 'false positives', text in the picture that was recognized but was not (part of) a subtitle, proved to be the biggest challenge in the project. Think of ticker tapes, straps, superimposed name titles, text elements in commercials, title scrolls, coming-up-next messages, programme schedules etc.

This final challenge was managed to a high degree by:

- Only allowing texts that remain stationary in the image (filters out ticker tapes).
- Only allowing correctly aligned text (e.g. picture centre only).
- Filtering OCR results which have a low confidence score from the language model, and/or do not contain normal words at all (filters out low quality texts and other texts like website links and email addresses).
- Not allowing caps-only texts (these occur very rarely in normal subtitles, but are very common for name titles).
- Tuning the threshold for reading only (near)-white characters containing a black border; not all subtitle texts have black borders all around, so some missing black is allowed. However, this prevents reading characters that do not have a border at all and are therefore likely to originate from other types of texts in the background.

In addition to an independent evaluation of the system performance, TNO also advised on the approach for handling the technical challenges that were encountered. This was achieved by deep analysis of the errors, discovery of structures and constructive discussion with Prime Vision and NPO, on how and to what level the structured errors should be handled. These discussions were possible due to trust, an open and creative attitude, short discussion lines and the drive to really solve the problem. Most importantly, they led to the required solution, where NPO's requirements and specifications were met:

- The total read rate, the amount of identified subtitles by the OCR process in time, is higher than 99%. On subtitle material that is fully compliant to the current NPO specification (i.e. correct font type, size, positioning, outline) the score is 100%.
- The word-error rate on recognized subtitles is below 1%.
- As a trade-off towards the highest possible read rate for programmes, some false positives were accepted on non-programme material only. The tuning of the system limits the number of cases where users are surprised by occasional invalid subtitles that are read aloud between the genuine subtitles within a programme.

LESSONS LEARNED FROM THE BROADCASTER PERSPECTIVE

In the project and following day-to-day video OCR practice, the following points became clear:

- We think that TV is about moving pictures … but static text in many forms is a major part of our video communications.

- Correct visual design is a critical success factor for overall OCR performance. This might be a challenge in a creative and imaginative on-air-marketing world. Technology should never take precedence over creative content, but the two can and must be tuned to each other.
- Attention to subtitling is required in the form of production policy and specification. Good subtitling is not a simple part of the (attic room) editing process; specific expertise is still needed. In the Netherlands, we therefore see the traditional subtitle text file delivered to the video editor being replaced by a pre-formatted transparent QuickTime file that will be overlaid during editing. In this way, a uniform and specialist-made presentation is guaranteed, to have the flexibility of anywhere, anytime, anyhow editing.

SUMMARY & CONCLUSIONS

LIVE SITUATION: 100% OF SUBTITLED PROGRAMMES AVAILABLE FOR SPEECH SYNTHESIS

Started as a pilot in 2011, NPO has now deployed the OCR solution on all three public premium television channels since March 2012.

The broadcast mixed text-to-speech solution, based on Nuance speech technology was developed and delivered by Ericsson in parallel to this work.

After some months, NPO concluded that the subtitle data generated by the OCR system was of such quality that it now should be the only spoken subtitle source. The technical equipment for deriving the data from subtitle file or ancillary data inserted by the subtitling systems in remote studio feeds or playout has been dismantled, thus removing single points of failure and increasing reliability in the playout chain.

The target audience is very enthusiastic and satisfied with the improved services. Complaints have been reduced to a minimum.

Nevertheless, system development does not stop: off-air, top-to-tail-cut output data will soon be used for metadata enrichment purposes; and within NPO's playout replacement project, a Softel 'Newfor' protocol implementation will provide for real time OCR in Miranda iTX playout communications. The authors wish that the live subtitling format coming out of the ongoing work on the EBU-TT part 3 specification [2] was already available for that but, as always, technological improvement comes step by step … or for subtitling maybe character by character...

REFERENCES

[1] http://en.wkipedia.org/Wiki/Levenshtein_distance

[2] http://tech.ebu.ch/ebu-tt

AUTHOR BIOGRAPHIES

MARCO SLIK

Marco Slik is a Senior Policy Adviser, Research & Development, for the Netherlands public broadcaster, NPO.

After an extensive career in broadcast project management, consultancy and architectural design for systems integrators and service providers, Marco has worked for NPO distribution and broadcasting since 2010. His primary field of expertise lies in TV playout and digital distribution, though he can find his way in audio and the connected world just as well. He leads R&D and infrastructural change projects and advises on innovation policy.

HANS JONGEBLOED

Hans Jongebloed works in the position of R&D consultant, at Prime Vision B.V., in the Netherlands.

Hans is an expert in speech recognition, language models and pattern recognition and has worked for KPN Research, Dutchear, and Prime Vision respectively since 1997. He has a degree in applied physics and artificial intelligence. His main work topics are R&D innovation and design of new solutions for new customer automation challenges in the area of speech technology and OCR.

MARK VAN STAALDUINEN

Mark van Staalduinen works in the area of Media and Network Services consultancy, at TNO in the Netherlands.

After a scientific career in computer vision and pattern recognition techniques applied to large databases, Mark now works as a consultant and project manager on innovation projects for TNO. His drive is to develop cutting edge solutions in the media or safety domains in cooperation with different partners from universities, government and commercial technology providers.

Published by the European Broadcasting Union, Geneva, Switzerland

ISSN: 1609-1469

Editor-in-Chief: Simon Fell

Managing Editor: Eoghan O'Sullivan

E-mail: osullivan@ebu.ch

Responsibility for views expressed in this article rests solely with the author(s).